FROM PARETO FRONTS TO STATISTICAL EVIDENCE: SELECTIVE INFERENCE FOR MODELS AND POLICIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-objective evaluation often uses the same noisy data both to select a Pareto shortlist and to support claims about the selected models or policies. This creates a post-selection problem: uncertainty statements that ignore how the shortlist was chosen need not provide the advertised reliability for the report that is actually published. We develop report-conditional inference for complete Pareto reports around a structural observation: the reported candidates themselves provide the dominance witnesses needed to reconstruct the entire selection event. Along an inferential line, each required dominance relation occupies an interval, so a report of size s among K candidates can be reconstructed from only $s(K-1)$ directed relations. This yields support construction in $O(Ksd + Ks\log(Ks+1))$ with near-optimal dependence on Ks for fixed d and sparse reports. Under joint Gaussianity, the reconstructed support gives exact conditional inference and enables scalar carving, which reuses selection data while preserving the same reported labels. Across Gaussian designs, model libraries, and policy evaluations, the method maintains report-level calibration while recovering substantially more information than matched splitting. In the primary image setting, carving reduces mean paired interval width by 33.6%; in a 1,024-configuration Bank library, report-sensitive reconstruction yields a $1.95\times$ end-to-end median acceleration over all-pair compression.

1 INTRODUCTION

Multi-objective evaluation preserves trade-offs among accuracy, robustness, calibration, risk and other performance goals. Researchers therefore publish Pareto shortlists so that users can compare relevant compromises rather than collapse all objectives into a single ranking (Feurer et al., 2023; Daulton et al., 2021). Our model pool makes this need concrete: AugMix ResNeXt achieves higher clean accuracy, Gauss50 is substantially stronger under Gaussian noise, and PRIME ResNet18 contributes a nondominated compromise while winning none of the five individual objectives (Table 31).

The statistical difficulty begins after the shortlist is selected. The same finite evaluation sample that makes a candidate appear nondominated is often reused to report its performance or support comparisons involving that candidate. Because the published labels are data dependent, uncertainty statements that ignore selection need not provide their advertised reliability for the report that is actually observed. This effect is practically relevant: in our controlled equal-means design, nominal 95% naive coordinate intervals attain only 90.7% coverage (Table 2). More subtly, aggregate protection and protection of a realized shortlist are different targets. For a two-label report occurring with probability 0.164, report-conditional coverage is 0.951 with carving but 0.878/0.889 for global/locally simultaneous procedures, even though the latter retain probability-weighted coverage 0.965/0.967 across reports (Table 3; Appendix N.1).

We therefore study *report-conditional evidence*: advertised coverage holds separately across repeated evaluations producing the same complete label set. Fixing the realized report also fixes the prespecified comparison family and supplies a common event for allocating report-level error. This complements simultaneous and locally simultaneous inference, which protect broader families of possible questions (Zrnic & Fithian, 2024; Rink & Brannath, 2025), while conditional inference explicitly incorporates the realized selection event (Lee et al., 2016; Liu et al., 2022).

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.

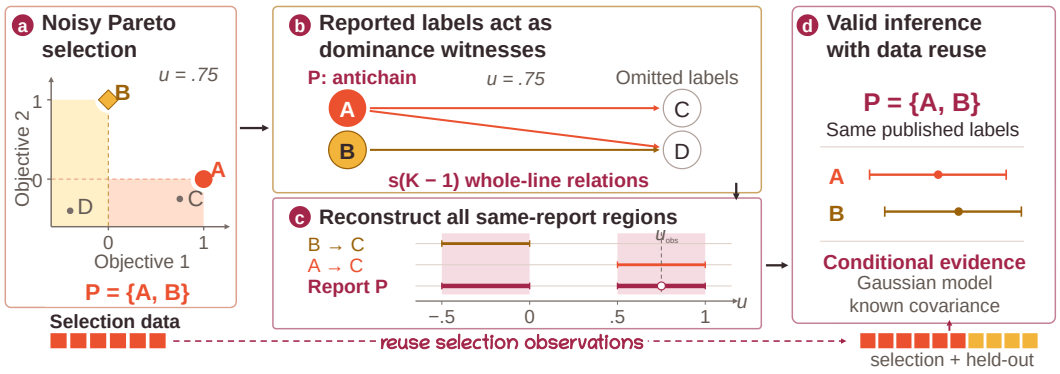

Figure 1: From noisy Pareto selection to statistical evidence. (a) Noisy scores select $P = \{A, B\}$ at $u = .75$. (b) Reported labels witness every omission. (c) With $A = (1, 0)$, $B = (0, 1)$, $C = (u, .5-u)$, $D = (-.4, -.4)$, the complete support is $[-.5, 0] \cup [.5, 1]$. (d) Same-report scalar carving reuses selection data, with conditional validity under the stated Gaussian model.

The main challenge is structural. Every reported candidate must survive domination, while every omitted candidate must have a dominator; moreover, the same complete report can occur on disconnected regions along an inferential line. A direct construction appears to require all candidate pairs. Our key observation is that a complete Pareto report is instead *self-certifying*. As Figure 1 illustrates, the reported labels form an antichain and themselves provide the dominance witnesses for every omission. For a report P of size s, only $s(K-1)$ directed relations are therefore required. Each occupies an interval along the inferential line, and combining them recovers every region producing exactly the same report, including disconnected components.

This report-centered representation has both computational and statistical consequences. We reconstruct the complete selective support in $O\{Ksd + Ks\log(Ks+1)\}$ operations per line, and an explicit-output lower bound shows near-optimal dependence on Ks for fixed $d \geq 2$ and sparse reports. Under joint Gaussianity with known covariance, the reconstructed support yields the exact conditional law for coordinates and prespecified comparisons. It also makes same-report scalar carving computationally tractable (Drysdale, 2023): the selection sample fixes the labels, while inference reuses its observations rather than discarding them as in matched splitting.

Our contributions are:

1. **Self-certifying report structure.** Only $s(K-1)$ reported-source dominance relations characterize the complete same-report selection event. Their interval representation recovers every region producing the report, including disconnected components.
2. **Near-optimal support reconstruction.** Reconstruction costs $O\{Ksd + Ks\log(Ks+1)\}$ operations per line. Supports with $(s-1)(K-s)+1$ components establish an $\Omega(Ks)$ explicit-output lower bound for $2 \leq s \leq K/2$, giving near-optimality for fixed $d \geq 2$.
3. **Exact inference from the complete support.** Under joint Gaussianity with known covariance, nuisance conditioning yields an exact finite-union truncated-normal law and computable report-conditional pivots. The support also enables same-report scalar carving and report-level error allocation.

Gaussian designs, image/tabular model libraries and policy evaluations assess the resulting calibration and information recovery. In the primary image setting, carving reduces mean paired marginal widths by 33.6% relative to matched splitting. All six fixed-report image/tabular budget settings improve mean precision and true-comparison yield over splitting. A 1,024-configuration Bank library yields a $1.95\times$ median end-to-end acceleration over all-pair compression.

2 Multi-Objective Evaluation and Validity Targets

The input is a fixed library of K candidates evaluated on d objectives, all oriented so that larger values are better. We write $[K] = \{1, \ldots, K\}$ and $[d] = \{1, \ldots, d\}$. Candidate k has population

performance vector $\theta_k \in \mathbb{R}^d$. Models, checkpoints, configurations or policies must be trained, and objectives chosen, outside evaluation.

The *selection view* $S \in \mathbb{R}^{K\times d}$ contains the scores used to choose the report. Candidate j dominates i, written $j \succ_S i$, when

$$S_{j,c} \geq S_{i,c} \quad \text{for every } c, \qquad S_{j,c} > S_{i,c} \quad \text{for some } c.$$

The empirical Pareto front retains all nondominated labels:

$$\widehat{\mathcal{P}}(S) = \{i \in [K] : \nexists j \neq i \text{ with } j \succ_S i\}.$$

Conditioning fixes the complete shortlist of candidate labels while scores remain random; nondominated ties are retained.

The *inference view* $T \in \mathbb{R}^p$ estimates performance; its relevant mean entries are $\theta_{k,c}$. It may use more data than S. After observing $\widehat{\mathcal{P}}(S) = P$, we consider a direction $\eta_{P,q}$ that is fixed for each possible reported set P and report item q. Its statistic and parameter are

$$U = \eta_{P,q}^\top T, \qquad \nu_{P,q} = \eta_{P,q}^\top \mathbb{E}[T].$$

A weight $+1$ selects a coordinate; weights $+1, -1$ compare coordinates. Fixed utilities and threshold margins are allowed. Directions may depend on P; further score-based selection requires extra conditioning. Comparisons to an unreported fixed reference are included when both estimates appear in T.

Exact Gaussian model. The finite-sample result assumes

$$X = \big(\operatorname{vec}(S), T\big) \sim \mathcal{N}(\mu_X, \Sigma_X), \qquad \Sigma_X \text{ known}, \qquad \operatorname{Var}(U) > 0.$$

Here vec stacks the score table. Singular covariance is allowed, including duplicate scores when $S = T$.

Gaussian approximation. Per-example averages and asymptotically linear policy values (Chernozhukov et al., 2018) motivate a Gaussian approximation. Proposition 2 transfers validity asymptotically under fixed K, d, non-rare reports and continuity.

Report-level goals. For $Q(P) = P \times [d]$ and $m_P = |P|d$, let C_q denote a reported interval and $V_P = \sum_{q\in Q(P)} \mathbf{1}\{\theta_q \notin C_q\}$. We study

$$\mathbb{E}\left[\frac{V_{\widehat{\mathcal{P}}}}{m_{\widehat{\mathcal{P}}}}\right] \quad \text{and} \quad \mathbb{P}\Big(\theta_q \in C_q \; \forall q \in Q(P) \mid \widehat{\mathcal{P}} = P\Big).$$

The false coverage rate (FCR) averages the within-report fraction of missed targets; conditional simultaneous coverage protects all coordinates of the realized label set.

3 From Report Structure to Exact Conditional Inference

We first characterize the complete report event, then establish its reconstruction cost and obtain the exact conditional law on its support. Fix a realized front P and one direction $\eta = \eta_{P,q}$. Let

$$\sigma^2 = \operatorname{Var}(U), \qquad \rho = \frac{\operatorname{Cov}(X, U)}{\sigma^2}, \qquad Z = X - \rho U.$$

Under the Gaussian model, $U \perp Z$; conditional on $Z = z$, $X(u) = z + \rho u$ and $U \sim \mathcal{N}(\nu, \sigma^2)$. Thus ρ determines how the selection scores move with the target statistic.

Complete reconstruction from reported sources. Let $s = |P|$ and $D_{p\to i}(z) = \{u : p \succ_{S(X(u))} i\}$ under the dominance rule in Section 2. Define the reported-source region $R_i^P(z) = \bigcup_{p\in P\setminus\{i\}} D_{p\to i}(z)$ for every label i. Reproducing P requires avoiding this region for reported labels and entering it for omissions.

Theorem 1 (Report-sensitive reconstruction and output complexity)**.** *For fixed z and nonempty P, the pointwise complete-report event is*

$$\mathcal{E}_P(z) = \{u : \widehat{\mathcal{P}}(S(X(u))) = P\} = \bigcap_{i \in P} (R_i^P(z))^c \cap \bigcap_{i \notin P} R_i^P(z).$$

Only $s(K-1)$ directed relations are required. Each occupies an interval, modulo finitely many endpoints, so $\mathcal{E}_P(z)$ is a finite interval union. For every $K > s \geq 2$, there is a two-objective affine line whose complete-report support has exactly $(s-1)(K-s)+1$ positive-length components. Under the stated Gaussian model, whenever its conditional mass is positive,

$$U \mid \{Z = z, \widehat{\mathcal{P}} = P\} \sim \mathrm{TN}\big(\nu, \sigma^2; \mathcal{E}_P(z)\big).$$

Here TN restricts and renormalizes a normal law over the full interval union; endpoints have zero Gaussian mass. Proofs, endpoint/component bounds and Algorithm 1 appear in Appendix A; implementation checks appear in Appendix D.

Near-optimal explicit reconstruction.

Proposition 1 (Complexity)**.** *Algorithm 1 computes $\mathcal{E}_P(z)$ in $O\{Ksd + Ks\log(Ks+1)\}$ arithmetic/comparison operations and $O(Ks)$ auxiliary words, excluding the supplied affine arrays.*

For $2 \leq s \leq K/2$, Theorem 1 gives examples with at least $Ks/4$ output components, requiring $\Omega(Ks)$ operations for explicit reconstruction in the worst case. For fixed $d \geq 2$, our bound is therefore optimal up to a logarithmic factor; at $s = K$, its order matches all-pair compression. The bound measures support construction; complete-output timings also include selection, covariance estimation and inversion.

Exact inference from a finite interval union. Write $\mathcal{E} = \mathcal{E}_P(z) = \bigcup_r (\ell_r, h_r)$ and let ϕ_σ denote the $\mathcal{N}(0, \sigma^2)$ density. The conditional CDF is

$$F^{\mathcal{E}}_{\nu,\sigma}(u) = \frac{\displaystyle\int_{\mathcal{E} \cap (-\infty, u]} \phi_\sigma(v - \nu)\, dv}{\displaystyle\int_{\mathcal{E}} \phi_\sigma(v - \nu)\, dv}.$$

Both integrals sum normal CDF differences over *all* components, including the disconnected regions in Figure 1c. At the true ν, $F^{\mathcal{E}}_{\nu,\sigma}(U)$ is uniform conditional on $(Z = z, \widehat{\mathcal{P}} = P)$. Inverting $\alpha/2 \leq F^{\mathcal{E}}_{\nu,\sigma}(u) \leq 1 - \alpha/2$ gives exact intervals; Appendix A.2 proves monotonicity and coverage.

Proposition 2 transfers validity asymptotically for fixed K, d, non-rare reports and the covariance/continuity conditions in Appendix B.

4 From Selective Support to Statistical Evidence

The exact law supplies target-wise inference; the common report event aggregates it into report-level guarantees (Benjamini & Yekutieli, 2005; Bakshi et al., 2026). Conditional miss probabilities at most $\alpha_q(P)$ give FCR at most α when $\alpha_q = \alpha$, and report-conditional simultaneous coverage at least $1 - \alpha$ when $\sum_q \alpha_q(P) \leq \alpha$. Corollary 1 and its proof appear in Appendix A.

Publishing one protected report. Prespecify the comparison rule and budgets $\alpha_C + \alpha_D \leq \alpha$. For P selected using S, reconstruct the prescribed supports and apply scalar carving to T (Section 4.1). Coordinate bands at $\alpha_C/(|P|d)$ and Holm claims at α_D (Holm, 1979) control any coordinate miss or false claim at α, conditional on $\widehat{\mathcal{P}} = P$, under the Gaussian carving model. Resource experiments use .025 each; marginal contrast intervals remain descriptive (Appendix K).

Reusing a protected report. Simultaneous bands $C_{k,c} = [L_{k,c}, U_{k,c}]$ protect every nonnegative utility through $w^\top L_k \leq w^\top \theta_k \leq w^\top U_k$ for all reported k and $w \geq 0$. Monotonicity gives $\mathrm{HV}_r(L_P) \leq \mathrm{HV}_r(\theta_P) \leq \mathrm{HV}_r(U_P)$ above a fixed reference r; lower endpoints certify fixed thresholds. Proofs and threshold comparisons appear in Appendices A and F.

4.1 SAME-REPORT DATA REUSE

The support enables scalar carving (Fithian et al., 2014; Tian & Taylor, 2018; Drysdale, 2023): a fraction f selects labels and all records estimate performance (Figure 1d). For nested Gaussian means with known full-sample covariance V, $S = T+\omega$, where $\omega \perp T$ and $\omega \sim \mathcal{N}(0, (1-f)V/f)$. Carving and matched splitting retain the same $\widehat{\mathcal{P}}(S)$. Write $\rho = V\eta/\sigma^2$, $Y = \eta^\top S$, and $Z_S = S - \rho Y$. Let $\mathcal{A}_P(Z_S) = \bigcup_j (a_j, b_j)$ be the complete-front support along $S(y) = Z_S + \rho y$. With $0 < f < 1$, $\sigma^2 = \eta^\top V \eta > 0$ and $\tau^2 = (1-f)\sigma^2/f$, integrating the selection scalar gives

$$f_\nu(u \mid Z_S, P) \propto \phi_\sigma(u-\nu) \sum_j \left[\Phi\left(\frac{b_j - u}{\tau}\right) - \Phi\left(\frac{a_j - u}{\tau}\right) \right].$$

One support sweep and one-dimensional quadrature evaluate this weight. Proposition 3 establishes the Gaussian pivot and the resulting report guarantees. Fixing ω gives Theorem 1's hard-conditioned law; independent fixed-scale offsets enter the selection intercept (Appendix J).

The finite-union bound of Kivaranovic & Leeb (2024, Theorem 3.1) gives shorter intervals than matched Gaussian splitting at the same budget under this model (proofs and residual integration: Appendix C).

4.2 QUANTITATIVE CONTROL FOR BOUNDED SCORES

Covariance estimation perturbs the variance, inferential line and support. The following bound controls the resulting scalar-carving pivot error.

Theorem 2 (Quantitative scalar-carving approximation)**.** *Under the bounded-i.i.d. setup of Appendix O.3, use the full-sample unbiased covariance and a target prescribed by P. Suppose its stated variance bounds hold and the complete report has probability $\pi_n \geq \pi_* > 0$. With the explicit nonnegative terms c, B, r defined there, the plug-in pivot at its true target satisfies*

$$\sup_{t\in[0,1]} |\mathbb{P}(G_n \leq t \mid \widehat{\mathcal{P}} = P) - t| \leq \Delta_{n,P,q} := \min\left\{1, \frac{c + 2B + 2r}{\pi_*}\right\}. \tag{1}$$

With fixed dimension, f bounded away from zero and one, and fixed positive variance, spectral and report-probability envelopes, the stated tuning gives $\Delta_{n,P,q} = O(n^{-1/4})$.

Here c bounds the Gaussian-kernel change, B the expected support-mass change, and r the coupling, covariance and report-change failure probability. Support sensitivity uses only $[\binom{s}{2} + s(K-s)]d$ incident comparisons. Averaging over Gaussian nuisance statistics cancels the local inverse selection mass; the bound retains the complete report's frequency. Appendix O gives all constants and the proof.

Calibrating reported errors. A certified constant envelope across candidate nulls permits equal-tail inversion at $\alpha_q - 2\Delta_q > 0$; otherwise use the parameter range. A uniform one-sided null bound gives conservative $\min\{1, p_q + \Delta_q\}$ values for Holm testing (Appendix O.4). The independent empirical-Bernstein option is a separate bounded-score mode (Appendix N.2).

5 EXPERIMENTS

Experiments test complete-support reconstruction, its scaling cost and report-conditional reliability, then measure information recovered by data reuse. Paired methods share records and reports; Monte Carlo uncertainty uses independent reports or training-library averages for policies.

5.1 COMPLETE RECONSTRUCTION AT REPORT-SENSITIVE COST

Preserving the entire selection event. Independent constructions agree on 318 supports and 270 target intervals, including ties, zero slopes, singular $S = T$ and recurrent reports (Appendix G). Selected contrast supports have two to four components in the controlled Gaussian study (Appendix K), exercising the disconnected regions required by Theorem 1.

Table 1: From sparse-support scaling to real-report computation. R/A: report-sensitive/all-pair seconds; times are medians of three paired runs. A: supplied covariance, $s = 4, d = 32$; Report includes all sd coordinate intervals. B: trained Bank library, $d = 5$; raw-row clocks start from loaded, warmed models. Ratios divide median times in panel A and use paired raw-row ratios in panel B.

A. Sparse reports with known covariance				
Output	K	R	A	Ratio
Support	128	0.014	0.101	7.3
Support	512	0.055	1.511	27.4
Support	1024	0.113	5.934	52.6
Support	2048	0.229	23.532	102.7
Report	1024	14.548	763.258	52.5
B. Complete reports from real Bank scores				
		Scores-to-report	Raw-rows-to-report	
K	s	R/A	R/A	Ratio
64	36–40	7.36/7.44	8.44/8.51	1.01
128	55–61	12.09/12.39	14.27/14.57	1.01
256	90–100	21.70/23.45	25.87/27.65	1.07
512	127–165	38.39/50.70	46.59/58.91	1.26
1024	180–234	58.06/128.94	74.21/145.08	1.95

Table 2: Exact-Gaussian calibration, 10,000 independent reports per cell with known covariance. Cov.: pooled-coordinate coverage (selective/naive). Selective FCR averages marginal miss fractions across reports; Sim. uses separate selective simultaneous bands. Full outputs: Table 5.

	Trade-off means			Equal means		
Covariance	Cov.	FCR	Sim.	Cov.	FCR	Sim.
Independent	.952/.945	.048	.958	.951/.922	.049	.953
Equicorrelated	.951/.953	.049	.956	.948/.931	.052	.955
Block	.949/.943	.051	.966	.952/.907	.049	.968
Near-singular	.950/.950	.050	.991	.950/.948	.050	.971

Scaling from sparse supports to real reports. With $s = 4, d = 32$, support gains increase from 7.3-fold at $K = 128$ to 102.7-fold at $K = 2048$; all-coordinate inference at $K = 1024$ gains 52.5-fold (Table 1A). Full grids, dense controls and numerical agreement appear in Appendix N.3. For real scores, we train 1,024 Bank configurations from four families and evaluate 15,211 rows on five objectives. Three paired reports per size use $|A| = 3000$, $|B| = 2000$. At $K = 1024$, $s = 180$–234: median paired gains are 4.84-fold for supports, 2.22-fold from scores to reports and 1.95-fold including fresh scoring (Table 1B). Both backends compute covariance, all coordinates and contrasts, adaptive refinement, Holm decisions, precision and exact hypervolume; shared kernels differ only in support sources, P versus $[K]$. Training, loading and compilation are timed separately (Appendix P).

5.2 From the Exact Law to Conditional Reliability

Table 2 retains all eight known-Gaussian cells: selective coordinate coverage is 0.948–0.952 and FCR is 0.048–0.052, including the equal-means/block cell where naive coverage falls to 0.907. Full outputs and numerical checks appear in Appendix E.

With 1,000 paired reports per covariance, scalar carving reduces mean simultaneous width by 12.6%–35.4% relative to splitting. Global and locally simultaneous procedures target unconditional reliability (Table 8; Appendix F). The following experiment isolates this distinction for a realized report.

When the realized report changes the reliability question. An additional Gaussian design compares means (δ, δ) and $(0, 0)$. For $\delta > 0$, a two-label report preserves an apparent trade-off involv-

Table 3: Same-report coverage at $\delta = 2, f = .9$: 4,000 conditional draws per row, analytic $\Pr(P)$, nominal error .05. Carving/splitting target conditional protection; global/LSI target aggregate protection. Full grids and MC intervals: Appendix N.1.

P	$\Pr(P)$	Carving	Splitting	Global	LSI
$\{1\}$	0.8284	0.952	0.951	0.982	0.983
$\{2\}$	0.0081	0.952	0.952	0.900	0.909
$\{1, 2\}$	0.1636	0.951	0.954	0.878	0.889

ing a population-dominated candidate. We sample all three reports separately to measure reliability *within the published shortlist*. At $\delta = 2, f = .9$, the two-label report occurs with probability .164: carving coverage is .951, versus .878/.889 for global/LSI (Table 3). Separate .05 threshold tests give false-report rates .024 versus .099/.089. Global/LSI retain probability-weighted coverage .965/.967 across reports. Thus their aggregate protection can coexist with a systematic deficit on a specific, non-negligible report. Proposition 4 establishes a limiting separation; Appendix N retains all 30 rare/common groups and uncertainty checks.

Independent confirmation within real shortlists. For each image/tabular pool below, we freeze three complete label sets using earlier frequencies and collect 2,000 new replications per set. With plug-in Gaussian covariance, carving's separate .05 simultaneous coverage is .958–.9765. Under the joint .025/.025 allocation, all four methods' adjusted Monte Carlo risk upper bounds are below .05 across these six groups; carving exceeds splitting in true-comparison yield in each (Appendix P.1). Event-matched learned-law baselines (Liu et al., 2022), including complete-output cost and approximation checks, appear in Appendix I.

5.3 Same-Report Data Reuse on Model Scores

The image pool contains 21 accessible pretrained models from a frozen 24-model RobustBench roster (Croce et al., 2021) and 10,000 CIFAR-10 images. Five paired scores measure clean accuracy, negative Brier and severity-3 Gaussian-noise, defocus-blur and brightness accuracy (Hendrycks & Dietterich, 2019) (Appendix L). Bank Marketing (Moro et al., 2014) uses 16 frozen models, 30,000 training and 15,211 evaluation rows, excluding call duration. Its five scores measure accuracy, probabilistic quality and positive/negative correctness yields. Fixed-pool resampling makes coverage and true comparisons measurable; application inference uses plug-in Gaussian covariance.

Table 4: Image-pool precision and comparisons, $n = 5000, f = .6$. A/B use independent batches of 1,000 reports; marginal, simultaneous and test budgets are separately .05. FCR is marginal; report coverage is simultaneous. R: paired width ratios to splitting, averaged within then across reports; Δ_C: carving's paired true-claim gain. Parentheses: report-based MC SE; false counts: totals. All methods use plug-in Gaussian inference; global/LSI protect unconditional families, unadjusted tests individual claims.

A. Coordinate precision and report coverage

Method	FCR	Report cov.	R_{marg}	R_{sim}
Carving	.050 (.003)	.961 (.006)	.664 (.002)	.666 (.002)
Splitting	.047 (.003)	.972 (.005)	1.000	1.000
Hard conditioning	.062 (.004)	.953 (.007)	1.955 (.245)	2.238 (.295)
Global max-T	—	.973 (.005)	—	.663 (.001)
LSI max-T	—	.967 (.006)	—	.649 (.001)

B. Independent fixed-reference comparison confirmation

Method	True/report	False claims	False reports	Δ_C
Carving + Holm	12.650 (.048)	1	1	—
Splitting + Holm	11.966 (.049)	2	2	.684 (.026)
Global Holm	12.393 (.045)	1	1	.257 (.018)
Global max-T	12.443 (.045)	1	1	.207 (.017)
Unadjusted	13.318 (.051)	10	10	−.668 (.025)

More precise and more informative reports. At $n = 5000, f = .6$, carving's mean target-paired width reductions are 33.6% for marginal intervals and 33.4% for simultaneous bands. The latter attain aggregate empirical report coverage 0.961 (Table 4A). A separate confirmation tests clean-accuracy noninferiority within two percentage points and corruption improvements above one point. Carving gains 0.684 true comparisons over splitting and 0.207 over global max-T per report (Table 4B; paired 95% MC intervals $[.633, .735]$ and $[.174, .240]$). Table 4B also reports false-claim counts; hypervolume projections appear in Table 17.

Holding the shortlist fixed across evaluation budgets. Each of 1,000 further replications per pool fixes the report using $|A| = 3000$ and independent jitter, with nested $|B| = 1000, 2000, 4000$. Carving reuses A; splitting estimates from B. Coordinates and reference comparisons share a $0.025/0.025$ budget. Precision requires full width $\leq .04$ times the known score range. Fixed labels isolate information recovery from report changes.

At $B = 1000$, gains reach 1.550 comparisons for CIFAR and 2.870 for Bank (Figure 2). Full grids and width diagnostics appear in Appendices J–K. Across all six fixed-report budget settings, carving improves mean paired widths and mean true-comparison counts over splitting.

Geometric work and information retention are distinct. In the uncorrelated all-trade-off Gaussian design ($K = 8, d = 3, |A| = 100, |B| = 50$), increasing the whole-layout multiplier from .5 to 2 raises mean directed relations per target from 27.636 to 39.970. Meanwhile, mean oracle log nuisance-conditioned support mass increases from -1.019 to $-.575$, and the mean paired carving/splitting width ratio falls from .812 to .753 (500 reports per setting; Appendix K). This controlled contrast separates the relation-count workload from the selective normalizer and attainable precision.

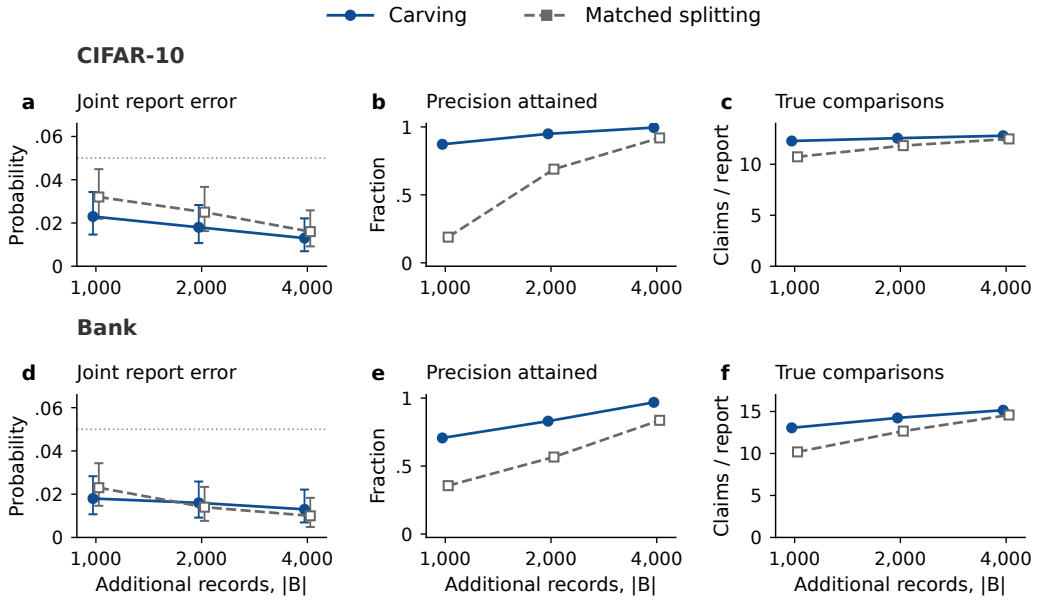

Figure 2: Same-report information recovery, 1,000 paired replications per pool; $|A| = 3000$, nested B. Joint error: any coordinate miss or false claim; precision: fraction meeting the width requirement; dotted line: .05 Gaussian target. Whiskers: pointwise 95% binomial intervals for error, mean ± 1.96 MC SE otherwise, using reports as units. Global/LSI results: Tables 24–25.

Additional score constructions and approximation checks. Independently trained policy libraries yield FCR .044–.054 across randomized-treatment designs (Table 7); checkpoint results appear in Appendix H. Appendix O evaluates covariance perturbations on real scores and exact-tail calibration in a bounded Bernoulli-factor family; Appendix N.2 evaluates independent empirical-Bernstein reports. Appendix M supplies complete report displays and figure sources.

6 Related Work

Multi-objective learning and evaluation. Noisy optimization, population-front identification and random Pareto surfaces support the search for useful trade-offs (Daulton et al., 2021; Kone et al., 2024; Tu et al., 2024). Pareto Testing and reliability graphs select risk-controlled configurations (Laufer-Goldshtein et al., 2023; Farzaneh & Simeone, 2025); SNPL jointly learns and evaluates policies (Cho et al., 2025). Our method provides conditional evidence for the population values and prespecified comparisons of labels in the complete empirical report.

Selection-aware inference. Simultaneous performance bounds and LSI protect global or localized question families (Rink & Brannath, 2025; Zrnic & Fithian, 2024). Winner and rank-based inference exploit ordered outputs (Andrews et al., 2024; Reid et al., 2017; Petrou-Zeniou & Shaikh, 2026; Bakshi et al., 2026). Black-box and flow methods learn selection probabilities or selected laws (Liu et al., 2022; Liu & Panigrahi, 2025). Our method directly calculates the complete support of the known Pareto rule, including disconnected regions.

Geometric foundations and data reuse. Dynamic skyline methods use dominance certificates (Frederickson & Rodger, 1990; Huang et al., 2006). Our report-centered characterization reconstructs the entire selective support and establishes near-optimal explicit-output complexity. Its finite-union representation supplies the conditional normalizer. Gaussian selective pivots, FCR control and scalar carving (Lee et al., 2016; Benjamini & Yekutieli, 2005; Drysdale, 2023) then convert this representation into report-level inference.

7 Discussion and Conclusion

What report structure makes possible. A complete Pareto report is *self-certifying*: its labels form an antichain and witness every omission. The $s(K-1)$ reported-source relations reconstruct the entire same-report support in $O\{Ksd + Ks\log(Ks+1)\}$ operations. Supports with $\Omega(Ks)$ components make reconstruction optimal up to a logarithmic factor for fixed $d \geq 2$ and $2 \leq s \leq K/2$ among explicit-output algorithms. Under the stated Gaussian model, nuisance conditioning yields an exact truncated-normal law on this finite interval union and computable report-conditional pivots.

Conditional versus aggregate protection. Report-conditional and simultaneous procedures answer different questions. Global and locally simultaneous methods protect broader families of possible queries, whereas our target asks whether advertised reliability holds for the particular complete report that is published. Our report-stratified experiments show that aggregate protection can coexist with substantially different coverage on a specific, non-negligible shortlist. When downstream claims concern the realized Pareto report itself, report-conditional validity directly matches that inferential target.

Scope and limitations. Finite-sample exactness of scalar carving assumes joint Gaussianity, known covariance and the stated nested-sample model. Asymptotic transfer assumes fixed K, d, continuity and non-rare reports; the quantitative correction additionally requires moment, variance and report-probability envelopes (Appendix O). The general envelope can be conservative, and the CIFAR/Bank experiments use a plug-in Gaussian approximation; some small-sample image settings exhibit undercoverage (Appendix J). Sharper computable corrections for broader score distributions and adaptive evaluation settings remain important directions.

Conclusion. A complete Pareto report supplies the dominance witnesses needed to reconstruct its own selection event. Report-sensitive reconstruction makes conditional inference and same-report data reuse computationally tractable, yielding narrower intervals and more supported comparisons than matched splitting while scaling to thousand-candidate libraries. More broadly, the framework turns a noisy Pareto shortlist into a report with explicit, selection-aware statistical evidence.

AI USE STATEMENT

Generative AI tools were used solely for language polishing and improving the readability of the manuscript. The authors reviewed all suggested edits and take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

The appendix provides proofs, numerical definitions and frozen evaluation protocols. It reports support-backend cross-checks, Monte Carlo uncertainty and approximation diagnostics. Implementation and audit details are specified alongside each experiment.

REFERENCES

Isaiah Andrews, Toru Kitagawa, and Adam McCloskey. Inference on winners. *The Quarterly Journal of Economics*, 139(1):305–358, 2024. doi: 10.1093/qje/qjad043.

Soham Bakshi, Lingjun Gao, Zijun Gao, and Snigdha Panigrahi. Flexible inference for winners with conditional validity, 2026. URL https://arxiv.org/abs/2607.18545.

Yoav Benjamini and Daniel Yekutieli. False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81, 2005. doi: 10.1198/016214504000001907.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

Brian Cho, Ana-Roxana Pop, Ariel Evnine, and Nathan Kallus. SNPL: Simultaneous policy learning and evaluation for safe multi-objective policy improvement, 2025. URL https://arxiv.org/abs/2503.12760.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. RobustBench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. URL https://openreview.net/forum?id=SSKZPJCt7B.

Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. Parallel bayesian optimization of multiple noisy objectives with expected hypervolume improvement. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Erik Drysdale. A parametric distribution for exact post-selection inference with data carving, 2023. URL https://arxiv.org/abs/2305.12581.

Amirmohammad Farzaneh and Osvaldo Simeone. Multi-objective hyperparameter selection via hypothesis testing on reliability graphs. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1324.

Matthias Feurer, Katharina Eggensperger, Edward Bergman, Florian Pfisterer, Bernd Bischl, and Frank Hutter. Mind the gap: Measuring generalization performance across multiple objectives. In *Advances in Intelligent Data Analysis XXI*, pp. 130–142. Springer, 2023.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection, 2014. URL https://arxiv.org/abs/1410.2597.

Greg N. Frederickson and Susan Rodger. A new approach to the dynamic maintenance of maximal points in a plane. *Discrete & Computational Geometry*, 5(4):365–374, 1990. doi: 10.1007/BF02187797.

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

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Zhiyong Huang, Hua Lu, Beng Chin Ooi, and Anthony K. H. Tung. Continuous skyline queries for moving objects. *IEEE Transactions on Knowledge and Data Engineering*, 18(12):1645–1658, 2006. doi: 10.1109/TKDE.2006.185.

Danijel Kivaranovic and Hannes Leeb. On the length of post-model-selection confidence intervals conditional on polyhedral constraints. *Journal of the American Statistical Association*, 116(534): 845–857, 2021. doi: 10.1080/01621459.2020.1732989.

Danijel Kivaranovic and Hannes Leeb. A (tight) upper bound for the length of confidence intervals with conditional coverage, 2024. URL https://arxiv.org/abs/2007.12448v3. Version 3.

Cyrille Kone, Emilie Kaufmann, and Laura Richert. Bandit pareto set identification: The fixed budget setting. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2548–2556. PMLR, 2024.

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

Bracha Laufer-Goldshtein, Adam Fisch, Regina Barzilay, and Tommi S. Jaakkola. Efficiently controlling multiple risks with pareto testing. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=cyg2YXn_BqF.

Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016. doi: 10.1214/15-AOS1371.

Sifan Liu and Snigdha Panigrahi. Flexible selective inference with flow-based transport maps, 2025. URL https://arxiv.org/abs/2506.01150.

Sifan Liu, Jelena Markovic-Voronov, and Jonathan Taylor. Black-box selective inference via bootstrapping, 2022. URL https://arxiv.org/abs/2203.14504.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL https://www.learningtheory.org/colt2009/papers/012.pdf.

Sérgio Moro, Paulo Rita, and Paulo Cortez. Bank marketing, 2014. URL https://doi.org/10.24432/C5K306.

Andreas Petrou-Zeniou and Azeem M. Shaikh. Inference after ranking with applications to economic mobility, 2026. URL https://arxiv.org/abs/2410.19212.

Martin Raič. A multivariate central limit theorem for lipschitz and smooth test functions, 2019. URL https://arxiv.org/abs/1812.08268v2. Version 2.

Stephen Reid, Jonathan Taylor, and Robert Tibshirani. Post-selection point and interval estimation of signal sizes in gaussian samples. *The Canadian Journal of Statistics*, 45(2):128–148, 2017. doi: 10.1002/cjs.11320.

Pascal Rink and Werner Brannath. Post-selection confidence bounds for prediction performance. *Machine Learning*, 114(3), 2025. doi: 10.1007/s10994-024-06632-w.

Xiaoying Tian and Jonathan E. Taylor. Selective inference with a randomized response. *The Annals of Statistics*, 46(2):679–710, 2018. doi: 10.1214/17-AOS1564.

Ben Tu, Nikolas Kantas, Robert M. Lee, and Behrang Shafei. Random pareto front surfaces, 2024. URL https://arxiv.org/abs/2405.01404.

Tijana Zrnic and William Fithian. Locally simultaneous inference. *The Annals of Statistics*, 52(3): 1227–1253, 2024. doi: 10.1214/24-AOS2391.

Tijana Zrnic and William Fithian. A flexible defense against the winner's curse. *The Annals of Statistics*, 53(6):2516–2535, 2025. doi: 10.1214/25-AOS2553.

A PROOFS

This appendix records the measure-theoretic qualification at comparison ties, the complete proof of Theorem 1, and the report-level corollaries. Throughout, intervals may be open, closed, half-open, unbounded, or empty. Two supports that differ only at finitely many points define the same conditional Gaussian law.

A.1 GAUSSIAN LINE DECOMPOSITION

Let $\mu_X = \mathbb{E}[X]$, $U = \eta^\top T$, $\nu = \mathbb{E}[U]$, and $\sigma^2 = \mathrm{Var}(U) > 0$. With

$$\rho = \mathrm{Cov}(X, U)/\sigma^2, \qquad Z = X - \rho U,$$

we have

$$\mathrm{Cov}(Z, U) = \mathrm{Cov}(X, U) - \rho\,\mathrm{Var}(U) = 0.$$

The vector (Z, U) is jointly Gaussian, so $Z \perp U$. This implication does not require Σ_X to be invertible. Conditional on $Z = z$,

$$U \sim \mathcal{N}(\nu, \sigma^2), \qquad X = z + \rho U.$$

In particular, the argument remains valid when $S = T$ and the stacked covariance of (S, T) is singular.

A.2 PROOF OF THE COMPLETE-FRONT SLICE THEOREM

Fix z. For an ordered pair (j, i) define

$$\delta_{jic}(u) = S_{j,c}(z + \rho u) - S_{i,c}(z + \rho u) = a_{jic}u + b_{jic}.$$

The weak comparison set $H_{jic} = \{u : \delta_{jic}(u) \geq 0\}$ is a half-line, $\mathbb{R}$, or $\varnothing$. Therefore

$$H_{ji} = \bigcap_{c=1}^{d} H_{jic}$$

is an interval or is empty. If at least one δ_{jic} is not identically zero, then on the interior of H_{ji} at least one coordinate comparison is strict, except possibly at finitely many roots. If every δ_{jic} is identically zero, the candidates are tied throughout the fiber and neither dominates the other. Thus the exact dominance event $D_{j\to i}$ agrees with an empty set or one positive-length interval after discarding finitely many points; it may also contain a singleton as discussed below.

For each fixed u, dominance is a finite strict partial order. Any nonmaximal label has a maximal strict dominator: repeatedly follow a dominator; finiteness and irreflexive transitivity prevent a cycle, so the chain ends at a maximal label. Hence if the full front is P, its members form an antichain and every omitted label is strictly dominated by a member of P.

Conversely, suppose these two conditions hold. All omitted labels are dominated. If a reported q had an omitted dominator i, choose $p \in P$ dominating i. Transitivity gives $p \succ q$, contradicting the antichain condition (or irreflexivity if $p = q$). No reported label has a reported dominator either. Thus the full front is exactly P, including at ties.

Define the exact reported-source incoming region

$$R_i^P = \bigcup_{p \in P \setminus \{i\}} D_{p\to i}.$$

There are $s(s-1)$ required internal directed relations and $s(K-s)$ reported-to-omitted relations, totaling $s(K-1)$. It follows pointwise that

$$\widehat{\mathcal{P}}(S(z + \rho u)) = P \quad \Longleftrightarrow \quad u \in \bigcap_{i \in P} (R_i^P)^c \cap \bigcap_{i \notin P} R_i^P. \tag{A.1}$$

Replacing each exact dominance event by its positive-length interval representative changes this identity only at finitely many points. A singleton dominance event need not be a complete tie:

differences $(u, -u, 1)$ are feasible only at zero with one strict coordinate. Such points are discarded solely because the inferential law is non-atomic, not because dominance is impossible. Persistent complete ties are instead handled on the entire line.

Every finite endpoint of a nonempty $D_{j\to i}$ is attained by an active affine coordinate constraint, and is therefore a root of one of the unordered-pair comparisons incident to P. There are $\binom{s}{2} + s(K-s)$ such pairs and at most two endpoints per required directed relation. Consequently, after separately merging the at most s intervals in each incoming union, the number $A_P = A_P(z)$ of distinct finite endpoints satisfies

$$A_P \le \min\left\{\left[\binom{s}{2} + s(K-s)\right] d,\ 2s(K-1)\right\}.$$

The interval-represented event is constant on each of the A_P+1 open cells cut out by these endpoints. A binary labeling of A_P+1 ordered cells contains at most $\lceil (A_P+1)/2 \rceil$ accepted runs after adjacent accepted cells are merged. Thus, modulo finitely many points, $\mathcal{E}_P(z)$ has at most $\lceil (A_P(z)+1)/2 \rceil$ interval components.

At an endpoint, several weak comparisons may be equal simultaneously. The weak-all/strict-one definition can then produce a front different from both neighboring open cells. The set of such endpoints is finite, however, and $U \mid Z = z$ has a continuous density. Their inclusion therefore has no effect on conditional probabilities or on the pivot.

Finally, conditional on $Z = z$, the event $\widehat{\mathcal{P}} = P$ is the restriction $U \in \mathcal{E}_P(z)$ up to that null set. Whenever its Gaussian mass is positive, the conditional density is

$$f_\nu(u \mid z, P) = \frac{\phi_\sigma(u-\nu)\mathbf{1}\{u \in \mathcal{E}_P(z)\}}{\int_{\mathcal{E}_P(z)} \phi_\sigma(v-\nu)\,dv}.$$

Writing $\mathcal{E} = \mathcal{E}_P(z) = \bigcup_r (\ell_r, h_r)$ and letting ϕ_σ denote the $\mathcal{N}(0,\sigma^2)$ density, the conditional CDF is

$$F^{\mathcal{E}}_{\nu,\sigma}(u) = \frac{\int_{\mathcal{E}\cap(-\infty,u]} \phi_\sigma(v-\nu)\,dv}{\int_{\mathcal{E}} \phi_\sigma(v-\nu)\,dv}.$$

Both integrals sum normal CDF differences over every support component. The probability integral transform makes $F^{\mathcal{E}_P(z)}_{\nu,\sigma}(U)$ uniform given $(Z = z, \widehat{\mathcal{P}} = P)$. For $\nu_2 > \nu_1$, the likelihood ratio on the common support is proportional to $\exp\{(\nu_2-\nu_1)u/\sigma^2\}$, so the family has a monotone likelihood ratio. Equal-tail inversion is therefore a one-dimensional root-finding problem and has exact coverage after averaging over Z. This proves the reconstruction and conditional-law parts of Theorem 1.

A.3 Output lower bound in Theorem 1

Fix $K > s \ge 2$, put $m = K - s$, $L = 4m$ and $D = L + 1$. The reported candidates and omitted candidates are respectively

$$a_p(u) = (pD + L, -pD), \quad 0 \le p < s; \qquad b_i(u) = (u - 2i, -u + 2i), \quad 0 \le i < m.$$

The a_p form an antichain. Their coordinate sum is $L > 0$, whereas the sum for every b_i is zero, so an omitted candidate cannot dominate a reported one. Moreover,

$$a_p \succ b_i \quad \Longleftrightarrow \quad u \in [pD + 2i, pD + L + 2i].$$

At either endpoint one comparison remains strict because $L > 0$. The complete report is precisely $P = \{a_0, \dots, a_{s-1}\}$ when all omitted candidates have a reported dominator. Intersecting their incoming unions gives

$$\mathcal{E}_P = [2m - 2, (s-1)D + L] \setminus \bigcup_{p=0}^{s-2} \bigcup_{i=0}^{m-1} (pD + L + 2i, (p+1)D + 2i).$$

Each removed gap has length one. Within a fixed p block the gaps are separated by length one; successive blocks and the two outer boundaries are separated by positive lengths $2m + 2$. Thus all $(s-1)m$ gaps are disjoint and strictly inside the common envelope, leaving exactly $(s-1)m + 1$ positive-length components.

Writing every component requires $\Omega((s-1)(K-s)+1)$ operations. For $2 \le s \le K/2$, $(s-1)(K-s) \ge Ks/4$. Padding with equal coordinates extends the construction to any $d \ge 2$. Together with Proposition 1, this proves that for fixed $d \ge 2$ the sparse-report bound is optimal up to a logarithmic factor among algorithms that explicitly output the interval union. The lower bound concerns this output representation, rather than the exact number of comparisons, an implicit representation, or growing-dimensional optimality.

Affinely map the common envelope, of length $W = (s-1)(4m+1) + 2m + 2$, onto $[-1, 1]$. The removed gaps then have total length $2m(s-1)/W < 1/2$, giving standard-normal support probability greater than $2\Phi(1) - 1 - (2\sqrt{2\pi})^{-1} > 0.48$ without changing the component count. This is the mass of the specified nuisance-conditioned slice, rather than a marginal report-frequency bound.

The rescaled example is also a Gaussian coordinate-target line. Add the common translation $(u - L, u)$ to every point. The first coordinate of a_0 then equals u, while all dominance relations stay unchanged. Write the stacked line as $a + bu$, with target index j satisfying $a_j = 0, b_j = 1$. For any $\epsilon > 0$,

$$V = bb^\top + \epsilon \operatorname{diag}(\mathbf{1}\{r \ne j\})$$

is positive definite: its quadratic form is $(b^\top x)^2 + \epsilon \sum_{r \ne j} x_r^2$, which vanishes only at zero. Also $Ve_j/V_{jj} = b$, as required by the Gaussian line decomposition. Choosing Gaussian mean a gives target mean zero and variance one. Thus the lower bound applies to genuine reported-coordinate inference lines as well as to abstract affine input arrays.

A.4 Proof of Proposition 1

Algorithm 1 Report-sensitive complete-report sweep

Require: affine selection line $S(u)$; target mask P
1: **for** $i \in [K]$ **do**
2: **for** $p \in P \setminus \{i\}$ **do**
3: stream the d comparison half-lines to obtain $D_{p \to i}$
4: **end for**
5: merge R_i^P; emit its labeled endpoints
6: **end for**
7: sort/group endpoints; initialize $\mathbf{1}\{u \in R_i^P\}$ at $u = -\infty$
8: maintain the number of labels whose selected state differs from P
9: sweep endpoint groups; output exactly the cells with zero mismatches
10: **return** merged output intervals (endpoint membership is immaterial)

Stream the d coordinate differences for each of the $s(K-1)$ required relations, maintaining two endpoints with constant scratch space. This costs $O(Ksd)$ arithmetic/comparison operations. For one candidate, sorting and merging at most s intervals costs $O\{s \log(s+1)\}$; over all candidates this is $O\{Ks \log(s+1)\}$. Store each merged component as two labeled endpoint events. At the leftmost cell, compute whether each candidate lies in its incoming union and initialize the number of candidate labels that disagree with P. Crossing a grouped endpoint toggles the state of only its attached candidate; the mismatch count is updated in constant time per incident event. Group simultaneous endpoints before evaluating the next open cell; intermediate states within a group do not represent actual cells. The sort and sweep cost $O\{Ks \log(Ks+1)\}$ and $O(Ks)$, with $O(Ks)$ stored intervals, events, and label states. The supplied affine arrays and any input conversion copies are excluded from this workspace claim; no length-d temporary difference vector is required. This proves the stated bounds in the exact-real operation model, not a bit-complexity or rounding-error bound.

For comparison, all-pair compression costs $O(K^2 d + K^2 \log K)$. The comparison-root sweep enumerates at most $\binom{K}{2} d$ affine coordinate crossings and costs $O(K^2 d \log(Kd))$ for the same support.

For $K = s = 1$ the support is the whole line and there are no relations. For $s = K$ the count equals the all-pair count, so there is no asymptotic improvement. Empty P is impossible for a nonempty finite strict order. Impossible or singleton-only reports have empty positive-measure support on

the supplied line. The reduction does not apply to nontransitive affine rules: with only $p \succ i$ and $i \succ q$, the putative report $\{p, q\}$ passes the reported-witness conditions but is not the actual maximal set $\{p\}$. All-pair interval logic remains available when pairwise interval geometry holds without transitivity. Prior skyline methods use dominance witnesses for query maintenance (Huang et al., 2006). Theorem 1 characterizes the complete selective support along an inferential line and establishes report-sensitive output complexity.

A.5 Report-level corollaries

Corollary 1 (Report-level guarantees)**.** *For every nonempty P, suppose the conditional miss probability of C_q is at most $\alpha_q(P)$ given $\widehat{\mathcal{P}} = P$. Then*

$$\mathbb{E}[V_P/m_P \mid \widehat{\mathcal{P}} = P] \le \frac{1}{m_P} \sum_{q \in Q(P)} \alpha_q(P),$$
$$\mathbb{P}(\theta_q \in C_q \; \forall q \in Q(P) \mid \widehat{\mathcal{P}} = P) \ge 1 - \sum_{q \in Q(P)} \alpha_q(P).$$

Setting $\alpha_q = \alpha$ gives FCR at most α after averaging over P; exact equal-tail intervals attain equality, apart from numerical error. Setting $\alpha_q = \alpha/(|P|d)$ gives simultaneous coverage at least $1 - \alpha$.

For a realized nonempty P, write $m_P = |P|d$ and $V_P = \sum_{q \in Q(P)} \mathbf{1}\{\theta_q \notin C_q\}$. Under the premise of Corollary 1,

$$\mathbb{E}\left[\frac{V_P}{m_P} \,\middle|\, \widehat{\mathcal{P}} = P\right] = \frac{1}{m_P} \sum_{q \in Q(P)} \mathbb{P}\{\theta_q \notin C_q \mid \widehat{\mathcal{P}} = P\} \le \frac{1}{m_P} \sum_{q \in Q(P)} \alpha_q(P).$$

Setting $\alpha_q(P) = \alpha$ and averaging over P proves FCR control. When every conditional miss probability equals α, so does the conditional average.

For the simultaneous statement of Corollary 1, with $\sum_q \alpha_q(P) \le \alpha$, conditional union bounding gives

$$\mathbb{P}\{\exists q \in Q(P) : \theta_q \notin C_q \mid \widehat{\mathcal{P}} = P\} \le \sum_{q \in Q(P)} \alpha_q(P) \le \alpha.$$

For a fixed reference r, $\mathrm{HV}_r(\theta_P)$ is the volume of the union of boxes from r to the selected labels' population values; a box is empty when any coordinate is below r. Increasing a coordinate can only enlarge this union. Thus simultaneous bounds imply $\mathrm{HV}_r(L_P) \le \mathrm{HV}_r(\theta_P) \le \mathrm{HV}_r(U_P)$ on the same coverage event. This can be wide and is not a direct selective hypervolume pivot. The band certifies a fixed safety vector t when $L_{k,c} \ge t_c$ for all c, and dominance of k over j when $L_{k,c} \ge U_{j,c}$ for every c with one strict inequality. Prespecified contrast intervals can be less conservative. The same simultaneous event implies $\sum_c w_c L_{k,c} \le \sum_c w_c \theta_{k,c} \le \sum_c w_c U_{k,c}$ for every $w_c \ge 0$, simply by multiplying the coordinate inequalities and adding them. Zero-weighted terms are omitted when endpoints are unbounded. This implication holds for all such weights on the original coverage event, including weights inspected after seeing the band. It is not an additional guarantee conditional on a later weight-selection event, and does not license adaptive pivot directions. Neither certificate identifies the population Pareto set without protecting comparisons to every relevant omitted label.

B Conditional transfer under Gaussian approximation

Many evaluation estimates have the expansion

$$\widehat{\theta} - \theta = \frac{1}{n} \sum_{r=1}^{n} \psi(O_r) + r_n, \qquad \sqrt{n}\, r_n = o_{\mathbb{P}}(1).$$

Here $\psi(O_r)$ is observation r's contribution and the remainder is negligible at the $n^{-1/2}$ scale. Fixed-dimensional Gaussian approximation before selection does not by itself establish approximation after conditioning.

Proposition 2 (Conditional transfer under non-rare report selection)**.** *Fix finite K, d, report P, and target loading q. Use common coordinates for the statistics and selection rule; any normalization must transform the selection event accordingly. Suppose $(X_n, \widehat{\Sigma}_n, \nu_n) \Rightarrow (X, \Sigma, \nu)$, where $X \sim \mathcal{N}(\mu, \Sigma)$, Σ is constant, $\nu = q^\top \mu$, and $q^\top \Sigma q > 0$. Let $A_P = \{\widehat{\mathcal{P}}(S(X)) = P\}$ have positive probability. Assume the plug-in covariance is positive semidefinite and its target variance and support mass are positive with conditional probability tending to one given report P. Assume the report indicator is almost surely continuous at X and the exact Gaussian pivot map is almost surely continuous at (X, Σ, ν) on A_P. Then the plug-in pivot G_n, evaluated at its true target ν_n, satisfies*

$$\mathbb{P}(G_n \le t \mid \widehat{\mathcal{P}}(S(X_n)) = P) \longrightarrow t, \qquad t \in [0, 1].$$

Equal-tail intervals have conditional coverage $1 - \alpha + o(1)$.

B.1 Proof of Proposition 2

Write $A_{n,P} = \{\widehat{\mathcal{P}}(S(X_n)) = P\}$ and $p_n = \mathbb{P}(A_{n,P})$. The statistics in the proposition are in common, fixed-dimensional coordinates; any normalization must also be reflected in the selection rule. The statement does not follow just by centering arbitrary fixed mean configurations without transforming their selection event.

Define the exact pivot map $g(x, H, m)$ by the formula in Section 3, with loading q, variance $q^\top H q$, and report P. It is defined on the selected event when the variance and support mass are positive. Set the gated map to zero off that event. Assume $\widehat{\Sigma}_n$ is positive semidefinite and the plug-in variance and mass are positive with probability tending to one on $A_{n,P}$; on exceptional undefined cases any fixed pivot convention may be used.

Use a common representation on which $(X_n, \widehat{\Sigma}_n, \nu_n) \to (X, \Sigma, \nu)$ almost surely. At almost every limit point the report indicator is continuous, so $\mathbf{1}\{A_{n,P}\} \to \mathbf{1}\{A_P\}$. On A_P, continuity of g gives $G_n \to G = g(X, \Sigma, \nu)$; off A_P, the gated maps eventually equal zero. Consequently the indicator and gated pivot converge jointly. In particular $p_n \to p = \mathbb{P}(A_P) > 0$. Theorem 1 gives $G \mid A_P \sim \mathrm{Unif}(0, 1)$, after integrating over the nuisance statistic. There are no conditional atoms, including at 0 and 1, so

$$\mathbb{P}(G_n \le t \mid A_{n,P}) = \frac{\mathbb{P}(\{G_n \le t\} \cap A_{n,P})}{p_n} \longrightarrow \frac{tp}{p} = t, \qquad 0 \le t \le 1.$$

The equal-tail coverage statement follows by taking the probability of $\alpha/2 \le G_n \le 1 - \alpha/2$. □

B.2 Sufficient continuity conditions and exclusions

The continuity assumption is substantive. One sufficient condition is that almost surely on A_P: the target variance is positive, the support has positive Gaussian mass, and no pair-coordinate comparison on the limiting selective line is identically zero. Also require the observed report indicator to be almost surely continuous (for example, no observed coordinate ties). The covariance itself need not be invertible.

To see the pivot continuity under these conditions, consider a convergent sequence (x_n, H_n, m_n). Positive limiting target variance implies convergence of the affine line coefficients and of its normal parameters. Each limiting comparison is a nonzero affine function, so it vanishes at only finitely many u, or never vanishes. Outside this finite collection, all comparison signs, and hence the complete-report indicators along the lines, converge pointwise. Normal densities with means and positive variances converging to finite limits have a common integrable envelope. Dominated convergence therefore applies to the support-mass denominator and the CDF numerator, including its moving upper limit. Their ratio converges because the limiting denominator is positive. This argument does not require the number of interval components to stay fixed.

Zero slopes with nonzero intercepts are allowed. Persistent zero comparisons require a separate stability argument, for example because the permitted perturbations preserve the equality exactly. Singular $S = T$ covariance is allowed when these conditions hold; singularity alone is not a continuity proof. Estimated covariance consistency does not ensure persistent-tie stability. The exact finite-sample theorem handles such ties, but the asymptotic transfer criterion is narrower.

B.3 WHY RARE REPORTS AMPLIFY APPROXIMATION ERROR

The amplification is a conditional-probability calculation, not a new CLT rate. For $p_n > 0$, put

$$e_n = \sup_{t\in[0,1]} |\mathbb{P}(\{G_n \le t\} \cap A_{n,P}) - tp|\,, \qquad d_n = |p_n - p|.$$

Then

$$\sup_{t\in[0,1]} |\mathbb{P}(G_n \le t \mid A_{n,P}) - t| \le \frac{e_n + d_n}{p_n}.$$

Indeed, subtract tp_n in the numerator and apply the triangle inequality. Thus a bound $e_n, d_n \le \delta_n$ would give $2\delta_n/p_n$. A marginal CLT supplies neither such a quantitative joint-event bound nor a uniform approximation over vanishing-probability reports. The proposition only treats fixed K, d and report probabilities with a positive limit; cross-fitted evaluations still require covariance consistency, negligible remainders, and the stated continuity conditions. For the scalar-integrated pivot, Theorem 2 and Appendix O supply explicit mean, covariance and report-boundary error terms under bounded-record assumptions.

C INTEGRATED SAME-REPORT CARVING

Under the nested Gaussian model of Section 4.1, write $\rho = V\eta/\sigma^2$, $Y = \eta^\top S$, $Z_S = S - \rho Y$, and $\mathcal{A}_P(Z_S) = \bigcup_j (a_j, b_j)$ for the full-report support.

Proposition 3 (Same-report scalar carving)**.** *For known V, $0 < f < 1$, and $\sigma^2 = \eta^\top V\eta > 0$, Z_S is independent of (U, Y) and $Y \mid U = u \sim \mathcal{N}(u, \tau^2)$, where $\tau^2 = (1-f)\sigma^2/f$. Conditional on Z_S and the entire front P,*

$$f_\nu(u \mid Z_S, P) \propto \phi_\sigma(u - \nu) \sum_j \left[\Phi\left(\frac{b_j - u}{\tau}\right) - \Phi\left(\frac{a_j - u}{\tau}\right)\right].$$

The resulting conditional CDF is an exact pivot under the Gaussian model. Its intervals inherit the FCR and simultaneous guarantees above.

C.1 PROOF OF PROPOSITION 3

Let $A \sim \mathcal{N}(\theta, V/f)$ and $B \sim \mathcal{N}(\theta, V/(1-f))$ independently, and put $S = A, T = fA + (1-f)B$. Then $\mathrm{Var}(S) = V/f$, $\mathrm{Var}(T) = V$, and $\mathrm{Cov}(S, T) = V$. For $Y = \eta^\top S, U = \eta^\top T, \sigma^2 = \eta^\top V\eta > 0$, and $\rho = V\eta/\sigma^2$, we obtain

$$\mathrm{Cov}(S - \rho Y, Y) = V\eta/f - \rho\sigma^2/f = 0, \qquad \mathrm{Cov}(S - \rho Y, U) = V\eta - \rho\sigma^2 = 0.$$

Joint Gaussianity makes $Z_S = S - \rho Y$ independent of (U, Y), including when V is singular. The full-front event conditional on $Z_S = z$ is $Y \in \mathcal{A}_P(z) = \bigcup_j (a_j, b_j)$ modulo null endpoints. Since $Y \mid U = u \sim \mathcal{N}(u, (1-f)\sigma^2/f)$, integrating over this support and applying Bayes' rule proves the density in the proposition. Its normalizing constant is

$$D_\nu = \sum_j \left[\Phi\left(\frac{\sqrt{f}(b_j - \nu)}{\sigma}\right) - \Phi\left(\frac{\sqrt{f}(a_j - \nu)}{\sigma}\right)\right].$$

The conditional probability integral transform proves pivot uniformity. The density ratio for $\nu_2 > \nu_1$ is a positive constant times $\exp\{u(\nu_2 - \nu_1)/\sigma^2\}$, so the CDF decreases in ν and equal-tail inversion gives a confidence interval. Averaging over each target's Z_S conditional on the common report P gives the report-level corollaries. A known selection offset is included in the affine intercept; an independent random offset may be conditioned on. Data-dependent jitter and estimated covariance remain approximations in our applications. □

For deterministic numerical evaluation, write $Z = (Y - \nu)\sqrt{f}/\sigma$. The pivot at x is

$$\mathbb{E}\left[\Phi\left(\frac{x - \nu - \sigma\sqrt{f}Z}{\sigma\sqrt{1-f}}\right) \,\middle|\, Z \in \bigcup_j \left(\frac{\sqrt{f}(a_j - \nu)}{\sigma}, \frac{\sqrt{f}(b_j - \nu)}{\sigma}\right)\right].$$

We normalize interval masses in the log domain and integrate the truncated-normal quantile in each interval with Gauss–Legendre nodes. This represents the standard normal-plus-truncated-normal carving law (Drysdale, 2023). The implementation uses 128 nodes, audited against 256 nodes on prespecified reports, rather than treating finite quadrature as exact arithmetic.

C.2 Applying the scalar-carving length bound

The length bound in the main text follows directly from Kivaranovic & Leeb (2024, Theorem 3.1), which allows a finite union of intervals. Conditional on Z_S, write $Y = U + R$ with independent $R \sim \mathcal{N}(0, \tau^2)$ and $\tau^2 = (1-f)\sigma^2/f > 0$. Their bound is $2\sigma\Phi^{-1}(1-\alpha_q/2)\sqrt{1+\sigma^2/\tau^2}$. Substitution gives the stated factor $1/\sqrt{1-f}$. The independent held-out estimator has variance $\sigma^2/(1-f)$, giving exactly this Gaussian split width for the same target and budget. This comparison fixes $\widehat{\mathcal{P}}(S)$ and does not keep an arbitrary $\widehat{\mathcal{P}}(T)$ unchanged. Replacing variances by estimates may preserve the formal length inequality but does not supply exact coverage.

C.3 Higher-dimensional residual integration

Suppose a selection fraction f produces S, while T is the estimator on all observations. In the exact Gaussian sample-mean model,

$$S = T + \omega, \qquad \omega \perp T, \qquad \omega \sim \mathcal{N}\left(0, \frac{1-f}{f} V\right).$$

Let $U = \eta^\top T$, set $\rho_T = \mathrm{Cov}(T, U)/\mathrm{Var}(U)$, and condition only on $Z_T = T - \rho_T U = z_T$. Bayes' rule and independence of T and ω give

$$f_\nu(u \mid z_T, \widehat{\mathcal{P}}(S) = P) \propto \phi_\sigma(u-\nu)\mathbb{P}_\omega\{\widehat{\mathcal{P}}(z_T + \rho_T u + \omega) = P\}. \tag{A.2}$$

For a fixed residual draw $\omega^{(m)}$, Theorem 1 returns an interval union $\mathcal{E}_P^{(m)}$. Hence a common-randomness approximation to the normalizing constant in (A.2) is

$$\frac{1}{M}\sum_{m=1}^{M} \int_{\mathcal{E}_P^{(m)}} \phi_\sigma(u-\nu)\, du,$$

and the numerator of the selective CDF replaces each support by $\mathcal{E}_P^{(m)} \cap (-\infty, u_{\mathrm{obs}}]$. This evaluates Gaussian interval masses analytically and uses Monte Carlo only for the residual integral. Quasi-Monte Carlo or importance sampling can reduce variance, but a finite ratio estimate is not itself the exact pivot. We therefore treat draw-count stability, repeated integration seeds, rare-front diagnostics, and a predeclared conservative fallback as part of the method's empirical validity audit.

Conditioning on the full orthogonal component of $X = (S, T)$ instead fixes the realized residual and returns the hard-truncation law of Theorem 1. It is exact under the Gaussian model but is a different, more conditioned procedure than (A.2).

D Implementation and Numerical Validation

We use two independent support constructors. The first builds ordered-pair dominance intervals, merges incoming intervals candidate by candidate, and sweeps labeled endpoints. The second sorts every comparison root, updates all coordinate signs, reconstructs the dominance graph on each open cell, and then merges cells with the target front. A third small-instance checker expands the Boolean condition in (A.1) directly. Tests compare supports by symmetric difference modulo breakpoints and recompute the Pareto front at interior query points.

The test suite includes constant comparisons, persistent ties, simultaneous and near-coincident roots, empty and unbounded dominance intervals, singular $S = T$ covariances, and supports with extremely small Gaussian mass. Gaussian interval probabilities and pivot inversions are evaluated with log-domain tail masses. The reported randomized checks use fixed seeds and fail on any open-cell or interval discrepancy.

Tail evaluation uses scaled complementary-error-function ratios and `expm1` to avoid cancellation. Support unions are preserved throughout inversion; an unbracketed root raises an explicit exception. Quadrature resolutions, endpoint tolerances and failure handling are specified with each experiment. High-precision checks and implementation history are recorded separately in the companion numerical-validation record.

E Experimental Audit and Full Results

E.1 Exact-Gaussian grid

Table 5 gives all eight cells behind the ranges in Section 5. Every cell uses known covariance and 10,000 independent replications. FCR is the replication-wise false coverage proportion averaged over reports; coordinate coverage instead pools reported coordinates. The hypervolume target is the population value of the selected labeled points.

Numerical revalidation recomputes all 525,080 marginal and simultaneous intervals in the four cells with archived bracket flags, resolving all 11 flagged intervals. Independent 80/120-digit calculations check every affected target. Coordinate, simultaneous-report and hypervolume coverage indicators are unchanged; the maximum recomputed CDF residual is 8.89×10^{-15}. The trade-off/independent cell's all-coordinate mean widths are 28.698655 (marginal) and 46.124842 (simultaneous); its marginal median and p90 are 2.582808 and 12.432660. The companion numerical record retains the recomputed widths for all four cells. Finite observed intervals can still have infinite expected length (Kivaranovic & Leeb, 2021).

Table 5: Complete exact-Gaussian grid, 10,000 reports per cell. The final column counts unresolved interval failures after numerical revalidation. Coverage and FCR retain every report and selected coordinate.

means	covariance	full coord.	naive coord.	FCR	naive FCR	full simult.	naive simult.	HV	failures
trade-off	independent	.952	.945	.048	.065	.958	.646	.999	0
trade-off	equicorrelated	.951	.953	.049	.049	.956	.764	1.000	0
trade-off	block	.949	.943	.051	.076	.966	.702	.996	0
trade-off	near-singular	.950	.950	.050	.050	.991	.908	.998	0
all equal	independent	.951	.922	.049	.091	.953	.636	.997	0
all equal	equicorrelated	.948	.931	.052	.077	.955	.764	.998	0
all equal	block	.952	.907	.049	.125	.968	.720	.991	0
all equal	near-singular	.950	.948	.050	.052	.971	.922	.991	0

E.2 Scalar-carving covariance and fraction audit

Table 6 reports every prespecified fraction and expanded covariance cell. The Gaussian grid uses known covariance, $K = 8, d = 2$, ten independent seeds, and 100 reports per seed. Simultaneous coverage ranges from 0.945 to 0.990; the two point estimates below 0.950 are within Monte Carlo uncertainty. Width gains vary systematically with covariance and selection fraction.

Main runs use 128 quadrature nodes, checked against 256 on the first 20 reports at a tolerance of 0.005 standard errors. All 16 cells pass this numerical precision check, with zero failures. These widths pool coordinates in their prespecified score units. The unit-invariant checkpoint reanalysis appears with its original coverage results in Appendix H.

E.3 Learned-policy design and complete sample-size grid

Covariates are standardized over a fixed empirical pool. The synthetic pool contains 20,000 independent ten-dimensional standard-normal records. For IHDP and ACIC we use their covariates only, one-hot encoding the three categorical ACIC variables before standardization. Let $a = \tanh x_1$, $b = \tanh x_2$, and $c = \tanh x_3$. Our two-objective mean surfaces are

$$\mu_0(x) = (0.4a + 0.3c,\ 0.3b - 0.2a), \qquad \mu_1(x) = \mu_0(x) + \tau(x),$$

$$\tau(x) = (0.7 \sin x_1 + 0.4b + 0.4ac + 0.15,\ -0.8 \sin x_1 + 0.3c - 0.25b + 0.1).$$

Table 6: Full scalar-carving grid, 1,000 reports per row. FCR SE uses reports. Mean and p95 widths describe marginal intervals; simultaneous coverage uses separate conditional Bonferroni intervals.

covariance	f	coord. cov.	FCR (SE)	simult. cov.	mean width	p95	reduction
block	0.2	0.950	0.049 (0.004)	0.964	1.070	1.095	2.3%
block	0.4	0.948	0.052 (0.003)	0.966	1.188	1.260	6.1%
block	0.6	0.953	0.047 (0.003)	0.962	1.366	1.536	11.8%
block	0.8	0.954	0.045 (0.003)	0.962	1.698	2.128	22.5%
equicorrelated	0.2	0.953	0.047 (0.003)	0.957	1.046	1.090	4.5%
equicorrelated	0.4	0.951	0.049 (0.003)	0.963	1.126	1.244	11.0%
equicorrelated	0.6	0.954	0.045 (0.003)	0.964	1.239	1.490	20.1%
equicorrelated	0.8	0.951	0.049 (0.003)	0.965	1.438	2.004	34.4%
independent	0.2	0.949	0.052 (0.003)	0.955	1.061	1.093	3.2%
independent	0.4	0.950	0.049 (0.003)	0.949	1.166	1.257	7.8%
independent	0.6	0.950	0.051 (0.003)	0.945	1.321	1.524	14.8%
independent	0.8	0.954	0.046 (0.003)	0.953	1.602	2.096	26.9%
near-singular	0.2	0.952	0.048 (0.006)	0.990	0.991	1.049	9.6%
near-singular	0.4	0.951	0.050 (0.006)	0.988	0.995	1.101	21.3%
near-singular	0.6	0.950	0.050 (0.006)	0.988	0.998	1.127	35.6%
near-singular	0.8	0.947	0.053 (0.006)	0.986	1.002	1.092	54.3%

Treatment is Bernoulli$(1/2)$ and independent of x. Outcome noise is Gaussian with covariance $0.25\left(\begin{smallmatrix}1 & 0.3\\ 0.3 & 1\end{smallmatrix}\right)$. Training and evaluation independently resample this covariate population and draw fresh treatments and noise. Repeated covariate values do not reuse training outcomes.

On each independent training sample of size 4,000, two extra-tree regressors (64 trees, minimum leaf size 20) estimate the arm-specific two-objective means $\widehat{Q}_0, \widehat{Q}_1$. Eight weights $w \in \{0, 1/7, \ldots, 1\}$ give $\pi_w(x) = \mathbf{1}\{(w, 1-w)^\top(\widehat{Q}_1(x) - \widehat{Q}_0(x)) > 0\}$. Both the policies and regressions are frozen before evaluation. The vector score is

$$\widehat{Q}_{\pi_w(X)}(X) + 2\mathbf{1}\{A = \pi_w(X)\}\{Y - \widehat{Q}_A(X)\}.$$

Known randomized treatment makes its mean the policy value conditional on training, regardless of regression misspecification. Its finite-sample distribution is not Gaussian. We estimate its full covariance and apply the same nested-sample approximation as in CIFAR. Reference policy values average the known surfaces over the covariate pool; reference values never enter an interval constructor.

Table 7: Complete custom randomized-treatment simulation grid, 1,000 reports per row and $f = 1/2$. FCR SE uses ten independent training-library averages. HV is the mean projection-interval width with fixed reference $(-1, -1)$; HV coverage is 1.000 throughout.

covariate pool	n	coord. cov.	FCR (SE)	simult. cov.	split FCR	reduction	HV width
synthetic	300	0.953	0.047 (0.003)	0.976	0.051	21.5%	0.883
synthetic	600	0.953	0.048 (0.003)	0.979	0.047	23.3%	0.617
synthetic	1000	0.951	0.049 (0.003)	0.970	0.051	24.2%	0.472
synthetic	2000	0.949	0.051 (0.003)	0.971	0.044	25.4%	0.330
IHDP	300	0.951	0.048 (0.003)	0.974	0.054	20.6%	0.910
IHDP	600	0.954	0.046 (0.003)	0.973	0.048	22.8%	0.632
IHDP	1000	0.952	0.048 (0.002)	0.968	0.049	24.4%	0.483
IHDP	2000	0.955	0.044 (0.003)	0.975	0.051	25.5%	0.338
ACIC	300	0.945	0.054 (0.003)	0.972	0.056	20.2%	0.887
ACIC	600	0.950	0.050 (0.003)	0.972	0.049	23.3%	0.612
ACIC	1000	0.951	0.049 (0.003)	0.966	0.052	25.3%	0.464
ACIC	2000	0.949	0.050 (0.004)	0.969	0.052	26.7%	0.323

All scalar, hard, and split intervals retain every replication; there are no numerical failures or infinite widths. The causal quadrature audit doubles 128 nodes to 256 on the first three evaluation replications of the first training seed, at every sample size; the largest audited marginal-endpoint difference is 4.7×10^{-5} standard errors. The full results include report-wise false coverage, training-seed uncertainty, paired FCR differences, and width quantiles.

F Closest Report-Level Baselines and Decision Audit

F.1 Paired Gaussian comparison protocol

Table 8 uses $K = 8, d = 2$, $f = 0.6$, coordinate standard error 0.25, and the trade-off means from the Gaussian grid. Each covariance has 1,000 independent outer reports. All methods use the identical nested views $S = A$, $T = fA + (1 - f)B$ and display exactly $\widehat{\mathcal{P}}(S)$. The all-coordinate equicorrelation is 0.5, or 0.97 in the near-singular cell. Block covariance is the Kronecker product of candidate and objective equicorrelations 0.15 and 0.55. The common error budget is $\alpha = 0.05$.

Carving and splitting use $\alpha/(|P|d)$ coordinate errors and have complete-report conditional simultaneous guarantees. The global Bonferroni baseline instead allocates $\alpha/(Kd)$ to all coordinates before report selection. Gaussian max-T calibrates their maximum absolute standardized error. Global, LSI, and zoom intervals provide unconditional simultaneous protection. The procedures share the nominal error budget while targeting different reliability criteria: conditional for each complete report, or unconditional over the selected reports.

LSI implementations. We implement two localization regions within the construction of Zrnic & Fithian (2024). Set $\nu = 0.005$. For coordinate-box localization, choose simultaneous radii r for the selection errors at level $1 - \nu$. The plausible labels are the union of Pareto sets over the box $[S - 2r, S + 2r]$. Label i belongs to this union exactly when no lower-bound competitor strictly dominates its upper-bound vector. This follows by setting candidate i to its upper bound and every competitor to its lower bound.

Difference localization instead bounds every candidate-pair objective error by r_{ijc}. It removes label i if one competitor satisfies $S_{j,c} - S_{i,c} \geq 2r_{ijc}$ in all objectives, strictly in at least one. The surviving labels form a safe superset of the feasible-label union, not necessarily the exact union. This construction cancels common translations. On the good localization event, each version supplies a deterministic intermediate family J_0 satisfying $P \subseteq J_0 \subseteq J^+(S)$. Monotone simultaneous intervals for $J^+(S)$ at error $\alpha - \nu$ then protect P, except for localization failure. The unconditional failure probability is at most $(\alpha - \nu) + \nu$.

For simulated critical values, two independent banks of 2,047 Gaussian errors are redrawn for every report, independently of the data. One bank localizes; the other calibrates inference. At error a, the critical value is the $\lceil 2048(1 - a) \rceil$th order statistic of 2,047 maxima, or infinity if that index is 2,048. Exchangeable ranks give conservative finite-bank calibration. Conditional on the localization bank, J_0 is fixed and the inference bank remains independent. Thus the LSI argument averages over calibration randomness; it is not a guarantee conditional on a fixed bank. We do not reuse one bank for both stages or clip different-view intervals to the smaller of global and local widths. The box-Bonferroni variant uses normal union-bound critical values instead of simulated maxima.

Zoom implementation. For each objective we apply the symmetric-tail construction at error α/d, using the common known marginal standard deviation. We use step-down upper bounds on the radii and the all-coordinate outer projections of the same confidence region (Zrnic & Fithian, 2025, Section 5.3). Taking each projection's interval hull preserves inclusion. Union bounding across objectives protects every coordinate before restricting to P. This is not a comparison against a variance-adaptive zoom implementation; all marginal standard errors in this experiment are equal.

Both LSI localizations retain all eight plausible labels in these trade-off cells, so their max-T outputs coincide. This does not make their screening rules equivalent in general. Carving has no numerical failures or infinite intervals in this comparison. Every raw record stores labels, interval endpoints, widths, coverage indicators, and source hashes. Audits reconstruct all metrics from these records. Timing excludes shared Gaussian calibration costs and is not used to claim a runtime advantage.

Table 9 gives Monte Carlo uncertainty across outer reports. Width comparisons use the ratio of mean report widths minus one. Their delta-method standard errors use the paired report-level residuals, not independently sampled coordinates. These are descriptive precision intervals, not simultaneous adjustments for selecting favorable designs.

Table 8: All simultaneous-baseline results. Each covariance has 1,000 reports. Entries are report coverage / mean report width. Naive intervals have no simultaneous guarantee; all other methods use nominal report error 0.05. The numerical values do not imply equal conditional guarantees.

method	independent	equicorrelated	block	near-singular
naive marginal (diagnostic)	.657 / .980	.771 / .980	.747 / .980	.910 / .980
global Bonferroni	.971 / 1.478	.980 / 1.478	.980 / 1.478	.994 / 1.478
global max-T	.971 / 1.476	.975 / 1.426	.977 / 1.467	.954 / 1.125
matched split	.954 / 2.116	.968 / 2.152	.959 / 2.077	.990 / 2.308
LSI box, Bonferroni	.971 / 1.494	.981 / 1.494	.980 / 1.494	.995 / 1.494
LSI box, max-T	.971 / 1.492	.977 / 1.444	.980 / 1.484	.958 / 1.147
LSI differences, max-T	.971 / 1.492	.977 / 1.444	.980 / 1.484	.958 / 1.147
zoom projections	.974 / 1.923	.984 / 1.895	.983 / 1.868	.995 / 1.895
scalar carving	.952 / 1.791	.955 / 1.720	.963 / 1.815	.992 / 1.491

Table 9: 95% Monte Carlo intervals. Coverage uses Wilson intervals; relative-width intervals use paired delta-method standard errors. Positive width change means carving is wider.

covariance	carving coverage	width vs global max-T	width vs split
independent	[.937,.964]	[+20.82%,+21.93%]	$[-15.56\%,-15.11\%]$
equicorrelated	[.940,.966]	[+20.06%,+21.16%]	$[-20.33\%,-19.84\%]$
block	[.949,.973]	[+23.05%,+24.38%]	$[-12.90\%,-12.29\%]$
near-singular	[.984,.996]	[+32.31%,+32.81%]	$[-35.49\%,-35.32\%]$

F.2 Complete-label conditional diagnostic

A separate equal-means design uses $K = 2, d = 2$, independent coordinate errors, standard error 0.25, $f = 0.6$, and 2,000 independent reports. The three possible label sets are $\{1\}$, $\{2\}$, and $\{1, 2\}$, with probabilities $1/4$, $1/4$, and $1/2$. We group by these exact sets, not report size. Carving's simultaneous coverage is 0.954 among 524 reports of $\{1\}$, 0.955 among 462 reports of $\{2\}$, and 0.942 among 1,014 reports of $\{1, 2\}$. Their 95% Wilson Monte Carlo intervals are $[0.933, 0.969]$, $[0.932, 0.970]$, and $[0.926, 0.955]$. This broad diagnostic is complemented by the targeted report-conditional study in Appendix N.1.

F.3 Direct safety certification with matched one-sided targets

We separately test whether reported candidates can be certified to exceed fixed thresholds $h = (0.25, 0.25)$ in both objectives. The design uses the same trade-off means, $K = 8, d = 2$, $f = 0.6$, and four covariance families, but standard error 0.08. Each family has 1,000 fresh reports. Thresholds, methods, and budgets were fixed before these simulations.

For the unsafe null $H_k = \bigcup_c \{\theta_{kc} \le h_c\}$, let p_{kc} be an upper-tail component p-value and set $p_k = \max_c p_{kc}$. If any component null is true, $\Pr(p_k \le t) \le t$ under its corresponding conditioning guarantee. Holm testing of the label-level p-values therefore controls false certification without a separate objective-wise Bonferroni factor. Carving uses report-conditioned Gaussian scalar p-values; splitting uses independent held-out values. Global Holm protects all K labels before report restriction. LSI uses its plausible family at level $\alpha - \nu$. Larger-family Holm rejections restricted to a smaller family are contained in the smaller-family rejections, so the localization argument still applies. One-sided max-T methods calibrate signed maxima, not absolute maxima.

Carving certifies more true-safe labels than splitting in every cell. All carving computations complete. Doubling 128 quadrature nodes to 256 on the first ten reports per cell changes component p-values by at most 3.34×10^{-6} and changes no certificate. Zero entries represent zero observed errors in 1,000 reports. All label-level raw decisions and paired uncertainty estimates are retained. These results assess threshold certification under the known-Gaussian model.

Table 10: Synthetic strict-safety audit, 1,000 reports per covariance. Each entry is mean correctly certified labels / probability of any false certification.

method	independent	equicorrelated	block	near-singular
global label Holm	.770 / .002	.872 / .000	.822 / .001	1.114 / .004
global one-sided max-T	.580 / .001	.759 / .000	.657 / .000	1.699 / .007
LSI box Holm	.739 / .002	.830 / .000	.784 / .001	1.053 / .004
LSI difference Holm	.739 / .002	.830 / .000	.784 / .001	1.053 / .004
LSI box one-sided max-T	.559 / .001	.739 / .000	.635 / .000	1.632 / .007
LSI difference one-sided max-T	.559 / .001	.739 / .000	.635 / .000	1.632 / .007
matched split Holm	.173 / .005	.272 / .011	.244 / .004	.393 / .006
scalar carving Holm	.347 / .004	.623 / .005	.415 / .006	1.114 / .004

G Report-Sensitive Computation and Complete-Label Diagnostics

G.1 Prespecified computational protocol

We fix seed 2026090814 and construct antichain reports of known size: their first two coordinates vary in opposite directions, and each omitted candidate lies strictly below a designated reported witness. All affine directions come from a supplied positive-definite low-rank-plus-diagonal covariance. This construction isolates support-reconstruction runtime. Three independent constructed instances per setting share all inputs across the comparison-root, all-pair, and report-sensitive backends, whose order is randomized. All 318 target supports and 270 complete-report target intervals agree across backends within the stated numerical tolerances (2×10^{-10} for support and 2×10^{-8} for interval endpoints).

The whole-report calculation uses fraction $f = 0.6$, 128 quadrature nodes and per-target budget $0.05/(sd)$. Initial front computation, line preparation, support construction and interval inversion are timed separately. Covariance is supplied; its estimation, candidate training and score generation are not included. Report totals are medians of sums, so their components' separate medians need not sum exactly. The script runs single-process with requested BLAS and OpenMP thread counts one, on a shared CPU machine. The single-target grid includes $K = 32, 64, 128, 256$, $d = 2, 8$, and $s = 4, K$; the four whole-report settings appear in Table 32.

A streamed validator implements the stated auxiliary-space bound. Peak Python-tracked allocation is measured separately from timing, excludes supplied arrays, and is not process RSS. Fewer required relations do not guarantee fewer *merged* endpoints or smaller measured memory on every instance: removing an omitted dominator can prevent unions from merging. The theorem improves a worst-case bound.

G.2 All complete-label groups

Before running, we fixed $K = 4, d = 2$, $V = 0.2^2 I_8$, $f = 0.6$, seed 2026090816, and population vectors

$$\theta_0 = (.75, .15), \quad \theta_1 = (.45, .45), \quad \theta_2 = (.15, .75), \quad \theta_3 = (.25, .25).$$

The middle trade-off label wins neither population objective. All methods use the same independent $A \sim N(\theta, V/f)$ and $B \sim N(\theta, V/(1-f))$, report $\widehat{\mathcal{P}}(A)$, and estimate with $T = fA + (1-f)B$ except for matched splitting, which uses B. We run 20,000 independent draws and retain all 15 possible nonempty reports. In 10,311 draws the observed report contains a member winning neither observed coordinate.

Global max-T uses the exact independent-Gaussian maximum quantile over eight coordinates. Pairwise-difference LSI uses localization budget $\nu = .005$ and independent banks of 2,047 localization and inference draws regenerated for every report. Carving and splitting use report Bonferroni budgets. Carving uses 128 nodes; 264 target intervals in the first 50 reports are checked at 256 nodes, with maximum discrepancy 0.000408 full-estimator standard errors. There are 101,346 reported coordinate targets and no numerical failures. LSI's plausible set contains all four labels in every replication; this experiment does not exercise effective candidate reduction by localization.

Table 11: All complete-label groups: conditional simultaneous coverage. Parentheses give unadjusted 95% Wilson Monte Carlo intervals, not inferential intervals for model performance. $N/20000$ is each report's empirical frequency. The prespecified $N \geq 2000$ precision flag applies only to $\{0, 2\}$ and $\{0, 1, 2\}$. No group is removed or pooled by report size.

P	N	global max-T	LSI	split	carve
$\{0\}$	390	.936 (.907,.956)	.944 (.916,.963)	.962 (.937,.977)	.972 (.950,.984)
$\{1\}$	700	.981 (.969,.990)	.981 (.969,.990)	.959 (.941,.971)	.963 (.946,.975)
$\{2\}$	368	.948 (.921,.967)	.948 (.921,.967)	.973 (.951,.985)	.951 (.924,.969)
$\{3\}$	56	.964 (.879,.990)	.964 (.879,.990)	.964 (.879,.990)	.982 (.906,.997)
$\{0, 1\}$	1954	.974 (.966,.981)	.977 (.969,.983)	.947 (.936,.956)	.954 (.944,.962)
$\{0, 2\}$	3409	.976 (.970,.981)	.979 (.973,.983)	.947 (.939,.955)	.946 (.938,.953)
$\{0, 3\}$	318	.978 (.955,.989)	.978 (.955,.989)	.943 (.912,.964)	.953 (.924,.971)
$\{1, 2\}$	1919	.972 (.964,.978)	.976 (.968,.982)	.951 (.940,.960)	.946 (.935,.956)
$\{1, 3\}$	264	.943 (.908,.965)	.947 (.913,.968)	.928 (.890,.953)	.939 (.904,.962)
$\{2, 3\}$	311	.936 (.903,.958)	.936 (.903,.958)	.939 (.907,.961)	.929 (.895,.953)
$\{0, 1, 2\}$	5553	.977 (.973,.981)	.981 (.977,.984)	.949 (.943,.954)	.952 (.946,.957)
$\{0, 1, 3\}$	704	.959 (.942,.971)	.963 (.946,.975)	.935 (.914,.951)	.955 (.937,.968)
$\{0, 2, 3\}$	1513	.969 (.959,.977)	.972 (.963,.979)	.950 (.938,.960)	.953 (.941,.963)
$\{1, 2, 3\}$	665	.965 (.949,.977)	.970 (.954,.980)	.950 (.931,.964)	.956 (.938,.970)
$\{0, 1, 2, 3\}$	1876	.960 (.950,.968)	.966 (.957,.973)	.955 (.945,.964)	.951 (.941,.960)

Overall simultaneous coverage is .97055 (global), .97390 (LSI), .94940 (split), and .95110 (carve). Mean report widths are 1.09080, 1.10589, 1.61895, and 1.26642, respectively. The paired carve-minus-split width difference is $-.35252$ (Monte Carlo SE $.000512$), or 21.8% shorter by the ratio of mean widths. Naive marginal intervals have report coverage .77410 and FCR .05370. This broad diagnostic is complemented by the targeted conditional-separation design in Appendix N and the independent real-report confirmation in Appendix P.1.

H Checkpoint Evaluation: Calibration, Widths and Comparisons

H.1 Training, scores and paired inference

Four convolutional configurations use standard augmentation at widths 8 and 16, Gaussian-noise augmentation at width 12, and blur augmentation at width 12. Each is trained on a fixed 5,000-image CIFAR-10 training subset; epochs $1, 2, 4, 7, 10$ supply 20 checkpoints (Krizhevsky, 2009). The 10,000-image test pool defines the reference population. Its five scores are clean accuracy and true-class log probabilities under clean, noise, blur and darkening conditions. Common noise and deterministic blur/darkening preserve cross-model dependence. These are local transformations, separate from official CIFAR-10-C.

Each replication samples 5,000 paired records with replacement. The first 2,500 select the randomized report; all records form the carving estimate, and the remaining 2,500 form the splitting estimate. Both methods report $\widehat{\mathcal{P}}(S)$; the original offset changes 0.5% of fronts. For the nested means, $\mathrm{Cov}(S, T) = \mathrm{Var}(T) = V$, giving line coefficient $\mathrm{Cov}(S, U)/\mathrm{Var}(U) = V_{:,q}/V_{q,q}$ for coordinate q. Pool means are used to assess coverage, independently of interval construction.

H.2 Calibration studies

Table 12 reports the frozen 1,000-replication hard-conditioning confirmation. The scalar-carving study uses a separate 1,000 draws over ten seed groups, with hard intervals replayed on those same draws (Table 13). These two tables represent separate experiments; the width reanalysis below uses the scalar-carving draws again.

On the frozen hard-conditioning draws, naive intervals cover 741 coordinates missed by selective intervals, with the reverse outcome for 376 coordinates. Global Bonferroni and reference-covariance max-t simultaneous coverage is 0.985 and 0.970, respectively. Full-pool covariance leaves hard-conditioning coverage essentially unchanged. A moment-matched Gaussian diagnostic instead gives

Table 12: Frozen CIFAR hard-conditioning confirmation (1,000 reports). The scalar-carving study uses different draws; only its replayed hard row is a paired comparison. Simultaneous columns use a separate Bonferroni allocation.

method	coord. cov.	FCR	simult. cov.	mean/median width
hard, plug-in V	.939	.061	.967	.232/.064
hard, oracle V	.940	.061	.969	.233/.064
split, same report	.951	.049	.966	.093/.078
naive marginal	.948	.052	.470	.066/.055

coordinate coverage 0.955 and FCR 0.044, supporting the two-view calculation under the Gaussian model.

Table 13: Earlier CIFAR-10 checkpoint resampling (1,000 reports over ten seeds, $n = 5{,}000$, $f = 0.5$). Simultaneous coverage uses separate Bonferroni intervals. $\bar{R}$ averages paired marginal width ratios over coordinates within reports, then over reports; SE is report-based. Dash: not reanalyzed. Plug-in V is estimated from the resample; oracle V uses the full score pool.

method	coord. cov.	FCR	simult. cov.	$\bar{R}$ (SE)
hard, plug-in V	.944	.057	.971	—
scalar, plug-in V	.950	.051	.968	.759 (.002)
scalar, oracle V	.950	.051	.970	—
split, same report	.948	.052	.965	1 (reference)

In the paired scalar-carving study, FCR is 0.0508 (MC SE 0.0027), versus 0.0524 (0.0029) for splitting; the paired difference is -0.00164 (0.00257). Hard FCR exceeds carving's by 0.0065 (paired SE 0.0030), with larger mean width. Naive coordinate coverage is 0.950 but whole-report coverage is 0.461, illustrating the distinction between marginal and report-level protection. These are plug-in Gaussian results on non-Gaussian score records.

H.3 Matched-width reanalysis

This paired reanalysis uses the same 1,000 scalar-carving reports as Table 13, with unchanged sources, seeds and numerical settings. All 1,000 saved reports reproduce, retaining 38,905 target-level width pairs. The archive has shape $10{,}000 \times 20 \times 5$. For each report, average carving/splitting ratios over its selected coordinates; then average equally over reports. Objective-specific summaries use the same ordering, and MC SE is the report-level SD divided by $\sqrt{1000}$. The ten seed groups are administrative partitions of this one checkpoint library.

Table 14: Carving/splitting width ratios from the original CIFAR draws. Each entry is a report-averaged paired ratio (report-level Monte Carlo SE). Both methods use the same marginal or simultaneous error budget.

objective	marginal ratio (MC SE)	simultaneous ratio (MC SE)
clean accuracy	.75722 (.00146)	.75888 (.00143)
clean log probability	.75955 (.00165)	.76146 (.00161)
noise log probability	.75415 (.00174)	.75572 (.00170)
blur log probability	.75632 (.00180)	.75811 (.00177)
dark log probability	.76699 (.00167)	.76894 (.00163)
all objectives	.75885 (.00159)	.76062 (.00156)

The 95% MC intervals for the overall marginal and simultaneous ratios are $[.75573, .76196]$ and $[.75756, .76368]$. All targets are retained: 41 marginal and 39 simultaneous ratios exceed one, with maxima about 1.025; no failure or zero denominator occurs. Paired ratios cancel changes of measurement units. The table below also gives precision in each objective's native units.

Table 15: Pooled selected-coordinate mean widths in each objective's own units, 7,781 targets per objective. Accuracy uses probability units; log scores use nats. These means weight coordinates, unlike the report-averaged ratios above.

objective	carving	split	carving simult.	split simult.
clean accuracy	.02898	.03821	.04773	.06279
clean log probability	.06291	.08253	.10364	.13561
noise log probability	.13223	.17328	.21791	.28489
blur log probability	.07394	.09766	.12181	.16047
dark log probability	.05803	.07563	.09559	.12426

The original mixed-unit pooled means were .0712151 for carving and .0934620 for splitting. Their 23.803% reduction and the paired-ratio summaries answer different aggregation questions. Coverage for the same draws appears once in Table 13.

H.4 Prespecified checkpoint-difference decisions

This extension asks whether the intervals support actual comparisons of the frozen checkpoint models. It uses the same 10,000-image score pool as Table 13, with 1,000 fresh outer draws (seed 2026090819), $n = 5,000$, $f = 1/2$ and the original five objectives for selection. It independently resamples the existing frozen checkpoint score pool.

The reference is the final checkpoint of the widest standard-training model, chosen by that fixed rule before the new draws. For each selected nonreference candidate k, the four tests concern $\theta_{k,c} - \theta_{r,c}$: clean-accuracy noninferiority with margin -0.02, and improvements greater than 0.05 nats in noise, blur and darkening log probabilities. All margins and the reference are fixed in the protocol. The selected family contains $4|P \setminus \{r\}|$ tests even when $r \notin P$. Paired image-level covariances include dependence on the reference and among all objectives. Carving conditions on the observed tie-breaking offset and the complete report, not just its members individually.

Holm tests at family error level 0.05 use carved or held-out one-sided p-values for the selected family. Global Holm and global one-sided max-T are applied to all 76 prespecified candidate–reference contrasts under the plug-in Gaussian working model before restricting to the report. Max-T uses the estimated contrast covariance and 2,047 independent Gaussian calibration draws. Unadjusted full-sample tests are reported separately and have no general family-wise guarantee. True and false claims are assessed against means of the entire fixed pool.

Decision yield and errors. Mean true claims per report (MC SE) are 9.059 (0.03485) for carving plus Holm, 9.015 (0.03385) for splitting plus Holm, 9.023 (0.03373) for global Holm and 9.034 (0.03397) for global one-sided max-T. Carving's paired gains are 0.044, 0.036 and 0.025, with approximate 95% Monte Carlo intervals $[0.0238, 0.0642]$, $[0.0202, 0.0518]$ and $[0.0095, 0.0405]$. Reports tie in 913 comparisons against splitting and 949 against each global comparator.

Carving and both global comparators have zero false-claim reports among 1,000; splitting has one. Zero observations give a one-sided exact 95% binomial upper bound 0.002991 on the false-report probability in this resampling mechanism. Unadjusted tests produce 9.631 true claims on average and 13 false-claim reports. The reference occurs in every observed report; unit tests additionally cover reports where it is absent.

Widths are a separate output. The mean report-averaged ratio of marginal 95% contrast widths is 0.77064 (MC SE 0.00166). Coordinate-weighted contrast coverage is 0.94677 for carving and 0.94398 for splitting; report-averaged marginal miss rates are 0.05361 (SE 0.00260) and 0.05626 (SE 0.00249). These marginal intervals use fixed nominal levels, separately from the Holm-adjusted decisions. Nineteen individual ratios exceed one. No numerical failures or empty families occur. The first ten reports' 128-to-256-node audit changes no Holm decision, with maximum p-value difference 2.66×10^{-6}. All results remain empirical plug-in evidence on this finite score population.

I EVENT-MATCHED BLACK-BOX CONDITIONAL INFERENCE

This experiment compares direct calculation of the selective law with learning its selection weight through a parametric-bootstrap implementation of Liu et al. (2022). Each method receives the same complete report, target and nuisance statistic Z_S. The design uses $K = 4, d = 2$, full-estimator coordinate SD 0.25, $f = 0.6$, and trade-off means with independent or equicorrelated Gaussian errors. The target is the first objective of the lowest-index reported candidate, fixed by the complete label set. Each covariance has 1,000 paired reports.

Bootstrap event and fitting. Write $U = \eta^\top T$, $\sigma^2 = \eta^\top V \eta$, $\rho = V\eta/\sigma^2$, $Y = \eta^\top S$, and $Z_S = S - \rho Y$. In the nested Gaussian model, $\epsilon = Y - U$ is independent of $\widetilde{Z} = Z_S + \rho U$ and has variance $(1-f)\sigma^2/f$. Bootstrap bases are centered at the observed $\widetilde{Z}$, with covariance

$$\mathrm{Var}(\widetilde{Z}) = V/f - \{(1-f)/f\}\sigma^2 \rho\rho^\top.$$

Independent residuals ϵ^* give labels $\mathbf{1}\{\widehat{\mathcal{P}}(\widetilde{Z}^* + \rho\epsilon^*) = P\}$. Features are scaled by their marginal bootstrap standard deviations. The fit and inversion use bootstrap data only; exact probabilities enter the subsequent accuracy checks.

Training budgets of 2,048 and 8,192 use nested bootstrap examples. Both classifiers have two 32-unit ReLU layers, unweighted cross-entropy, Adam step size 10^{-3}, batch size 256 and a 200-epoch cap. The primary variant stops using classification accuracy on an internal 20% holdout. A paired variant selects checkpoints by validation cross-entropy with patience 20; all replayed fits use the same 32-unit architecture. The loss-stopped variant reuses the primary reports and changes the optimizer and internal-split streams. An external 1,024-draw set measures log loss and Brier score.

Numerical evaluation. The learned weight defines a weighted-normal exponential family on an initial standardized grid $[-12, 12]$ with spacing 0.01. The range expands until the Gaussian outside-mass bound divided by the inside normalizer is below 10^{-7} at the null and both interval endpoints. The first ten reports per covariance and budget use half the spacing and at least twice the initial range. Maximum endpoint changes are 7.80×10^{-5} for the primary fits and 1.13×10^{-4} for the loss-stopped replay. All fits complete. Additional held-out bootstrap checks use 2,048 proposals and at most 32 selected records per checked fit, each with its own nuisance and target center.

Target-wise results. Table 16 retains both fitting rules and all primary comparison methods. Widths are in full-estimator target-SD units. Increasing the primary bootstrap budget reduces mean absolute pivot error from 0.05935 to 0.03664 under independence and from 0.04243 to 0.02133 under equicorrelation. Validation log losses fall from 0.3096 to 0.2061 and from 0.3132 to 0.1779, respectively. The prespecified Monte Carlo screen requires a one-sided 95% binomial miss upper bound at most .07 and no numerical failure. In the equicorrelated cell, the exact method and the primary high-budget learner have upper bound .07822; the loss-stopped high-budget learner has .08148. Naive intervals exceed the screen in both cells. This screen measures empirical calibration at a fixed tolerance.

Table 16: Event-matched target-wise comparison, 1,000 reports per covariance. Entries give coverage / mean width in target-SD units. The loss-stopped variants are paired replays of the primary reports. Every method uses the same report event and target rule; this study evaluates one target per report.

Method	Independent	Equicorrelated
Exact scalar carving	.948 / 4.7692	.936 / 4.4069
Matched splitting	.951 / 6.1980	.947 / 6.1980
Naive	.940 / 3.9199	.934 / 3.9199
BB accuracy, 2,048	.966 / 4.4914	.946 / 4.2350
BB accuracy, 8,192	.952 / 4.6460	.936 / 4.3330
BB loss, 2,048	.960 / 4.6932	.945 / 4.3345
BB loss, 8,192	.946 / 4.6916	.933 / 4.3646

A separate exact-only confirmation uses 10,000 new reports per covariance and a seed frozen after the primary comparison. Exact coverage is .9526/.9528, with 95% binomial intervals

$[.94825, .95668]$ and $[.94846, .95687]$. Splitting covers at .9510/.9477 and naive intervals at .9504/.9506. The exact method's miss upper bounds are .05104/.05084, meeting the same screen. All methods complete. The primary comparison and this independent confirmation remain separate batches.

Complete-output computation. The whole-report study in Appendix K.8 applies both fitting rules and budgets to every prescribed coordinate and contrast. Table 30B reports common-scope costs together with endpoint, pivot and Holm-decision agreement. Direct support calculation supplies the reference conditional law without selection-probability fitting. The comparison measures attained numerical accuracy at the specified training budgets.

J Official-Model Evaluation with Independent Selection

J.1 Selection and inference protocol

The fixed 21-model score pool and its five objectives are described in Appendix L. Independently resample records with replacement, then assign prespecified portions to selection A and held-out inference B. For objective c, use its known range $[\ell_c, h_c]$, with $\ell = (0, -2, 0, 0, 0)$ and $h = (1, 0, 1, 1, 1)$:

$$S^{\xi}_{k,c} = \overline{X}_{A,k,c} + \xi_{k,c}, \qquad \xi_{k,c} = 10^{-3}\frac{h_c - \ell_c}{2\sqrt{|A|}} G_{k,c}, \quad G_{k,c} \overset{\text{iid}}{\sim} N(0,1).$$

The offset stream is independent of sampling and calibration. Selection therefore depends on A, fixed ranges and independent noise; held-out observations remain independent of the selected labels.

Condition on the realized offset and apply scalar carving along $S^{\xi}(y) = Z_S + \xi + \rho y$, using Proposition 3. The offset shifts the support; the conditional Gaussian covariance remains the nested sampling covariance. Hard and membership calculations retain the identical offset and labels. Application intervals use estimated covariance and a Gaussian approximation.

J.2 Frozen settings and numerical checks

Reporting uses seed 2026091051, ten seed groups of 100 reports per cell, $n \in \{500, 1000, 2500, 5000\}$ and $f \in \{.4, .6, .8\}$. The primary cell is $n = 5000$, $f = .6$. Fractions share sampled records within each n and replication; Monte Carlo summaries are computed within cells. The separate decision confirmation uses seed 2026091061 and 1,000 primary-setting reports. Its four fixed-reference contrasts test clean-accuracy differences above $-.02$ and noise/blur/brightness differences above .01 relative to `Standard`. Negative Brier contributes to selection.

All methods share the covariance estimator, variance ridge, 2,047-draw calibration banks and LSI allocation $\nu = .005$. Both reporting and decisions use 256-node quadrature. The first ten reports per cell receive 512-node checks at tolerance 0.002 full standard errors; decision checks also compare Holm vectors. The protocols were frozen before the confirmation runs.

Dependency tests vary B while fixing A and the offset: production workers retain identical selection views, hashes and labels. Offset checks cover coordinate and contrast targets. These checks belong to the 180-test regression suite. Post-run accounting verifies frozen source hashes, complete replication counts and numerical failures. Summaries retain every planned report.

J.3 Primary results and information cost

Table 4 reports the coordinate and independent decision confirmations in the main text.

All 1,000 primary reports contain 26,640 selected coordinates. Carving FCR is 0.050332, with pointwise 95% MC interval $[0.044107, 0.056556]$. Its paired difference from splitting is 0.003519, $[-0.003458, 0.010495]$. The corresponding width reduction is 33.61% for marginal intervals and 33.43% for simultaneous intervals. These application results are evaluated on the fixed score population under the plug-in Gaussian working model.

Carving minus global max-T has paired simultaneous width-ratio difference 0.002839, $[-0.000321, 0.005999]$; carving minus LSI max-T gives 0.016502, $[0.013362, 0.019643]$. Here the

held-out covariance is estimated separately; the theoretical length bound uses the same covariance for both estimators.

Hard marginal FCR is 0.061961, $[0.054014, 0.069908]$. Its coordinate width-ratio median and 99th percentile are 0.639388 and 35.735626; the report-mean ratio is 1.955495. These are different aggregation units. An independent audit reconstructs the nine largest marginal coordinate ratios and solves their 18 endpoints at 70 and 110 decimal digits. Summing Gaussian mass over the full support union gives maximum stored-endpoint CDF residual 4.71×10^{-16} and root error 1.32×10^{-14} in native units. The reported width distributions include all targets and tails.

J.4 From performance intervals to comparisons and summaries

The paired carving-minus-global max-T gain is 0.207 true comparisons per report (95% MC interval $[0.173944, 0.240056]$); 215 reports gain comparisons, 758 tie and 27 lose comparisons. The corresponding splitting gain is 0.684, $[0.633388, 0.734612]$. The small numbers of false reports leave relative error rates uncertain. The row audit reconstructs selection from the frozen sampling streams, independently recomputes nonselective contrast statistics, and verifies all Holm and max-T decisions given recorded selective p-values and calibrated cutoffs. A separate doubled-node audit checks numerical sensitivity on the prescribed first ten reports.

Table 17: Projection to the hypervolume of the selected labels' population values. Simultaneous boxes are intersected with fixed physical ranges, reference $(0, -2, 0, 0, 0)$ and upper bound $(1, 0, 1, 1, 1)$. Coverage intervals are Wilson intervals, including all-hit cases. Width parentheses are report-based MC SE.

method	covered	Wilson 95%	mean width
Carving	999/1000	$[0.994357, 0.999823]$	0.162252 (0.000445)
Splitting	1000/1000	$[0.996173, 1.000000]$	0.243250 (0.000153)

The mean paired hypervolume width ratio is 0.667028, $[0.663534, 0.670521]$. An independent inclusion–exclusion calculation on all 1,000 primary reports agrees with the production projection to 1.34×10^{-15}. Near-unit coverage reflects the conservatism of projecting simultaneous coordinate bounds onto the reported labels' hypervolume.

J.5 Complete grid, label groups and denominator checks

Across all six $n \geq 2500$ cells, carving FCR is 0.050–0.056 and simultaneous coverage is 0.954–0.961. At $n = 500$–1000, the respective ranges are 0.067–0.087 and 0.903–0.943. 6 of the six smaller-sample cells have FCR intervals above 0.05; 3 have Wilson upper simultaneous-coverage limits below 0.95. All checks are pointwise and descriptive, not a general sample-size cutoff.

Table 18: Complete-label groups observed at least 100 times under the prespecified descriptive display rule. All 14 groups remain in aggregate results and the archive. Labels are zero-based positions in the included 21-model pool, not original roster indices. C/S: carving/splitting simultaneous coverage. These intervals are not simultaneous across groups; rare and unseen reports remain unvalidated.

labels	reports	cov. C	Wilson C	cov. S
14,15,16,19,20	653	0.968	$[0.951, 0.979]$	0.974
14,15,16,17,19,20	160	0.950	$[0.904, 0.974]$	0.956
14,15,16,18,19,20	113	0.965	$[0.913, 0.986]$	0.991

At $n = 500, f = 0.8$, 56 of 29,695 selected coordinates in 32 reports have all-success held-out binary scores. The fixed 10^{-12} ridge then gives splitting widths near 3.92×10^{-6}. These coordinates contribute 97.44% of the mean carving width ratio 26.0839, but only 0.000985 to carving FCR. The denominator effect explains extreme width ratios, not most undercoverage. No ratio is trimmed and no variance floor is retuned.

The final audit retains all 12,000 reporting draws and 3,147,885 method-coordinate intervals, all 12 cells, all 1,000 separate decision draws and every model mean. Each selection view and offset is

reconstructed from its independent random stream and checked bitwise against recorded hashes. No interval failure, nonfinite width, or theoretical same-covariance length-bound violation is recorded. Numerical completion is distinct from statistical calibration.

J.6 All prespecified reporting cells

Tables 19–20 contain all prespecified cells, each with 1,000 reports. Parentheses denote report-level Monte Carlo SE. Fractions share sampled records, so cross-fraction comparisons are paired. All methods use the same complete report within a cell.

Table 19: Complete official-model grid, 1,000 reports per cell. FCR uses marginal intervals; C and S coverage columns use separate simultaneous carving and splitting intervals. $\bar{R}_C$ is the matched marginal width ratio. Large ratios with large SE arise from nearly zero held-out variance; they are retained, not treated as stable efficiency factors.

n	f	FCR C (SE)	FCR S (SE)	cov. C	cov. S	$\bar{R}_C$ (SE)
500	0.4	0.070 (0.004)	0.062 (0.003)	0.915	0.916	0.894 (0.002)
500	0.6	0.074 (0.004)	0.070 (0.004)	0.912	0.882	0.799 (0.004)
500	0.8	0.087 (0.004)	0.087 (0.004)	0.903	0.776	26.084 (5.682)
1000	0.4	0.067 (0.003)	0.060 (0.003)	0.943	0.933	0.862 (0.002)
1000	0.6	0.069 (0.003)	0.058 (0.003)	0.939	0.932	0.742 (0.003)
1000	0.8	0.071 (0.004)	0.069 (0.004)	0.940	0.870	0.797 (0.214)
2500	0.4	0.055 (0.003)	0.051 (0.003)	0.954	0.951	0.828 (0.002)
2500	0.6	0.054 (0.003)	0.054 (0.003)	0.957	0.947	0.697 (0.002)
2500	0.8	0.056 (0.003)	0.060 (0.004)	0.954	0.927	0.517 (0.003)
5000	0.4	0.051 (0.003)	0.053 (0.003)	0.959	0.966	0.803 (0.001)
5000	0.6	0.050 (0.003)	0.047 (0.003)	0.961	0.972	0.664 (0.002)
5000	0.8	0.051 (0.003)	0.053 (0.003)	0.961	0.959	0.479 (0.002)

Table 20: Global/local comparison across all official cells. G: global max-T; L: difference-localized LSI max-T; C: carving. All widths are simultaneous ratios to the identical-report splitting interval. Different theoretical conditioning targets are not equated.

n	f	cov. G	cov. L	$\bar{R}_C$	$\bar{R}_G$	$\bar{R}_L$
500	0.4	0.919	0.924	0.895	0.801	0.809
500	0.6	0.918	0.922	0.803	0.660	0.667
500	0.8	0.918	0.922	26.232	17.645	17.833
1000	0.4	0.958	0.961	0.863	0.803	0.812
1000	0.6	0.961	0.962	0.745	0.658	0.665
1000	0.8	0.960	0.960	0.808	0.694	0.699
2500	0.4	0.968	0.968	0.829	0.807	0.810
2500	0.6	0.966	0.963	0.700	0.661	0.657
2500	0.8	0.965	0.960	0.521	0.472	0.465
5000	0.4	0.972	0.969	0.804	0.810	0.798
5000	0.6	0.973	0.967	0.666	0.663	0.649
5000	0.8	0.974	0.968	0.482	0.470	0.460

J.7 Primary native-unit width distributions

K Resource–Evidence Evaluation of a Fixed Complete Report

This prospective experiment measures the precision and comparison evidence recovered by reusing selection observations at fixed evaluation budgets. It is separate from the earlier marginal-FCR and decision experiments: its coordinate and comparison outputs share a single error budget. All protocols, seeds, numerical tolerances, models and target families were frozen before their confirmation runs. Software-only sanity runs use distinct seeds. The three evaluation budgets are fixed in advance.

Table 21: Numerical completion for the full grid. J: reports changed by jitter; F: recorded interval failures; I: nonfinite widths; E: largest 256/512-node endpoint discrepancy in full-SE units, among the first ten reports per cell; B: violations of the theoretical same-covariance width bound. The fixed endpoint tolerance is 0.002.

n	f	J	F	I	E	B
500	0.4	354	0	0	0.000226	0
500	0.6	278	0	0	0.000511	0
500	0.8	224	0	0	0.000667	0
1000	0.4	236	0	0	0.000208	0
1000	0.6	149	0	0	0.000419	0
1000	0.8	120	0	0	0.000760	0
2500	0.4	106	0	0	0.000169	0
2500	0.6	64	0	0	0.000371	0
2500	0.8	50	0	0	0.000668	0
5000	0.4	32	0	0	0.000169	0
5000	0.6	27	0	0	0.000315	0
5000	0.8	14	0	0	0.000552	0

Table 22: Primary marginal widths in native objective units, not mixed-unit averages. Objectives 1–5 are clean accuracy, negative Brier, noise accuracy, blur accuracy and brightness accuracy. I counts nonfinite widths. Quantiles pool selected coordinates within each objective, not independent reports. Gaussian intervals are not clipped to physical parameter bounds in these width tables.

obj.	method	mean	median	p90	p99	I
1	Carving marginal	0.01334	0.01338	0.01518	0.01929	0
1	Split marginal	0.02010	0.01990	0.02274	0.02385	0
1	Hard marginal	0.03890	0.01349	0.01783	0.68778	0
1	Membership marginal	0.01844	0.01258	0.01435	0.03231	0
2	Carving marginal	0.01835	0.01783	0.02256	0.02861	0
2	Split marginal	0.02770	0.02720	0.03472	0.03702	0
2	Hard marginal	0.05148	0.01791	0.02541	0.89620	0
2	Membership marginal	0.02552	0.01729	0.02217	0.04710	0
3	Carving marginal	0.02116	0.02186	0.02619	0.03346	0
3	Split marginal	0.03142	0.03033	0.03771	0.03850	0
3	Hard marginal	0.07591	0.02210	0.03891	0.92486	0
3	Membership marginal	0.03181	0.01918	0.02388	0.07168	0
4	Carving marginal	0.01480	0.01457	0.01759	0.02248	0
4	Split marginal	0.02231	0.02106	0.02738	0.02859	0
4	Hard marginal	0.04564	0.01492	0.02001	0.89578	0
4	Membership marginal	0.02161	0.01333	0.01742	0.03037	0
5	Carving marginal	0.01408	0.01392	0.01598	0.02025	0
5	Split marginal	0.02133	0.02074	0.02432	0.02552	0
5	Hard marginal	0.03679	0.01412	0.01800	0.70088	0
5	Membership marginal	0.01856	0.01302	0.01538	0.02983	0

K.1 Joint Error Accounting

Given the complete label set P, let E_C denote any coordinate miss and E_D any rejection of a false improvement claim. The coordinate intervals use error $\alpha_C/(|P|d)$ each. The prespecified reference contrasts have conditional superuniform one-sided p-values under their composite nulls; Holm's procedure at α_D therefore controls $\mathbb{P}(E_D \mid P)$ without requiring independence (Holm, 1979). A union bound gives

$$\mathbb{P}(E_C \cup E_D \mid P) \leq \mathbb{P}(E_C \mid P) + \mathbb{P}(E_D \mid P) \leq \alpha_C + \alpha_D.$$

Different targets may have different nuisance statistics; integrating each target's conditional guarantee over its nuisance statistic leaves the same event P for this argument. The target-selection rule may depend on P, but cannot inspect additional continuous scores to choose hypotheses. We set $\alpha_C = \alpha_D = 0.025$. The Gaussian guarantee belongs to carving and independent splitting under

Table 23: Primary simultaneous widths in native objective units, not mixed-unit averages. Objectives 1–5 are clean accuracy, negative Brier, noise accuracy, blur accuracy and brightness accuracy. I counts nonfinite widths. Quantiles pool selected coordinates within each objective, not independent reports. Gaussian intervals are not clipped to physical parameter bounds in these width tables.

obj.	method	mean	median	p90	p99	I
1	Carving simultaneous	0.02123	0.02134	0.02430	0.03059	0
1	Split simultaneous	0.03190	0.03165	0.03607	0.03793	0
1	Global max-T	0.02110	0.02070	0.02368	0.02447	0
1	LSI max-T	0.02067	0.02029	0.02319	0.02397	0
2	Carving simultaneous	0.02921	0.02844	0.03596	0.04566	0
2	Split simultaneous	0.04396	0.04319	0.05515	0.05897	0
2	Global max-T	0.02908	0.02858	0.03644	0.03803	0
2	LSI max-T	0.02848	0.02800	0.03571	0.03729	0
3	Carving simultaneous	0.03366	0.03470	0.04184	0.05286	0
3	Split simultaneous	0.04987	0.04818	0.05991	0.06165	0
3	Global max-T	0.03299	0.03178	0.03958	0.04038	0
3	LSI max-T	0.03232	0.03114	0.03876	0.03962	0
4	Carving simultaneous	0.02356	0.02328	0.02796	0.03557	0
4	Split simultaneous	0.03542	0.03346	0.04347	0.04542	0
4	Global max-T	0.02343	0.02192	0.02875	0.02969	0
4	LSI max-T	0.02295	0.02146	0.02815	0.02914	0
5	Carving simultaneous	0.02241	0.02227	0.02563	0.03221	0
5	Split simultaneous	0.03385	0.03296	0.03859	0.04061	0
5	Global max-T	0.02238	0.02152	0.02540	0.02630	0
5	LSI max-T	0.02192	0.02108	0.02487	0.02578	0

their stated assumptions. Global max-T and LSI have their respective unconditional guarantees; using the same observed error metric does not make the guarantees identical.

The coordinate box also supplies hypervolume and nonnegative-utility projections, with no further error allocation. For bounded scores, we intersect it with the known physical bounds before hypervolume projection. An empty intersection is recorded as failure to cover, not discarded. Marginal contrast intervals at error 0.025 are diagnostic outputs; only Holm-adjusted directional claims enter E_D.

K.2 Paired Nested Budgets

For each of 1,000 independent outer replications per score pool, we sample A with 3,000 records and draw its independent, fixed-range offset once. This fixes the complete report P for all methods and budgets. An independently sampled stream of 4,000 records supplies nested prefixes $B_{1000} \subset B_{2000} \subset B_{4000}$. Sampling is with replacement from the frozen score pool. The same image's clean and corrupted scores stay in one record. The bank task similarly retains all model scores for one row together.

At each B budget, splitting uses only B for estimation, while carving uses $A \cup B$ and the appropriate fraction $f = 3000/(3000 + |B|)$. Means and plug-in covariances are recomputed at each budget; no held-out covariance determines the selection offset. There are ten independent seed groups of 100 replications, and all methods share each complete draw. The independent auditor recreates A, every B prefix, the offset, and the report and checks their hashes against the raw output.

Four methods are retained at every budget: carving, matched splitting, global max-T, and LSI max-T. Carving/splitting use coordinate Bonferroni and contrast Holm. Global coordinate and contrast calibration protect all candidate targets; LSI uses coordinate localization budget 0.005 within the coordinate budget 0.025, and the same global contrast family. Calibration uses 2,047 independent Gaussian draws. Truth from the full pool is used only to score outputs, never to choose the report, targets, offsets, regularization, confidence levels or inference method.

Precision and evidence. The full coordinate width requirement is 0.04 times each objective's known score range, without clipping intervals to make them appear precise. We report the within-

report fraction of coordinates meeting it, the event that all coordinates meet it, and the number of true and false individual comparison claims. Efficiency first averages the coordinatewise carving/splitting width ratios within each report, then averages over outer replications. This avoids mixing measurement units or giving larger reports extra weight.

K.3 Independent Frozen Tabular Task

We use the original 45,211-row `bank-full.csv` from the UCI Bank Marketing archive (Moro et al., 2014), rather than its later 41,188-row variant. A fixed random permutation reserves 30,000 training rows and a disjoint 15,211-row evaluation pool. Call duration is excluded. Numeric standardization and categorical one-hot encoding are fitted on training rows only, with unseen categories ignored.

The roster is fixed at 16 models: logistic regression, histogram gradient boosting, random forest and extremely randomized trees, each with positive class weight $1, 2, 4, 8$. Logistic regression uses $C = 1$ and 2,000 iterations; boosting uses 200 iterations, 15 leaves, learning rate 0.05, $\ell_2 = 1$ and no early stopping; both forests use 256 trees, minimum leaf size five, and square-root feature subsampling. Family seeds and all settings are frozen, with no evaluation-pool tuning or model removal.

For probability p, label y, and fixed classification threshold 0.5, the five scores are accuracy, $-(p - y)^2$, normalized clipped log score, $\mathbf{1}\{\widehat{y} = 1, y = 1\}$ and $\mathbf{1}\{\widehat{y} = 0, y = 0\}$. The log score clips probabilities to $[10^{-6}, 1 - 10^{-6}]$ and divides by $\log(10^6)$, giving range $[-1, 0]$. The last two scores are joint correctness yields, not recall or specificity; their sum is accuracy. The joint score vector preserves this exact redundancy. Comparisons to the fixed unweighted logistic reference are accuracy noninferiority within 0.02, Brier-score improvement exceeding 0.01, and each yield improvement exceeding 0.01.

Independent auditing reconstructs all 1,216,880 scores, verifies frozen model parameters, feature preprocessing, row identities and score ranges, and reloads all 16 models on 257 predetermined evaluation rows. The maximum score-formula discrepancy is 5.6×10^{-17}; replayed probabilities match exactly. This supplies an independent scoring source relative to CIFAR. The sampling population is the finite row pool; repeated-contact and temporal dependence fall outside this resampling design.

K.4 Numerical implementation and validation

Resource experiments use 512 quadrature nodes, checked against 1024 on the first ten replications at the frozen 0.002-SE endpoint tolerance. A failed target receives a whole-line interval or $p = 1$ and remains in the denominator. Failures prevent precision attainment and finite-only report-width summaries. Numerical completion is independently recomputed. Statistical errors and operational failures, and their union, are reported separately.

K.5 All Fixed-Pool Resource Results

The following tables retain every frozen budget and method. Uncertainty uses independent outer reports, not coordinates or budget rows.

K.6 Controlled Geometry and Information Experiment

Eight known-Gaussian cells cross two mean layouts, two separation multipliers and two candidate correlations at $K = 8, d = 3$. A two-anchor layout places other candidates below the anchors; an all-trade-off layout spreads all eight along opposing first and second objectives. Multipliers 0.5 and 2 scale the entire mean layout in selection-sample standard errors, not an isolated domination margin. Candidate correlations are 0 or 0.6; objective correlation is 0.25. Each cell uses 500 paired replications, $|A| = 100$, and independent nested $|B| = 50, 200$, with known covariance and no jitter. The same $0.025/0.025$ budget and fixed-reference zero-margin contrasts are used. Hard intervals are a coordinate-only diagnostic.

Table 24: CIFAR-10: joint coordinate–comparison reports (1,000 outer replications; $|A| = 3000$). Joint, coordinate and false-claim columns count erroneous reports; they overlap and must not be added. Ready is the mean fraction of coordinates meeting the frozen precision. Width is the paired ratio to splitting at the same B.

$\lvert B\rvert$	Method	Joint	Coord.	Claim	Ready	True claims	Width
1000	Carving	23	23	0	0.872	12.283	0.548
1000	Splitting	32	32	0	0.189	10.733	1.000
1000	Global max-T	13	13	0	0.919	12.209	0.529
1000	LSI max-T	16	16	0	0.919	12.209	0.525
2000	Carving	18	18	0	0.949	12.559	0.670
2000	Splitting	25	23	2	0.689	11.830	1.000
2000	Global max-T	14	14	0	0.944	12.392	0.667
2000	LSI max-T	14	14	0	0.949	12.392	0.662
4000	Carving	13	13	0	0.995	12.799	0.781
4000	Splitting	16	15	1	0.919	12.486	1.000
4000	Global max-T	8	8	0	1.000	12.624	0.796
4000	LSI max-T	8	8	0	1.000	12.624	0.790

Table 25: Bank: joint coordinate–comparison reports (1,000 outer replications; $|A| = 3000$). Joint, coordinate and false-claim columns count erroneous reports; they overlap and must not be added. Ready is the mean fraction of coordinates meeting the frozen precision. Width is the paired ratio to splitting at the same B.

$\lvert B\rvert$	Method	Joint	Coord.	Claim	Ready	True claims	Width
1000	Carving	18	16	2	0.707	13.057	0.598
1000	Splitting	23	21	2	0.356	10.187	1.000
1000	Global max-T	28	27	1	0.915	14.658	0.457
1000	LSI max-T	23	22	1	0.887	14.658	0.468
2000	Carving	16	16	0	0.830	14.234	0.709
2000	Splitting	14	13	1	0.567	12.666	1.000
2000	Global max-T	24	24	0	0.964	15.066	0.576
2000	LSI max-T	19	19	0	0.953	15.066	0.590
4000	Carving	13	13	0	0.969	15.156	0.810
4000	Splitting	10	10	0	0.837	14.580	1.000
4000	Global max-T	24	24	0	1.000	15.518	0.688
4000	LSI max-T	18	18	0	1.000	15.518	0.705

Table 26: CIFAR-10: supporting diagnostics under the joint $0.025/0.025$ allocation. FCR is report-weighted; all-ready, HV errors and failures are counts out of 1,000. HV is a conservative projection. These are not the earlier marginal-0.05 experiments.

$\lvert B\rvert$	Method	FCR	MC SE	All ready	HV errors	HV width	Failures
1000	Carving	0.0014	0.0003	0	0	0.201	0
1000	Splitting	0.0021	0.0004	0	0	0.365	0
1000	Global max-T	0.0007	0.0002	0	0	0.193	0
1000	LSI max-T	0.0008	0.0002	0	0	0.192	0
2000	Carving	0.0013	0.0004	243	0	0.173	0
2000	Splitting	0.0013	0.0003	0	0	0.259	0
2000	Global max-T	0.0010	0.0003	3	0	0.173	0
2000	LSI max-T	0.0010	0.0003	3	0	0.171	0
4000	Carving	0.0007	0.0002	877	0	0.143	0
4000	Splitting	0.0015	0.0006	0	0	0.183	0
4000	Global max-T	0.0004	0.0002	996	0	0.146	0
4000	LSI max-T	0.0004	0.0002	997	0	0.145	0

Table 27: Bank: supporting diagnostics under the joint $0.025/0.025$ allocation. FCR is report-weighted; all-ready, HV errors and failures are counts out of 1,000. HV is a conservative projection. These are not the earlier marginal-0.05 experiments.

$\|B\|$	Method	FCR	MC SE	All ready	HV errors	HV width	Failures
1000	Carving	0.0006	0.0002	0	0	0.031	0
1000	Splitting	0.0011	0.0003	0	0	0.052	0
1000	Global max-T	0.0011	0.0003	0	0	0.023	0
1000	LSI max-T	0.0008	0.0002	0	0	0.024	0
2000	Carving	0.0007	0.0002	0	0	0.026	0
2000	Splitting	0.0007	0.0003	0	0	0.037	0
2000	Global max-T	0.0014	0.0005	63	0	0.021	0
2000	LSI max-T	0.0011	0.0004	7	0	0.021	0
4000	Carving	0.0009	0.0004	236	0	0.021	0
4000	Splitting	0.0004	0.0002	0	0	0.026	0
4000	Global max-T	0.0017	0.0006	1000	0	0.018	0
4000	LSI max-T	0.0013	0.0005	1000	0	0.018	0

An independent 100,000-selection-draw table per cell estimates the frequency of every complete label set. This frequency is distinct from the Gaussian mass of an individual target's support given its nuisance statistic. The auditor checks the latter probability from the stored interval union; this integral check is not an independent reconstruction of its geometry. An unseen label mask is not assigned true probability zero. Frequency distributions are descriptive conditional on the frequency table. Solver failures remain separate from nonfinite endpoints returned without an exception; neither is silently removed from hard-width summaries.

Computation and information measure different aspects of the report. All 4,000 outer pairs and 8,000 budget records are retained (Table 28). Carving's mean coordinate width ratio to same-budget splitting ranges from 0.645 to 0.815 at $B = 50$ and 0.850 to 0.931 at $B = 200$. Observed carving joint errors range from 12 to 22 out of 500. The largest count has a one-sided 95% CP upper bound of 0.0622.

Increasing the mean-layout multiplier can increase required relations while decreasing information loss. For the uncorrelated all-trade-off layout, the multiplier change $0.5 \to 2$ raises mean relations per target from 27.636 to 39.970, while mean coordinate log nuisance mass increases from -1.019 to -0.575 and the $B = 50$ width ratio falls from 0.812 to 0.753. This contrast changes the whole mean layout. The same selection geometry, label frequencies and nuisance masses recur at both B budgets; increasing independent data changes precision without changing that geometry. Relation counts describe reconstruction work; conditional support mass describes the information retained after selection.

Every coordinate support in this design has one component, whereas selected contrast supports include two to four components. Coordinate geometry is therefore not a summary of all target directions. All hard-coordinate intervals are finite in the observed runs, but report-average hard/split width ratios have long tails: in the all-trade-off, multiplier-0.5, uncorrelated $B = 200$ cell, the mean, median and p99 are 8.036, 2.337 and 126.023. Finite observed endpoints do not prove finite population mean length. The numerical audit has no failures or changed Holm vectors; its maximum 512/1024 endpoint discrepancy is 0.000141 SE in the prespecified first ten replications per cell.

K.7 All Controlled Gaussian Resource Cells

K.8 Complete-output cost and numerical agreement

Sequential one-thread timing uses the first ten frozen reports per score pool at each budget, yielding 60 budget reports. Each backend computes all coordinates, reference contrasts, Holm decisions, precision flags and hypervolume. All three backends agree within the numerical tolerances on these outputs. Table 30A includes selection, sample means, covariance estimation, inversion and mandatory 512/1024-node refinement. Cached model training, scoring and the shared warm-up are outside these clocks. Nested budgets remain paired through the same ten reports per pool.

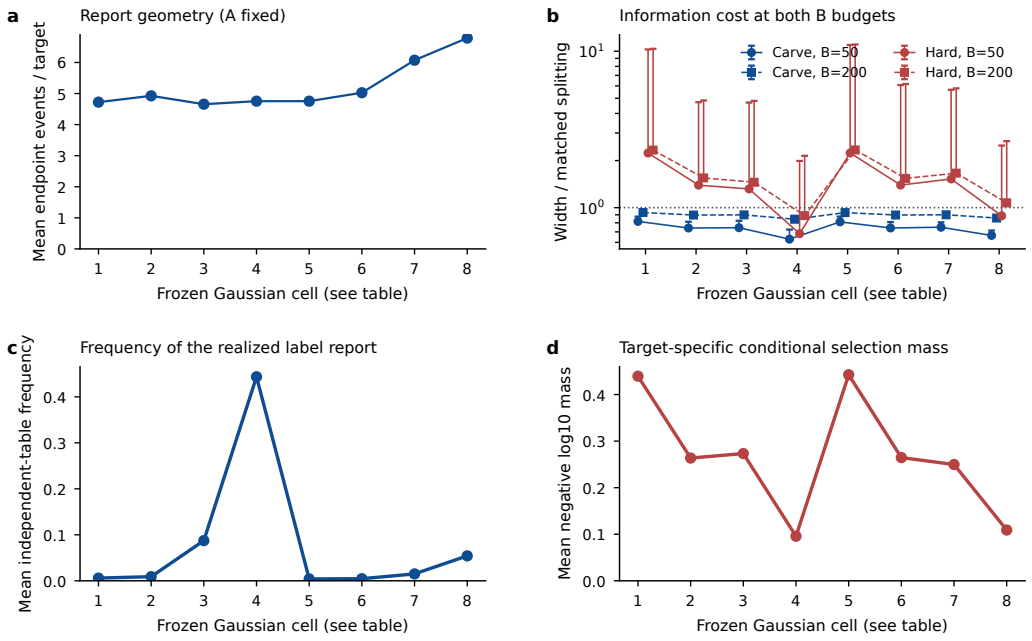

Figure 3: Geometry and information respond differently to the frozen design. All eight cells have 500 paired reports at each B. (a) Coordinate endpoint events; selection geometry is identical across budgets. (b) Median report-average coordinate width ratios; upper whiskers are p90, not uncertainty intervals. Hard conditioning is a coordinate-only diagnostic. (c) Mean independent-table frequency of the realized complete label set. (d) Mean negative log10 nuisance-conditioned coordinate support mass. The last two panels are distinct, truth-based diagnostics, not tuning inputs. Cell indices refer to Table 29; connecting lines do not represent a continuous intervention.

Table 28: Eight Gaussian cells, each with 500 fixed reports and two budgets. Endpoint counts summarize coordinate geometry. C/S/G/L columns count joint errors for carving/splitting/global/LSI. Width ratios use matched splitting; hard is a coordinate-only diagnostic, with solver failures retained. A dash means the full mean is unavailable, not silently trimmed.

Cell	B	Mean s	Endpoints	C	S	G	L	Carve width	Hard width	Hard fails
1	50	3.91	4.72	14	20	8	8	0.815	6.909	0
1	200	3.91	4.72	16	13	7	6	0.931	7.022	0
2	50	4.04	4.93	15	15	9	6	0.742	3.009	0
2	200	4.04	4.93	15	20	13	9	0.898	3.157	0
3	50	3.41	4.66	13	14	5	2	0.747	3.735	0
3	200	3.41	4.66	13	18	6	5	0.900	3.884	0
4	50	2.44	4.75	17	17	4	3	0.645	2.452	0
4	200	2.44	4.75	12	13	6	4	0.850	2.653	0
5	50	3.95	4.75	18	19	14	14	0.812	7.923	0
5	200	3.95	4.75	21	16	13	11	0.930	8.036	0
6	50	4.32	5.03	16	17	13	11	0.743	3.169	0
6	200	4.32	5.03	22	17	19	18	0.898	3.315	0
7	50	5.71	6.07	14	15	7	7	0.753	3.517	0
7	200	5.71	6.07	20	22	14	10	0.902	3.657	0
8	50	6.57	6.79	17	21	14	13	0.667	1.641	0
8	200	6.57	6.79	14	14	10	9	0.860	1.828	0

Table 29: Gaussian cell definitions and distinct selection diagnostics. Frequency is the average independent-table frequency of the realized complete report; log mass is the mean coordinate-specific nuisance-conditioned log probability. Neither is an estimated tuning input. Separation scales the full mean layout.

Cell	Layout	Separation	Correlation	Label frequency	Log conditional mass
1	two anchors	0.5	0.0	0.0059	-1.012
2	two anchors	0.5	0.6	0.0090	-0.607
3	two anchors	2.0	0.0	0.0874	-0.629
4	two anchors	2.0	0.6	0.4436	-0.221
5	all tradeoffs	0.5	0.0	0.0040	-1.019
6	all tradeoffs	0.5	0.6	0.0046	-0.609
7	all tradeoffs	2.0	0.0	0.0150	-0.575
8	all tradeoffs	2.0	0.6	0.0542	-0.251

A separate 40-report Gaussian study uses $K = 4, d = 2$, $f = .6$ and 20 reports per independent/equicorrelated design. The accuracy- and loss-stopped bootstrap learners from Appendix I each fit every selected coordinate and candidate-minus-reference contrast, with reference candidate 0. Coordinate and Holm budgets are 0.025 each; both objective contrasts have zero margins. Each method evaluates 404 target directions in total. Table 30B includes setup, bootstrap, fitting, inversion, numerical refinement and Holm processing. For the four learned configurations, external validation adds mean 0.0166–0.0171 seconds and oracle auditing adds 0.0273–0.0275 seconds per report.

Whole-report approximation checks require every endpoint discrepancy at most 0.05 target SD, checked pivot discrepancies below 0.01, and full Holm agreement. The numerical-refinement tolerance is 0.002 SD, using grid steps 0.01/0.005 for the learned laws. Oracle values are used only for these post-fit checks. All fits complete, but the learned laws exceed the approximation tolerances in every report. Costs are therefore shown alongside attainment, rather than as a matched-precision speed ranking. This is a timing and numerical study, separate from the coverage confirmations.

Table 30: Completed whole-output timings in seconds, retaining all attempted reports. A: mean standalone times over ten frozen reports per pool and budget; the three budgets reuse the same reports. B: mean and median standalone costs over 40 Gaussian reports, including all target fits and refinements. Attained counts reports meeting every numerical, approximation and decision check. All methods complete with zero target failures. Standalone costs exclude cached scoring and separately recorded external/oracle diagnostics.

A. Paired real-score report replay

Pool	$\|B\|$	Report-sensitive	All-pair	Root sweep
CIFAR	1,000	1.174	1.239	1.501
CIFAR	2,000	1.147	1.252	1.517
CIFAR	4,000	1.159	1.262	1.523
Bank	1,000	3.005	3.081	3.391
Bank	2,000	3.171	3.132	3.526
Bank	4,000	3.288	3.338	3.707

B. Gaussian full-report cost and attainment

Method	Mean	Median	Attained
Exact report-sensitive	0.201	0.200	40/40
Exact all-pair	0.202	0.198	40/40
Exact root sweep	0.202	0.199	40/40
BB accuracy, 2,048	3.887	3.906	0/40
BB accuracy, 8,192	16.574	15.685	0/40
BB loss, 2,048	7.111	7.209	0/40
BB loss, 8,192	23.850	23.624	0/40

L Official Model Pool and Scoring Protocol

L.1 Candidates and paired scores

The frozen RobustBench roster contains 24 pretrained models (Croce et al., 2021). All 21 models with accessible authentic weights and successful loading checks were scored. The unavailable entries are Xu2023Exploring_WRN-28-10, Chen2024Data_WRN_34_10 and Bartoldson2024Adversarial_WRN-82-8. Inclusion was determined by availability before full-pool scoring. The original roster, transport records and pinned registry revision are retained.

Each model receives all 10,000 CIFAR-10 test images, with 1,000 per class. Five fixed maximization scores are clean accuracy, negative clean Brier score, and Gaussian-noise, defocus-blur and brightness accuracy. Corruptions use official CIFAR-10-C arrays at severity 3. Negative Brier is $-\sum_{a=1}^{10}(p_a - \mathbf{1}\{a = y\})^2$; accuracy is binary. All variants of one original image form one joint record across models and objectives. The array contains $10{,}000 \times 21 \times 5$ finite scores. A 64-image GPU replay per model reproduces all five stored scores exactly.

Means of this fixed score pool define the performance and comparison targets. Paired resampling preserves dependence across models, objectives and corruption variants while making coverage and false claims directly measurable. Current selection, covariance estimation and calibration settings are given in Appendix J.

L.2 Complete evaluated model roster

Table 31: All 21 included official models and their full-pool means. Index is one-based in the original 24-model list, so gaps denote unavailable weights. A: clean accuracy; B: negative Brier; N/D/L: Gaussian-noise, defocus-blur and brightness accuracy at severity 3. Scores use all 10,000 paired images, not a selected validation subset.

index	model	A	B	N	D	L
1	Standard	0.9478	-0.0867	0.3919	0.8898	0.9378
2	Andriushchenko2020Understanding	0.7984	-0.3209	0.7568	0.7645	0.7867
3	Wong2020Fast	0.8334	-0.2705	0.7692	0.7938	0.8235
4	Carmon2019Unlabeled	0.8969	-0.2309	0.8188	0.8632	0.8884
5	Hendrycks2019Using	0.8711	-0.2303	0.7862	0.8213	0.8701
6	Zhang2019Theoretically	0.8492	-0.2497	0.7887	0.8082	0.8403
7	Sehwag2021Proxy_R18	0.8459	-0.2502	0.7867	0.8134	0.8374
8	Rebuffi2021Fixing_R18_ddpm	0.8353	-0.3208	0.7792	0.8026	0.8200
9	Rade2021Helper_R18_ddpm	0.8686	-0.2691	0.8033	0.8296	0.8546
10	Addepalli2021Towards_RN18	0.8024	-0.3816	0.7615	0.7626	0.7862
11	Addepalli2022Efficient_RN18	0.8571	-0.3957	0.8039	0.8178	0.8500
12	Debenedetti2022Light_XCiT-S12	0.9006	-0.2487	0.8417	0.8683	0.8990
13	Wang2023Better_WRN-28-10	0.9244	-0.2285	0.8529	0.8932	0.9175
15	Cui2023Decoupled_WRN-28-10	0.9216	-0.2497	0.8544	0.8924	0.9149
18	Hendrycks2020AugMix_WRN	0.9508	-0.0747	0.7955	0.9458	0.9447
19	Hendrycks2020AugMix_ResNeXt	0.9583	-0.0639	0.7577	0.9508	0.9502
20	Kireev2021Effectiveness_Gauss50percent	0.9324	-0.1089	0.9064	0.8902	0.9173
21	Kireev2021Effectiveness_AugMixNoJSD	0.9497	-0.0800	0.7522	0.9444	0.9424
22	Kireev2021Effectiveness_RLAT	0.9310	-0.1097	0.8571	0.8962	0.9215
23	Kireev2021Effectiveness_RLATAugMix	0.9475	-0.0804	0.8631	0.9400	0.9447
24	Modas2021PRIMEResNet18	0.9306	-0.1523	0.8902	0.9165	0.9250

M FIGURE SOURCE DATA AND COMPLETE REPORT DISPLAYS

Figure 1 uses editable PowerPoint objects; other figures use Python. Vector PDFs and source CSV files are retained. The algebraic support illustration is checked against independent sweeps. Quantitative source files retain model identities, uncertainty definitions and all observations used in each figure.

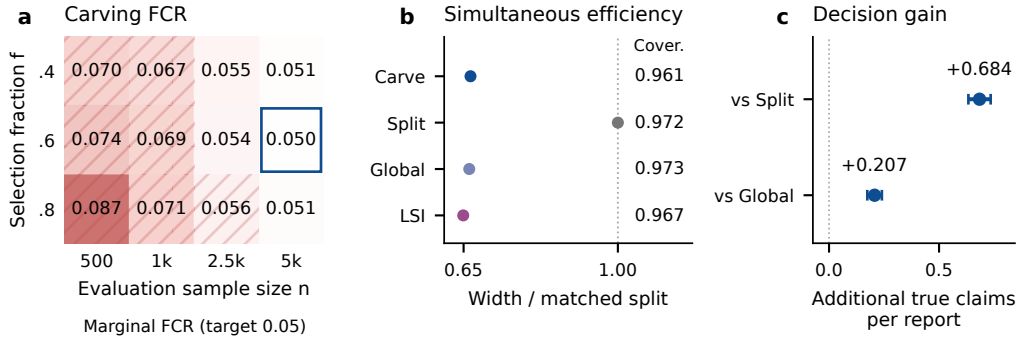

Figure 4: Retained official-model experiment, separate from the new joint resource protocol. (a) All 12 carving marginal FCR cells; hatching uses pointwise 95% MC intervals without across-grid correction. (b) Primary simultaneous widths and report coverage; global/LSI target unconditional protection. (c) Carving's paired gain in true individual claims, not models. Whiskers in (b,c): mean ± 1.96 report-based MC SE. Each cell has 1,000 reports from the fixed pool. Marginal, simultaneous and decision studies here each use their original separate error budgets; they are not joint $0.025/0.025$ results.

Table 32: Known-covariance whole-report time in seconds (median of three paired designs). R: comparison-root sweep; D: all-pair dominance compression; P: report-sensitive compression. Last two columns split P's support and inversion costs; totals also include initial-front and line preparation. The medians of components need not sum to the median of totals.

K	d	s	R total	D total	P total	P support	P invert
32	2	3	.116	.088	.051	.007	.043
64	5	4	1.730	.705	.234	.061	.172
128	8	4	15.580	3.563	.526	.227	.296
16	2	16	.387	.361	.352	.057	.294

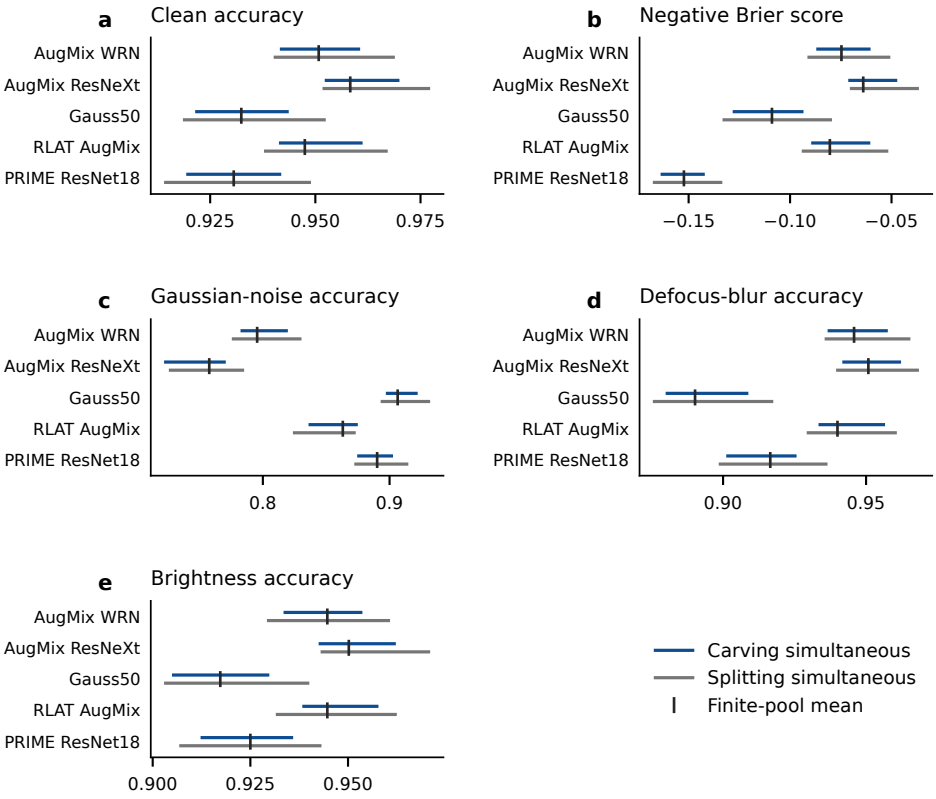

Figure 5: A complete real-score report from primary replication 0 ($n = 5000, f = .6$). Every reported label and all five objectives appear in native units. Each method allocates $0.05/(5|P|)$ error per coordinate. Colored horizontal segments are the stored confidence intervals; black marks show the fixed score-pool means that define the targets. Full candidate names and their original roster indices accompany every interval in the source CSV. This display illustrates the complete set of outputs for one report.

N RECONSTRUCTION SCALE AND REPORT-SPECIFIC RELIABILITY

This appendix gives the new conditional-reliability design, the independent bounded-score option, and the expanded computational grid. The reconstruction lower bound and its Gaussian realization are proved in Appendix A.3. All existing calibration, application and timing studies retain their original protocols and observations.

N.1 A SEPARATION BETWEEN GLOBAL AND REPORT-CONDITIONAL RELIABILITY

Proposition 4 (Global protection and report-specific error)**.** *Let $K = d = 2$, $\theta_1 = (\delta, \delta)$ and $\theta_2 = (0, 0)$, with $\delta \geq 0$. Fix $f \in (0, 1)$ and independent Gaussian vectors $A \sim \mathcal{N}(\theta, I_4/f)$, $B \sim \mathcal{N}(\theta, I_4/(1-f))$. Select $\widehat{\mathcal{P}}(A)$ and estimate with $T = fA + (1-f)B$. The probability of report $\{2\}$ is $[1 - \Phi(\delta\sqrt{f/2})]^2$. As $\delta \to \infty$, global Gaussian max-T bands have simultaneous coverage tending to zero conditional on this report; global one-sided max-T tests have probability tending to one of making at least one false claim $\theta_{2,c} > 0$ conditional on the same report. Their unconditional family-wise guarantees remain valid. Carving and independent splitting retain their report-conditional guarantees.*

Proof. For each objective, the difference and midpoint are independent:

$$D_c = A_{2,c} - A_{1,c} \sim \mathcal{N}(-\delta, 2/f), \qquad M_c = (A_{1,c} + A_{2,c})/2 \sim \mathcal{N}(\delta/2, 1/(2f)).$$

The two objectives are independent. Report $\{2\}$ occurs exactly when both differences are positive, giving its stated probability. For $\delta > 0$ and every $x \geq 0$, the normal-tail ratio gives

$$\Pr(\delta D_c > x \mid D_c > 0) = \frac{1 - \Phi((\delta + x/\delta)\sqrt{f/2})}{1 - \Phi(\delta\sqrt{f/2})} \longrightarrow e^{-fx/2}.$$

Thus $D_c = O_P(\delta^{-1})$ after conditioning; the independent midpoint keeps its original law. Consequently, $A_{2,c} = \delta/2 + O_P(1)$ and $T_{2,c} = f\delta/2 + O_P(1)$ conditional on $\widehat{\mathcal{P}}(A) = \{2\}$. The global two-sided and one-sided critical values over four coordinates are fixed finite constants. Both asserted conditional limits follow. Unconditionally, $T - \theta \sim \mathcal{N}(0, I_4)$, so global calibration is unchanged. Proposition 3 and independence of B give the remaining report-conditional guarantees for every finite δ. □

Frozen conditional experiment. The complete grid uses $\delta \in \{0, 1, 2, 3, 4\}$, $f \in \{.6, .9\}$ and all three reports $\{1\}, \{2\}, \{1, 2\}$. Each group receives 4,000 independent conditional draws, for 120,000 draws in total. With $r = 1 - \Phi(\delta\sqrt{f/2})$, the report probabilities are $(1-r)^2, r^2, 2r(1-r)$. These probabilities, rather than the equal conditional sample sizes, determine unconditional averages.

The sampler independently generates M_c and the appropriate truncated D_c, then sets $A_{1,c} = M_c - D_c/2$ and $A_{2,c} = M_c + D_c/2$. For the two-label report, the two opposite-sign configurations are equally likely. The held-out B is drawn independently afterward. The confirmation seed is 2026091701, with separate streams for each $(\delta, 10f, \text{report})$; software checks use a distinct seed.

Simultaneous coordinate bands and positive-threshold tests each have a *separate* error budget .05. Carving and splitting use coordinate Bonferroni bands and selected-family Holm tests. Global max-T calibrates the four independent standard-normal errors exactly. Difference-localized LSI uses localization budget $\nu = .005$, the exact maximum of the two absolute standardized difference errors, twice the resulting localization radius, and inference at $.05 - \nu$. On the localization event its fixed intermediate family lies between the report and the plausible family, so the usual unconditional LSI argument applies. Both candidates remain plausible in every simulated draw; for the two-label report this inclusion is required by the observed report itself.

Tables 33–34 retain all groups, including null, rare and common reports. At $\delta = 2, f = .9$, the two-label report has probability .16356. Its carving/global/LSI coverage estimates are .95100/.87825/.88850; the corresponding false-threshold-report rates are .024/.099/.089. For global and LSI, even Clopper–Pearson intervals using error $.05/120$ per interval have coverage upper limits .89579/.90534 and false-report lower limits .08309/.07388. These adjustments apply separately

Table 33: All 30 complete-report groups: conditional simultaneous coverage. C/S/G/L denote carving, splitting, global max-T and LSI. Each group has 4,000 conditional draws. Analytic $\Pr(P)$, not group sample size, determines the frequency of that report in an unconditional evaluation. Individual coverage Monte Carlo standard errors are at most .008.

δ	f	P	$\Pr(P)$	C	S	G	L
0	0.6	$\{1\}$	0.25	0.947	0.949	0.975	0.979
0	0.6	$\{2\}$	0.25	0.956	0.944	0.974	0.975
0	0.6	$\{1,2\}$	0.5	0.951	0.953	0.952	0.957
0	0.9	$\{1\}$	0.25	0.949	0.957	0.974	0.976
0	0.9	$\{2\}$	0.25	0.946	0.952	0.974	0.977
0	0.9	$\{1,2\}$	0.5	0.948	0.959	0.947	0.953
1	0.6	$\{1\}$	0.50135	0.956	0.954	0.985	0.987
1	0.6	$\{2\}$	0.08523	0.954	0.950	0.963	0.966
1	0.6	$\{1,2\}$	0.41342	0.946	0.944	0.937	0.943
1	0.9	$\{1\}$	0.56075	0.950	0.948	0.983	0.984
1	0.9	$\{2\}$	0.063085	0.951	0.945	0.950	0.954
1	0.9	$\{1,2\}$	0.37616	0.954	0.953	0.937	0.942
2	0.6	$\{1\}$	0.74535	0.948	0.958	0.981	0.982
2	0.6	$\{2\}$	0.018676	0.948	0.950	0.933	0.939
2	0.6	$\{1,2\}$	0.23597	0.953	0.947	0.921	0.931
2	0.9	$\{1\}$	0.82836	0.952	0.951	0.982	0.983
2	0.9	$\{2\}$	0.0080741	0.952	0.952	0.900	0.909
2	0.9	$\{1,2\}$	0.16356	0.951	0.954	0.878	0.889
3	0.6	$\{1\}$	0.90217	0.950	0.948	0.978	0.980
3	0.6	$\{2\}$	0.0025174	0.947	0.943	0.888	0.899
3	0.6	$\{1,2\}$	0.095313	0.958	0.954	0.883	0.896
3	0.9	$\{1\}$	0.95632	0.950	0.956	0.980	0.983
3	0.9	$\{2\}$	0.00048778	0.953	0.956	0.754	0.772
3	0.9	$\{1,2\}$	0.043196	0.950	0.950	0.731	0.750
4	0.6	$\{1\}$	0.97174	0.955	0.952	0.981	0.982
4	0.6	$\{2\}$	0.00020249	0.951	0.949	0.810	0.825
4	0.6	$\{1,2\}$	0.028055	0.952	0.958	0.792	0.813
4	0.9	$\{1\}$	0.99272	0.950	0.953	0.975	0.980
4	0.9	$\{2\}$	1.3287e-05	0.950	0.946	0.536	0.557
4	0.9	$\{1,2\}$	0.0072638	0.953	0.949	0.467	0.498

to the 120 coverage probabilities and to the 120 false-report probabilities. Probability-weighted global/LSI coverage in this slice is .96458/.96715. Thus the conditional separation coexists with the intended unconditional protection.

Carving uses 256 quadrature nodes. For the first 20 draws per group, 512-node refinement changes pivots by at most 2.31×10^{-6} and endpoints by at most 1.18×10^{-4} target standard deviations; no Holm vector changes. All 120,000 draws complete without target failures. Width summaries use only the prespecified first 100 draws per group. Independent replay verifies the sampling streams, labels, stored-pivot coverage indicators and global/LSI decisions. Group-wise binomial intervals and width summaries accompany the retained raw results.

N.2 Finite-sample independent protection for bounded scores

Proposition 5 (Bounded-score report protection)**.** *Let the selection data and its independent offset determine a complete report P and the prescribed target family. Let B contain $n_B \geq 2$ i.i.d. evaluation records independent of selection. For each target q, suppose its score lies in a known interval of length R_q and has mean ν_q. Let $\bar{B}_q$ and $\widehat{v}_q$ be its held-out sample mean and unbiased sample variance. Set*

$$C_q^B = \bar{B}_q \pm r_q(\alpha_q), \qquad r_q(\alpha_q) = \sqrt{\frac{2\widehat{v}_q \log(4/\alpha_q)}{n_B}} + \frac{7R_q \log(4/\alpha_q)}{3(n_B - 1)}.$$

These intervals have miss probability at most α_q conditional on selection. One-sided empirical-Bernstein p-values are conditionally superuniform for the corresponding fixed threshold nulls. Co-

Table 34: False threshold-claim reports among 4,000 conditional draws in each group. C/S use selected-family Holm; G/L use one-sided max-T. Threshold testing has its own .05 error budget, separate from coordinate bands. A count divided by 4,000 is the conditional error estimate.

δ	f	P	C	S	G	L
0	0.6	$\{1\}$	208	201	195	181
0	0.6	$\{2\}$	194	226	180	164
0	0.6	$\{1,2\}$	195	198	198	175
0	0.9	$\{1\}$	196	171	192	178
0	0.9	$\{2\}$	196	194	207	189
0	0.9	$\{1,2\}$	220	197	214	185
1	0.6	$\{1\}$	0	0	0	0
1	0.6	$\{2\}$	180	177	260	245
1	0.6	$\{1,2\}$	108	102	166	146
1	0.9	$\{1\}$	0	0	0	0
1	0.9	$\{2\}$	203	206	387	360
1	0.9	$\{1,2\}$	103	86	214	186
2	0.6	$\{1\}$	0	0	0	0
2	0.6	$\{2\}$	204	206	489	453
2	0.6	$\{1,2\}$	114	125	275	252
2	0.9	$\{1\}$	0	0	0	0
2	0.9	$\{2\}$	163	208	681	622
2	0.9	$\{1,2\}$	96	111	396	356
3	0.6	$\{1\}$	0	0	0	0
3	0.6	$\{2\}$	183	205	774	705
3	0.6	$\{1,2\}$	140	141	437	401
3	0.9	$\{1\}$	0	0	0	0
3	0.9	$\{2\}$	188	171	1511	1437
3	0.9	$\{1,2\}$	122	95	818	767
4	0.6	$\{1\}$	0	0	0	0
4	0.6	$\{2\}$	195	203	1207	1121
4	0.6	$\{1,2\}$	168	149	675	618
4	0.9	$\{1\}$	0	0	0	0
4	0.9	$\{2\}$	208	208	2471	2376
4	0.9	$\{1,2\}$	167	114	1592	1511

ordinate Bonferroni allocation at α_C and Holm testing at α_D therefore satisfy

$$\Pr(E_C \cup E_D \mid \widehat{\mathcal{P}} = P) \le \alpha_C + \alpha_D$$

for every positive-probability report, without a Gaussian assumption.

Proof. Condition on all selection data and its independent randomization. The labels, targets and budgets are now fixed, while the law of B is unchanged. Rescaling a target to $[0,1]$ and applying Maurer & Pontil (2009, Theorem 4) to both signs, each at error $\alpha_q/2$, gives the two-sided radius

$$r_q(\alpha_q) = \sqrt{\frac{2\widehat{v}_q \log(4/\alpha_q)}{n_B}} + \frac{7R_q \log(4/\alpha_q)}{3(n_B-1)}.$$

The same theorem for the upper deviation of the mean, with $\log(2/t)$ in place of $\log(4/\alpha_q)$, gives a level-t test of $\nu_q \le h_q$. Inverting this nested family gives a superuniform p-value for the composite null. Holm and a union bound then give the displayed error bound conditional on all selection data. Averaging over those data given P proves the report-conditional result. Independence among different targets is unnecessary. □

P-values and comparison ranges. For $x = (\bar{B}_q - h_q)_+$, let $a = \sqrt{2\widehat{v}_q/n_B}$ and $b = 7R_q/[3(n_B-1)]$. If $x = 0$, set $p_q = 1$. Otherwise the positive solution of $ay + by^2 = x$ is

$$y = \frac{2x}{a + \sqrt{a^2 + 4bx}}, \qquad p_q = \min\{1, 2e^{-y^2}\}.$$

Table 35: Independent empirical-Bernstein report option: paired replay of all 1,000 existing resource reports per pool at each budget. Coordinates and Holm comparisons share .025/.025. Joint errors count any coordinate miss or false claim. Width ratios use the same-report Gaussian splitting bands in Tables 24–25; parentheses are report-level MC SE. Zero observed errors give a one-sided 95% binomial upper bound .00299 per cell.

Pool	$\|B\|$	Joint errors	True claims/report	Width/split
CIFAR	1000	0	5.380	2.059 (0.0012)
CIFAR	2000	0	7.637	1.815 (0.0006)
CIFAR	4000	0	10.681	1.644 (0.0003)
Bank	1000	0	0.000	2.959 (0.0021)
Bank	2000	0	1.779	2.444 (0.0011)
Bank	4000	0	8.557	2.082 (0.0006)

Table 36: Independent empirical-Bernstein replay of the six small-sample image cells, 1,000 original reports per cell. Marginal FCR and simultaneous report errors use separate .05 procedures. The final column averages simultaneous interval width divided by the known objective range within and then across reports. Intervals are not clipped.

n	f	Marginal FCR	Report errors	Normalized width
500	0.4	0.000000	0	0.235
500	0.6	0.000000	0	0.321
500	0.8	0.000000	0	0.561
1000	0.4	0.000065	0	0.141
1000	0.6	0.000050	0	0.189
1000	0.8	0.000000	0	0.320

A zero-range target is deterministic and is handled directly. For a paired difference, the score is formed per record before computing its variance; its known range length is at most the sum of the two coordinate range lengths. This gives $2R_c$ for same-objective comparisons. All coordinate and reference-contrast dependencies are thereby retained. Mode choice and budgets are fixed from selection-side information before inspecting held-out scores. Binary enumeration checks cover probabilities .001 to .999 and sample sizes 2 to 500, together with inversion and scale-equivariance tests.

Paired application replay and precision cost. We replay the exact frozen selection and held-out draws from Appendices J–K. Archived streams reproduce the draws; hash checks verify resource draws and selection views. This is a new method evaluated on existing draws, rather than an independent confirmation batch. Tables 35 and 36 give every planned cell. The six resource settings use the original joint $.025/.025$ allocation. Each has zero observed joint errors among 1,000 reports, with mean width ratios 1.644–2.959 relative to Gaussian splitting. The six small-sample image settings use separate .05 marginal and simultaneous procedures; all have zero observed simultaneous misses. The conservative option therefore quantifies the precision and comparison cost of finite-sample protection under bounded independent sampling. It uses B alone for inference and leaves the Gaussian-approximation status of full-data plug-in carving unchanged.

N.3 Expanded candidate and objective scaling

The support grid crosses $K \in \{128, 512, 1024, 2048\}$, $d \in \{8, 32\}$ and $s \in \{4, 32\}$. Dense controls use $(K, d, s) = (32, 8, 32), (128, 8, 128)$. Whole-coordinate-report settings are $(128, 8, 4), (512, 8, 4), (1024, 8, 4), (1024, 32, 4)$. Each setting has three constructed designs, with seed 2026091702 and independent setting-specific streams. This grid extends the known-covariance computational study, rather than the number of trained application models.

Reported candidates have opposing first and second coordinates from .5 to 1; their remaining coordinates equal .75. Each omitted candidate lies independently .05–.25 below a designated reported witness in every objective. For a $(Kd) \times 4$ standard-normal loading matrix L, set

Table 37: Large-library computation, $s = 4$: median seconds over three paired designs. Support times measure one exact support; report times include selection, line preparation and all sd coordinate inversions. Known covariance is supplied in factorized form. All supports and endpoints agree. Full grids, ranges and dense controls appear in Appendix N.3.

Output	K	d	Report-sensitive	All-pair	Ratio
Support	128	32	0.014	0.101	7.3
Support	512	32	0.055	1.511	27.4
Support	1024	32	0.113	5.934	52.6
Support	2048	32	0.229	23.532	102.7
Report	512	8	1.184	47.884	40.4
Report	1024	8	2.131	190.174	89.2
Report	1024	32	14.548	763.258	52.5

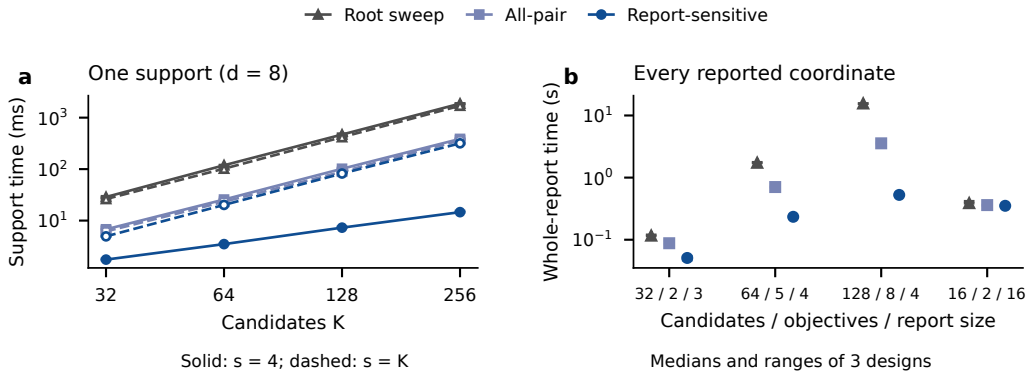

Figure 6: Report-sensitive computation. Centers and whiskers are medians and min–max ranges of three paired designs. (a) Single-support time, with dense reports shown as dashed curves. (b) Known-covariance whole-report time includes selection, line preparation, support and inversion. Shared-CPU Python reference implementations.

$v_r = \sqrt{.3\|L_{r,:}\|^2/4 + .7}$ and supply the known covariance

$$V = .25^2 \operatorname{diag}(v^{-1})(.3LL^\top/4 + .7I) \operatorname{diag}(v^{-1}).$$

Its diagonal equals $.25^2$. Each regression column is computed directly from the factors, so a dense $(Kd) \times (Kd)$ covariance is unnecessary for either backend. Factor storage is distinct from input-line storage and auxiliary support-construction space.

Backends receive identical lines, with randomized execution order in a sequential one-thread run on an Intel Core i9-9900K CPU using Python 3.13.12. The single-support clock measures only support construction. Whole-coordinate-report clocks include initial selection, line preparation, every selected coordinate support, and 128-node scalar-carving inversion at $f = .6$ and error $.05/(sd)$. These constructed inputs measure computation, not sampling coverage. They exclude covariance estimation, model scoring, contrasts, Holm tests and hypervolume projection; the full application-output replay remains Table 30A with its original scope. Support and interval checks use absolute/relative tolerances 2×10^{-10} and 2×10^{-8}, respectively. Root-sweep checks additionally use the prespecified $K = 128, d = 8$ and dense controls. The support table includes their median times; dashes mark settings outside that root-sweep grid. All root-run ranges remain in the source records.

All 66 runs complete, with agreeing supports and interval endpoints. A separate post-timing audit doubles 128 to 256 quadrature nodes on all 384 coordinates in the three $K = 1024, d = 32, s = 4$ designs. The largest endpoint change is 1.54×10^{-8} target standard deviations, below the prespecified .002-SD tolerance. This refinement cost is outside the timing comparison; it checks numerical integration, not statistical coverage.

Table 38: Full single-support timing grid, including dense $s = K$ controls. Seconds are median [minimum, maximum] over three paired designs; the last column divides the all-pair median by the report-sensitive median. Both backends calculate the same support and, for whole reports, the same intervals.

K	d	s	Report-sensitive	All-pair	Root	Ratio
32	8	32	0.005 [0.005,0.005]	0.006 [0.006,0.006]	0.025	1.12
128	8	4	0.007 [0.007,0.008]	0.096 [0.094,0.106]	0.474	13.26
128	8	32	0.039 [0.037,0.041]	0.111 [0.107,0.115]	0.475	2.82
128	8	128	0.076 [0.076,0.076]	0.087 [0.087,0.087]	0.403	1.14
128	32	4	0.014 [0.014,0.014]	0.101 [0.101,0.102]	–	7.33
128	32	32	0.078 [0.076,0.079]	0.142 [0.140,0.143]	–	1.81
512	8	4	0.029 [0.029,0.034]	1.493 [1.477,1.565]	–	50.96
512	8	32	0.168 [0.167,0.176]	1.623 [1.621,1.623]	–	9.66
512	32	4	0.055 [0.053,0.057]	1.511 [1.498,1.567]	–	27.44
512	32	32	0.362 [0.362,0.367]	1.717 [1.702,1.736]	–	4.74
1024	8	4	0.058 [0.057,0.061]	5.913 [5.895,5.919]	–	102.75
1024	8	32	0.330 [0.328,0.334]	6.076 [6.064,6.137]	–	18.42
1024	32	4	0.113 [0.108,0.116]	5.934 [5.855,5.946]	–	52.56
1024	32	32	0.736 [0.729,0.740]	6.364 [6.347,6.427]	–	8.65
2048	8	4	0.115 [0.115,0.117]	23.559 [23.541,23.863]	–	204.03
2048	8	32	0.668 [0.660,0.682]	23.936 [23.929,24.060]	–	35.86
2048	32	4	0.229 [0.222,0.230]	23.532 [23.503,23.604]	–	102.67
2048	32	32	1.482 [1.480,1.513]	24.552 [24.527,24.706]	–	16.57

Table 39: All-target coordinate-report timing grid. Seconds are median [minimum, maximum] over three paired designs; the last column divides the all-pair median by the report-sensitive median. Both backends calculate the same support and, for whole reports, the same intervals.

K	d	s	Report-sensitive	All-pair	Ratio
128	8	4	0.509 [0.507,0.509]	3.320 [3.285,3.323]	6.53
512	8	4	1.184 [1.184,1.223]	47.884 [47.760,48.039]	40.44
1024	8	4	2.131 [2.126,2.137]	190.174 [190.076,190.254]	89.24
1024	32	4	14.548 [14.469,14.754]	763.258 [758.764,764.187]	52.47

O Quantitative Approximation for Full-Data Scalar Carving

Report geometry also determines sensitivity to distributional approximation. We first propagate perturbations through the scalar-integrated pivot of Proposition 3, then instantiate a finite-sample bound for bounded paired records. The same derivation specifies how a certified error envelope changes the reported intervals and tests.

O.1 Scalar-pivot and support perturbations

Standardize the selection coordinate as $W = \sqrt{f}(Y - \nu)/\sigma$, and write its complete-report support as E. The Gaussian scalar-carving CDF at u is

$$F(a, E) = \frac{\int_E \Phi(a - b_f w)\phi(w)\, dw}{\gamma(E)}, \qquad a = \frac{u - \nu}{\sigma\sqrt{1 - f}}, \quad b_f = \sqrt{\frac{f}{1 - f}}, \tag{2}$$

where γ is standard Gaussian measure. For a perturbed argument and support, put $D = \gamma(E \triangle \widetilde{E})$ and $q = \gamma(E) > 0$. Then

$$|F(a, E) - F(\widetilde{a}, \widetilde{E})| \le \frac{|a - \widetilde{a}|}{\sqrt{2\pi}} + \frac{2D}{q}. \tag{3}$$

To prove this, denote the numerators by $N, \widetilde{N}$ and let $\widetilde{q} = \gamma(\widetilde{E}) > 0$. The CDF kernel is $1/\sqrt{2\pi}$-Lipschitz in a and takes values in $[0, 1]$, giving $|N - \widetilde{N}| \le q|a - \widetilde{a}|/\sqrt{2\pi} + D$ and $|q - \widetilde{q}| \le D$. Use $\widetilde{N} \le \widetilde{q}$ when subtracting the two ratios. If $\widetilde{q} = 0$, any pivot convention in $[0, 1]$ still satisfies the bound, since $D \ge q$.

For a fixed report of size s, only $M_P = [\binom{s}{2} + s(K - s)]d$ unordered pair–coordinate comparisons are needed. On the standardized line write comparison j as $A_j + B_j w$. If coefficient changes are bounded by h_{0j}, h_{1j}, then, for every $R > 0$,

$$D \le 2\overline{\Phi}(R) + \sum_{j=1}^{M_P} \mathbb{P}\{|A_j + B_j W| \le h_{0j} + R h_{1j}\}, \qquad W \sim \mathcal{N}(0, 1). \tag{4}$$

Indeed, on $|w| \le R$ a changed sign requires proximity to its comparison boundary. Agreement of all incident signs preserves the report by Theorem 1; union bounding gives the result. Each term is a one-dimensional Gaussian interval probability. A zero slope gives a deterministic indicator. Persistent equalities may be removed when the perturbation preserves them exactly. Discrete observations at a boundary are included in the bad event, rather than discarded as Gaussian null points.

O.2 Averaging line sensitivity over the complete report

Let A_n, A_G denote the real and Gaussian-reference report events, with probabilities π_n, π_G. Write G_n, G_G for their pivots at the true target and Z_G for the Gaussian nuisance statistic. Suppose outside a bad event of probability at most r, the report indicators agree and the selected pivots satisfy

$$|G_n - G_G| \le c + \frac{2J(Z_G)}{q(Z_G)}, \qquad q(Z_G) = \mathbb{P}(A_G \mid Z_G), \quad \mathbb{E}J(Z_G) \le B,$$

where $c \ge 0$ is deterministic and $J \ge 0$ depends only on Z_G. Gaussian uniformity conditional on (Z_G, A_G) gives

$$\sup_{t \in [0,1]} |\mathbb{P}(G_n \le t \mid A_n) - t| \le \min\left\{1, \frac{c\pi_G + 2B + 2r}{\pi_n}\right\}. \tag{5}$$

For the upper bound, remove the bad event at cost r, apply the conditional uniform CDF at $t + c + 2J/q$, and average:

$$\mathbb{P}(A_n, G_n \le t) \le \mathbb{E}[q \min\{1, t + c + 2J/q\}] + r \le t\pi_G + c\pi_G + 2B + r.$$

The lower bound uses $\max\{0, t - c - 2J/q\}$. Finally $|\pi_n - \pi_G| \le r$; subtract $t\pi_n$ and divide. The event $q = 0$ has zero selected reference probability. Thus the local $1/q$ sensitivity in (3) is weighted by its probability q of being selected. The theorem needs no uniform lower bound on every nuisance-specific mass. The remaining $1/\pi_n$ factor concerns the occurrence of the complete label report, a different probability.

O.3 AN EXPLICIT BOUND FOR BOUNDED PAIRED RECORDS

Positively rescale each objective by its known range, so that records $X_i \in [0,1]^p$, $p = Kd$, are i.i.d. with mean θ and covariance Σ. Use groups of sizes $a, b \geq 2$, $n = a + b$ and $f = a/n$. For this proof use common coordinates $S = \sqrt{n}\overline{X}_A$, $T = \sqrt{n}\overline{X}_{A\cup B}$, $\mu = \sqrt{n}\theta$, and scale the selection offset accordingly. This is a change of units of both the statistic and event. Condition on a fixed offset ξ; an independent random offset can be averaged when the stated bounds hold uniformly in it. The Gaussian reference has independent group means $S_G \sim \mathcal{N}(\mu, \Sigma/f)$ and $B_G \sim \mathcal{N}(\mu, \Sigma/(1-f))$, with $T_G = fS_G + (1-f)B_G$.

Fix P and its prescribed η, and set $t_\eta = \|\eta\|_1$, $\sigma^2 = \eta^\top \Sigma \eta$. Let $v_j^2 = d_j^\top \Sigma d_j$ be the record variance of incident comparison j, where d_j subtracts its two score coordinates. Zero-variance comparisons must be preserved by the covariance estimator: $d_j^\top \widehat{\Sigma} = 0$. This holds for the unbiased sample covariance because the record difference is almost surely constant. A singular covariance or linear score redundancy is therefore allowed. Arbitrary diagonal regularization requires separate treatment when it breaks such an equality.

Choose $e, L, R > 0$ and $\beta \in (0,1)$. Let r_Σ be covariance rank, λ_+, λ_- its largest and smallest positive eigenvalues, and $\beta_3 = \mathbb{E}\|\Sigma^{+1/2}(X_i - \theta)\|_2^3$. The following quantities are explicit:

$$C_r = 11.1 + .83 \log r, \qquad r_m = \min\left\{1, \frac{C_{r_\Sigma} \beta_3 \sqrt{\lambda_+}}{e\sqrt{n}} \left(\frac{1}{f} + \frac{1}{1-f}\right)\right\}, \tag{6}$$

$$u_n = \sqrt{\frac{\log\{2(p+p^2)/\beta\}}{2n}}, \qquad h = \frac{3n}{n-1} u_n + \frac{1}{4(n-1)}. \tag{7}$$

Certified upper bounds may replace β_3 and λ_+; a certified positive lower bound may replace λ_- in the boundedness envelope. Boundedness gives $\lambda_+ \leq p/4$ and $\beta_3 \leq r_\Sigma \sqrt{p/\lambda_-}$. Define $v = t_\eta^2 h$. Given a lower bound $\sigma \geq \sigma_* > 0$ with $v \leq \sigma_*^2/2$, put

$$\begin{aligned}
R_0 &= \frac{t_\eta}{4\sigma_*^2}, \qquad R_1 = R_0 + D_\rho, \\
D_\rho &= \frac{2t_\eta h}{\sigma_*^2} + \frac{t_\eta v}{2\sigma_*^4}, \\
c &= \frac{1}{\sqrt{\pi(1-f)}} \left(\frac{t_\eta e}{\sigma_*} + \frac{Lv}{\sigma_*^2}\right). \\
H_0 &= e(1 + t_\eta R_1) + \frac{D_\rho L t_\eta}{2\sqrt{f}}, \\
H_1 &= \frac{D_\rho t_\eta/2 + R_1 v/\sigma_*}{\sqrt{f}}, \\
h_\ell &= 2(H_0 + RH_1).
\end{aligned}$$

Summing only positive-variance incident comparisons, define

$$\begin{aligned}
r_s &= \min\left\{1, \sqrt{\frac{2}{\pi}}\, 2e\sqrt{f} \sum_{j: v_j > 0} \frac{1}{v_j}\right\}, \\
B &= \min\left\{1, 2\overline{\Phi}(R) + \sqrt{\frac{2}{\pi}}\, h_\ell \sqrt{f} \sum_{j: v_j > 0} \frac{1}{v_j}\right\}, \\
r &= \min\{1, r_m + \beta + 4\overline{\Phi}(L) + r_s\}.
\end{aligned}$$

These are the terms in Theorem 2. The report probability is $\pi_n = \mathbb{P}(\widehat{\mathcal{P}}(S + \xi) = P)$; require $\pi_n \geq \pi_* > 0$. Use the full-sample unbiased covariance, a target prescribed by the complete report, and any fixed $[0,1]$ convention for an undefined pivot. For the theorem's fixed-dimensional rate, choosing $e = n^{-1/4}$, $\beta = n^{-1/2}$ and $L = R = \sqrt{2\log n}$ gives the $O(n^{-1/4})$ bound with fixed positive variance, spectral and report-probability envelopes and f bounded away from zero and one.

Proof: coupled sample means and covariance. Whiten on the positive covariance range. Equation (3.5) of Raič (2019) bounds the Wasserstein distance of an m-record standardized mean by $C_{r_\Sigma}\beta_3/\sqrt{m}$. Transforming back multiplies it by at most $\sqrt{\lambda_+}$. Couple the two independent group means independently. The coupling can be extended to the raw records conditional on their means, preserving their joint law with $\widehat{\Sigma}$. Markov's inequality gives

$$\mathbb{P}\{\|S - S_G\|_\infty > e \text{ or } \|T - T_G\|_\infty > e\} \le r_m.$$

Specifically, the two $\sqrt{n}$ group-mean expected errors are bounded by $C_{r_\Sigma}\beta_3\sqrt{\lambda_+}/(f\sqrt{n})$ and the analogous $(1-f)$ expression. Controlling both also controls their convex combination. For the covariance, Hoeffding bounds all p averages and p^2 raw second moments to error u_n with probability at least $1-\beta$. The product-of-means error is at most $2u_n$, since the means lie in $[0,1]$. Correcting the biased covariance by $n/(n-1)$ gives $\|\widehat{\Sigma} - \Sigma\|_{\max} \le h$ from (7). Independence of covariance and means is not assumed. A known estimator perturbation may be added to h, with the zero-variance preservation requirement unchanged.

Proof: propagate covariance to the selection line. On these good events, $|\widehat{\sigma}^2 - \sigma^2| \le v$, $\widehat{\sigma} \ge \sigma_*/\sqrt{2}$, and $|\widehat{\sigma} - \sigma| \le v/\sigma_*$. Since $\|\Sigma\eta\|_\infty \le t_\eta/4$, subtracting $\widehat{\rho} = \widehat{\Sigma}\eta/\widehat{\sigma}^2$ and $\rho = \Sigma\eta/\sigma^2$ gives $\|\widehat{\rho} - \rho\|_\infty \le D_\rho$, $\|\rho\|_\infty \le R_0$ and $\|\widehat{\rho}\|_\infty \le R_1$. Exclude the Gaussian tails $|U_G - \nu| > L\sigma$ or $|Y_G - \nu| > L\sigma/\sqrt{f}$, at cost at most $4\overline{\Phi}(L)$. The two kernel arguments in (2) then differ by at most $\sqrt{2\pi}c$. The standardized selection lines are

$$S + \xi + \widehat{\rho}(\nu - \eta^\top S + \widehat{\sigma}w/\sqrt{f}), \quad S_G + \xi + \rho(\nu - Y_G + \sigma w/\sqrt{f}).$$

Their coordinatewise intercept and slope errors are at most H_0, H_1: expand the intercept into the error in S, the ρ error times $\nu - Y_G$, and $\widehat{\rho}$ times the error in Y. Expand the slope into $(\widehat{\rho} - \rho)\sigma$ and $\widehat{\rho}(\widehat{\sigma} - \sigma)$; use $\sigma \le t_\eta/2$. Pairwise subtraction doubles these errors, giving h_ℓ on $|w| \le R$.

Proof: control boundaries and integrate. Outside the coupling failure, a changed observed comparison needs $|d_j^\top(S_G + \xi)| \le 2e$. Its Gaussian standard deviation is $v_j/\sqrt{f}$; the maximum density times the interval length bounds its probability by $\sqrt{2/\pi}\,2e\sqrt{f}/v_j$. Zero-variance comparisons are identical in the two experiments, so r_s bounds report changes on the coupled good event. For support changes, choose $J(Z_G)$ as the right side of (4), capped at one, with common threshold h_ℓ. Gaussian independence of Z_G and Y_G implies that $Z_G + \rho(\nu + \sigma W/\sqrt{f}) + \xi$, for fresh independent W, has the same distribution as $S_G + \xi$. The same one-dimensional density bound therefore gives $\mathbb{E}J \le B$. Apply (3) and (5), and use $\pi_G \le 1$ and $\pi_n \ge \pi_*$. This proves (1). The stated choices give $r_m, e = O(n^{-1/4})$; covariance and Gaussian-tail terms are smaller, proving the rate. □

O.4 Turning a certified error envelope into inference

Let $\Delta_q(\nu)$ be a deterministic upper bound on conditional pivot error at each candidate null value, uniform over the specified nuisance class. Retain ν if

$$\Delta_q(\nu) \ge \alpha_q/2 \quad \text{or} \quad \alpha_q/2 - \Delta_q(\nu) \le G_q(\nu) \le 1 - \alpha_q/2 + \Delta_q(\nu). \tag{8}$$

At the true parameter, each excluded tail has probability at most $\alpha_q/2 - \Delta_q + \Delta_q$, proving conditional coverage. A constant informative bound recovers usual inversion at $\alpha_q - 2\Delta_q$; a null-dependent bound can give disconnected sets, whose interval hull preserves coverage. For a raw one-sided p-value satisfying $\mathbb{P}(p_q \le t \mid P) \le t + \Delta_q$ uniformly over its composite null, $\min\{1, p_q + \Delta_q\}$ is conditionally superuniform. Indeed, its rejection event is empty below Δ_q; above it, evaluate the raw bound at $t - \Delta_q$. These corrected targets support the coordinate–Holm budgets of Section 4. If only simple-null bounds are available, composite tests require an appropriate supremum over the null, which can be conservative.

These guarantees require valid envelopes, not unchecked plug-in estimates of moments, eigenvalues or report frequency. An uncertain estimated envelope needs a separate allowance for underestimation. The formula evaluator returns an uninformative certificate when the variance premise or effective level fails; it does not use a diagnostic to choose the shortest of competing intervals.

O.5 Bounded discrete calibration with both data groups

To make finite-sample error and correction directly evaluable, consider the specified one-factor family $X_i = c + dZ_i$, with known score map, $Z_i \sim \text{Bernoulli}(p)$ and unknown $p \in (0,1)$. Its covariance is singular and estimated from the full sample; $\widehat{\rho} = d/(\eta^\top d)$, so estimation changes scale rather than rotating the line. This family isolates discreteness and scale estimation; the paired score study below additionally exercises changing supports. For the complete report let

$$\mathcal{J}_P = \{j \in \{0,\dots,a\} : \widehat{P}(c + dj/a) = P\}, \quad C_P(k) = \sum_{j \in \mathcal{J}_P} \binom{a}{j}\binom{b}{k-j},$$

where invalid count terms are zero. The support calculation supplies $\mathcal{J}_P$, with discrete endpoint membership retained. For total count $N = \sum_{i=1}^{n} Z_i$,

$$\mathbb{P}_p(N = k \mid P) = \frac{C_P(k)\exp\{k \operatorname{logit} p\}}{\sum_l C_P(l)\exp\{l \operatorname{logit} p\}}. \tag{9}$$

Thus the actual plug-in Gaussian pivot $g_{P,p}(k)$ has a finite conditional CDF obtained by summing these weights over $g_{P,p}(k) \le t$. Sorting all pivot values and checking both sides of each jump computes its exact-arithmetic $\Delta_{P,n}(p)$. The tested p is a null argument, not an unknown truth supplied to inference.

Additive correction and exact-tail calibration have distinct roles. The computed CDF error makes (8) operational for simple nulls. Near $p = 0, 1$, conditional counts concentrate and Δ becomes large; positive levels at interior nulls alone do not ensure useful interval hulls over the entire parameter space. We therefore separately evaluate exact conditional tails from (9). This is a discrete selective-law specialization, rather than a Gaussian approximation guarantee. For $p_2 > p_1$, the likelihood ratio is proportional to $\exp\{k(\operatorname{logit} p_2 - \operatorname{logit} p_1)\}$, increasing in k. The inclusive upper tail at p_0 is consequently a superuniform test of $p \le p_0$; the lower tail handles $p \ge p_0$. At the boundary null, superuniformity follows by taking the largest tail rejection segment of probability at most the cutoff; stochastic ordering extends it over the composite null. Keeping parameters for which both tails are at least $\alpha/2$ gives a connected conditional interval of coverage at least $1 - \alpha$. Both groups contribute through N. Log-space count coefficients support stable tail evaluation even for very rare reports. At $p = 0, 1$, limiting conditional laws provide numerical root brackets; ordinary conditioning is defined only when the report has positive probability. Affine target transformations reverse interval endpoints and testing directions when their slope is negative.

Complete validation design. The two candidates have score maps $(.25 + .6z, .60 - .1z)$ and $(.55 + .1z, .50 + .1z)$. The second candidate alone is reported for $z \in [.5, .6]$; both are reported outside, producing a disconnected support. Enumerate $n \in \{500, 1000, 5000\}$, $f \in \{.6, .8\}$, $p \in \{.50, .55, .60\}$ and both possible complete reports. The plug-in record variance is $n\widehat{p}(1-\widehat{p})/(n-1) + 10^{-12}$, recomputed at every possible total count. All 36 conditional laws, including probabilities as small as 1.81×10^{-10}, are retained (Table 40). The additive correction has positive levels in 15/36 marginal and 3/36 coordinate-Bonferroni cells. Its conditional miss probabilities respect the corresponding budgets; nonpositive levels return no rejection. The maximum 256/512-node discrepancy is 1.38×10^{-6}, which is a numerical check rather than an interval-arithmetic certificate.

We additionally evaluate a separately specified exact-tail extension. It retains the same 36 cells and has conditional marginal misses .0383–.0497. The table's interval widths use the conditional median count in each cell; they are not average lengths. Both marginal and coordinate-budget intervals, and one-sided null tests, meet their error allocations. This verifies non-Gaussian full-data conditional inference within the known one-factor family; the original image and Bank libraries are not assumed to follow that model.

O.6 Separating score-law and covariance-estimation effects

Using the unchanged image and Bank score pools, we independently sample 500 reports for each $n \in \{500, 1000, 5000\}$ and $f \in \{.6, .8\}$. All objectives are first rescaled to $[0, 1]$ by known ranges.

Table 40: Complete bounded Bernoulli-factor enumeration. $P = 1$ denotes the singleton second candidate; $P = 01$ retains both. π is the exact report probability at the tested parameter p. Δ is the plug-in carving pivot's conditional CDF error. Raw and KS columns give conditional miss probabilities at marginal level .05 before and after the additive correction; a dash means the effective level is nonpositive and the corrected test never rejects. Tail is the separately developed exact conditional-count calibration. W_{50} is its interval width at the conditional median count, not a mean width. All 36 cells are retained.

n	f	p	P	π	Δ	Raw	KS	Tail	W_{50}
500	0.6	0.50	1	0.5228	0.0378	0.0544	–	0.0416	0.1144
500	0.6	0.50	01	0.4772	0.0387	0.0518	–	0.0456	0.1044
500	0.6	0.55	1	0.9281	0.0207	0.0497	0.0102	0.0497	0.1006
500	0.6	0.55	01	0.0719	0.0272	0.0600	–	0.0408	0.0478
500	0.6	0.60	1	0.5217	0.0402	0.0502	–	0.0433	0.1115
500	0.6	0.60	01	0.4783	0.0414	0.0536	–	0.0413	0.1025
500	0.8	0.50	1	0.5199	0.0444	0.0578	–	0.0422	0.1302
500	0.8	0.50	01	0.4801	0.0453	0.0514	–	0.0453	0.1108
500	0.8	0.55	1	0.9608	0.0197	0.0508	0.0121	0.0383	0.1010
500	0.8	0.55	01	0.0392	0.0363	0.0551	–	0.0473	0.0386
500	0.8	0.60	1	0.5190	0.0470	0.0539	–	0.0388	0.1261
500	0.8	0.60	01	0.4810	0.0485	0.0536	–	0.0413	0.1089
1000	0.6	0.50	1	0.5163	0.0269	0.0504	–	0.0453	0.0799
1000	0.6	0.50	01	0.4837	0.0272	0.0478	–	0.0439	0.0795
1000	0.6	0.55	1	0.9877	0.0134	0.0531	0.0247	0.0452	0.0653
1000	0.6	0.55	01	0.0123	0.0241	0.0549	0.0020	0.0455	0.0268
1000	0.6	0.60	1	0.5155	0.0286	0.0464	–	0.0464	0.0780
1000	0.6	0.60	01	0.4845	0.0292	0.0474	–	0.0434	0.0784
1000	0.8	0.50	1	0.5141	0.0315	0.0557	–	0.0481	0.0914
1000	0.8	0.50	01	0.4859	0.0320	0.0505	–	0.0403	0.0876
1000	0.8	0.55	1	0.9960	0.0132	0.0488	0.0248	0.0488	0.0647
1000	0.8	0.55	01	0.0040	0.0326	0.0573	–	0.0455	0.0223
1000	0.8	0.60	1	0.5134	0.0334	0.0519	–	0.0444	0.0887
1000	0.8	0.60	01	0.4866	0.0341	0.0504	–	0.0463	0.0861
5000	0.6	0.50	1	0.5073	0.0121	0.0512	0.0267	0.0488	0.0354
5000	0.6	0.50	01	0.4927	0.0121	0.0499	0.0259	0.0477	0.0356
5000	0.6	0.55	1	1.0000	0.0059	0.0498	0.0380	0.0482	0.0278
5000	0.6	0.55	01	3.26×10^{-08}	0.0519	0.0525	–	0.0478	0.0101
5000	0.6	0.60	1	0.5069	0.0129	0.0487	0.0239	0.0464	0.0347
5000	0.6	0.60	01	0.4931	0.0130	0.0491	0.0241	0.0473	0.0349
5000	0.8	0.50	1	0.5063	0.0141	0.0515	0.0211	0.0483	0.0405
5000	0.8	0.50	01	0.4937	0.0142	0.0494	0.0201	0.0494	0.0406
5000	0.8	0.55	1	1.0000	0.0059	0.0498	0.0380	0.0482	0.0278
5000	0.8	0.55	01	1.81×10^{-10}	0.0690	0.0518	–	0.0485	0.0146
5000	0.8	0.60	1	0.5060	0.0150	0.0494	0.0197	0.0462	0.0397
5000	0.8	0.60	01	0.4940	0.0152	0.0489	0.0195	0.0471	0.0400

The independent selection offset is $10^{-3}G/(2\sqrt{a})$; the target is the first objective of the lowest-index reported label, a rule depending only on P. For each real sample, compute scalar pivots with full-pool covariance and with the sample covariance. A separate moment-matched Gaussian experiment uses the same two covariance modes. Within each source, both modes retain the identical complete report. Gaussian group means and within/between-group Wishart scatter generate the same sufficient-statistic distribution as explicitly sampling all Gaussian records. Population covariance and mean enter only these diagnostics.

All 6,000 paired replications complete, giving 24,000 pivot evaluations (Table 41). The 12,000 oracle/estimated pairs satisfy (3) within the fixed 4×10^{-5} tolerance; the bound saturates at one in 1,859 pairs. The first ten replications per cell receive doubled-node checks, covering 480 pivots with maximum difference 1.23×10^{-7}. Every observed label group and its count is retained in the source data, with pointwise binomial uncertainty.

Table 41: Scalar-carving source/covariance diagnostic, 500 independent replications per cell. Entries are marginal miss counts for one label-prespecified coordinate per report; they are not FCR or simultaneous-error counts. R/G denote real paired records/moment-matched Gaussian statistics, O/E population/estimated covariance. D_R, D_G average absolute oracle–estimated pivot changes on the same report. Each source retains every complete label set; oracle quantities are diagnostic inputs only.

Pool	n	f	R–O	R–E	G–O	G–E	D_R	D_G
Bank	500	0.6	36	41	23	22	0.0832	0.0396
Bank	500	0.8	38	37	11	11	0.1570	0.0561
Bank	1000	0.6	31	34	23	24	0.0857	0.0277
Bank	1000	0.8	54	47	22	20	0.1389	0.0320
Bank	5000	0.6	31	27	20	21	0.0328	0.0117
Bank	5000	0.8	26	27	17	17	0.0387	0.0107
CIFAR	500	0.6	31	28	26	26	0.0755	0.0258
CIFAR	500	0.8	29	40	23	22	0.0735	0.0350
CIFAR	1000	0.6	29	30	29	27	0.0369	0.0162
CIFAR	1000	0.8	36	33	21	21	0.0531	0.0149
CIFAR	5000	0.6	29	28	24	28	0.0107	0.0030
CIFAR	5000	0.8	25	26	28	27	0.0176	0.0030

Bank's population- and estimated-covariance modes miss 216 and 213 times among 3,000 target-wise reports; the known-Gaussian control misses 116 times. This comparison separates score-law effects from covariance estimation (Table 41). At the present application scales, the explicit certified envelope remains conservative and gives the trivial bound even in the oracle-assisted $n = 5000$ diagnostic. That diagnostic uses $\pi_* = 1$, the largest coordinate variance, one incident comparison and 81 coupling radii from 10^{-5} to 10^3 to assess conservatism. The application intervals therefore use the Gaussian approximation; Appendix N.2 provides an independent finite-sample bounded-score alternative.

Table 42: Exact label sets used in the independent confirmation. Indices are zero-based within the original 21-model image and 16-model Bank pools, not the new 1,024-configuration timing library.

Report	Complete labels
CIFAR 1	$\{14, 15, 16, 19, 20\}$
CIFAR 2	$\{14, 15, 16, 18, 19, 20\}$
CIFAR 3	$\{15, 16, 19, 20\}$
Bank 1	$\{1, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15\}$
Bank 2	$\{0, 1, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15\}$
Bank 3	$\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15\}$

Table 43: Independent complete-report confirmation, 2,000 accepted draws per frozen label set. Each entry is the number of joint-error reports / multiplicity-adjusted upper bound (%). Coordinates and Holm comparisons share .025/.025. Bounds are two-sided Clopper–Pearson intervals with error .05/72 across the prespecified group–method–endpoint family. $\widehat{\pi}_P$ uses a separate 10,000-draw frequency table per pool.

Report	$\widehat{\pi}_P$ (%)	Carving	Splitting	Global	LSI
CIFAR 1	65.42	32 / 2.79	57 / 4.34	21 / 2.06	21 / 2.06
CIFAR 2	11.85	43 / 3.48	51 / 3.97	23 / 2.20	25 / 2.33
CIFAR 3	2.00	51 / 3.97	63 / 4.70	43 / 3.48	48 / 3.79
Bank 1	7.50	28 / 2.53	20 / 2.00	47 / 3.73	36 / 3.04
Bank 2	3.03	34 / 2.91	29 / 2.59	58 / 4.40	42 / 3.42
Bank 3	1.07	26 / 2.40	17 / 1.79	61 / 4.58	47 / 3.73

P Independent Real-Report Confirmation and a Thousand-Candidate Library

P.1 Reliability conditional on an exact published shortlist

This confirmation tests the realized-report reliability question on the original 21-model image and 16-model Bank score pools. Group selection uses only label frequencies from the previous resource experiment: among reports observed at least ten times, order by decreasing frequency with a lexicographic tie-break, and select the first, middle and last distinct groups. The six sets are frozen before the new draws (Table 42). A separate 10,000-selection-draw table per pool estimates their probabilities; these estimates are independent of the confirmation outcomes.

For each group, rejection sampling draws the 3,000-record selection sample and its independent fixed-range offset until the complete label set matches. Only then are 2,000 independent held-out records and independent Gaussian calibration banks drawn. Each group contains 2,000 accepted replications, divided into four independent streams of 500. Acceptance and stopping depend on labels and the fixed accepted count, not on interval coverage or tests. Carving, splitting, global max-T and LSI receive the same paired records. Pool means are used only to score coverage and the truth of comparisons.

The primary report uses the joint .025/.025 coordinate–comparison budget. Two secondary endpoints use separate .05 coordinate bands and .05 comparison families. These secondary procedures are not combined into a .05 joint report. We retain all six groups, four methods and three endpoints; the source tables contain pointwise Clopper–Pearson intervals and simultaneous intervals allocating .05/72 to each group–method–endpoint combination. For each group, the first 40 draws also receive full interval inversion and the first eight receive 512/1024-node numerical checks. An independent replay reconstructs 18 draws under both budgets using the original complete inference pipeline.

Results within the frozen label sets. All 12,000 replications complete without numerical failure. Carving's joint error estimates range from .013 to .0255, and its separate .05 coordinate simultaneous coverage ranges from .9580 to .9765. For the primary joint endpoint, all 24 method–group multiplicity-adjusted Monte Carlo upper bounds are below .05; the largest is .04697. Thus this confirmation supports reliability within these tested reports for all four procedures.

Table 44: Separate .05 coordinate procedures: empirical simultaneous coverage within each exact label set. These budgets are distinct from the joint reports in Table 43. Complete pointwise and multiplicity-adjusted Monte Carlo intervals accompany the source CSV.

Report	Carving	Splitting	Global	LSI
CIFAR 1	0.9645	0.9540	0.9825	0.9780
CIFAR 2	0.9680	0.9655	0.9770	0.9725
CIFAR 3	0.9580	0.9545	0.9630	0.9575
Bank 1	0.9765	0.9790	0.9535	0.9560
Bank 2	0.9755	0.9785	0.9495	0.9530
Bank 3	0.9760	0.9830	0.9445	0.9495

Table 45: Mean true comparisons per conditional report under the joint .025/.025 allocation. Global and LSI use the same prespecified global contrast family; LSI localizes coordinate bands.

Report	Carving	Splitting	Global	LSI
CIFAR 1	11.890	11.162	11.685	11.685
CIFAR 2	12.870	12.118	12.728	12.728
CIFAR 3	8.768	8.220	8.735	8.735
Bank 1	14.324	12.675	15.360	15.360
Bank 2	14.489	12.608	15.113	15.113
Bank 3	15.185	13.835	16.299	16.299

Carving supports more true comparisons than matched splitting in all six exact-report groups, while relative yield versus unconditional procedures varies by pool (Table 45). LSI uses the same global contrast family as global max-T, explaining their identical claim counts. The checked groups account for estimated selection probability .7927 in the image pool and .1160 in Bank. Equal group sizes estimate conditional risks, not an unconditional average or reliability of unobserved reports. These finite-pool findings complement the controlled Gaussian separation in Appendix N.

P.2 A FROZEN LIBRARY OF 1,024 GENUINELY TRAINED CONFIGURATIONS

The computational study retains the Bank training/evaluation partition, preprocessing, excluded call-duration feature and five score definitions from Appendix K. All 1,024 frozen configurations are trained. Fitting and preprocessing use only training rows; evaluation labels are used only for scoring. Four families each supply 256 configurations:

- Logistic regression crosses eight positive-class weights $\{1, 1.5, 2, 3, 4, 6, 8, 12\}$, eight C values from 10^{-3} to 10^{4} and training fractions $\{.4, .6, .8, 1\}$; the iteration cap is 2,000.
- Histogram gradient boosting crosses the same weights, maximum leaf counts $\{7, 15, 31, 63\}$, learning rates $\{.03, .1\}$ and ℓ_2 penalties $\{0, 1, 10, 100\}$, with 100 iterations and no early stopping.
- Random forests and extremely randomized trees each cross the same weights, depths $\{4, 8, 16, \text{unlimited}\}$, minimum leaf sizes $\{1, 5, 20, 50\}$ and feature rules $\{\sqrt{p}, .8p\}$, using 64 trees.

The reference is full-training, unweighted logistic regression with $C = 1$. A frozen family-balanced order defines nested libraries of $K \in \{64, 128, 256, 512, 1024\}$. No candidate is filtered by evaluation performance. Distinct configurations and training seeds produce 1,024 distinct probability vectors and 1,002 distinct binary prediction vectors. These are configurations from four model families; the five scores retain their original dependencies, including accuracy as the sum of the two correctness yields.

The complete array contains 77,880,320 finite scores. Independent recalculation of every score from saved probabilities and labels agrees within 1.11×10^{-16}. Reloading every fitted model on 257 fixed rows reproduces its probabilities exactly. All frozen parameters, training-row identities, model and prediction hashes, and warnings are retained; every model completes without a training warning or failure. Training/scoring costs are recorded separately from the inference benchmark. This library

extends the computational test to real trained candidates; statistical calibration remains a separate question from agreement of exact support backends on the supplied plug-in Gaussian model.

P.3 Paired complete-output clocks and implementation attribution

Three frozen resampling streams each provide 3,000 selection and 2,000 held-out records. The same sampled row indices and offsets are used at all five nested library sizes. Backend order is randomized within each pair; library sizes run in ascending order. Both backends compute every selected coordinate interval and every prescribed candidate–reference contrast, Holm decisions, coordinate-precision flags and hypervolume bounds. Every target starts with a 512/1024-node check at .002 target standard deviations. If adjacent-order endpoints differ more, both backends follow the same doubling sequence up to 4096/8192 nodes. The output uses the lower order of the first accepted pair; all attempts are timed, and an unresolved target retains its conservative fallback and failure flag. This is a convergence diagnostic, rather than a rigorous quadrature-error bound. Output agreement also requires matching integration orders, labels, support unions, intervals, decisions, precision flags and summaries.

Both support constructors use one compiled float64 interval/sweep kernel with fast-math disabled. The permitted domination sources are P or $[K]$; all remaining operations are shared. Both use exactly the required sample-covariance columns and the original common variance ridge. This avoids forming unused columns without introducing a low-rank approximation. Separate checks against the original dense covariance and three Python support constructions precede timing.

Three clocks distinguish support construction, scores-to-complete-report and raw-rows-to-complete-report. The latter recomputes preprocessing, predictions and scores on all 5,000 sampled rows for each backend; it does not read cached predictions within the clock. Scores-to-report includes selection, means, estimated covariance, support, inversion, numerical refinement and all summaries. Training, model loading, two-row prediction warm-up and compilation are recorded separately. Numerical thread pools are fixed at one on an Intel Core i9-9900K CPU, with no other experiment running. The completed timings describe inference from fitted models, including fresh evaluation.

Shared summary implementation. Both backends use the same compiled coordinate-slab hypervolume routine. It agrees with a recursive reference and independent inclusion–exclusion on 140 continuous/discrete small instances, and reproduces both endpoints of seven saved complete reports. A complete-report check also matches the reference inference pipeline. All 15 planned pairs use the same final implementation and adaptive refinement rule. The within-implementation comparison isolates the contribution of reported-source reconstruction; summary computation and numerical precision requirements are shared.

Measured scale and common costs. All 15 paired reports pass the numerical and output checks (Tables 46–47), comprising 14,647 target instances computed by each backend. Thirteen paired target instances use 1024/2048 nodes; all others pass at 512/1024. The largest accepted adjacent-order change is .001978 SE. The independent audit reconstructs sampling streams and reports, binds target records to aggregate outputs, and verifies paired supports, numerical work and decisions.

At $K = 1024$, the three reports have sizes 180, 234 and 190. Median report-sensitive/all-pair times are 58.06/128.94 seconds from scores to the complete report and 74.21/145.08 seconds including fresh prediction and scoring. Median paired ratios are 4.84 for support construction, 2.22 for scores-to-report and 1.95 for raw-rows-to-report. The last ratios range from 1.89 to 1.96 across the three reports. At $K = 64$ and 128, median complete-pipeline ratios are about 1.01; at $K = 512$ the ratio is 1.26. These stage-specific measurements explain why a faster support constructor gives a smaller full-pipeline gain when inference and scoring are common costs.

Table 46: Complete paired timing records (seconds), part 1. R: report-sensitive; A: all-pair. Support sums every prescribed target. Total includes raw-row prediction and scoring; score-to-report also includes common preparation and all report summaries. Checks require every target refinement and all output-agreement conditions.

K	Rep	s	Method	Targets	Support	Score-to-report	Total	Checks
64	0	40	R	356	0.089	8.02	9.09	yes
64	0	40	A	356	0.112	8.13	9.21	yes
64	1	38	R	338	0.086	7.36	8.44	yes
64	1	38	A	338	0.107	7.44	8.51	yes
64	2	36	R	320	0.078	7.29	8.36	yes
64	2	36	A	320	0.099	7.27	8.33	yes
128	0	61	R	545	0.320	12.24	14.43	yes
128	0	61	A	545	0.517	12.39	14.57	yes
128	1	59	R	527	0.303	12.09	14.27	yes
128	1	59	A	527	0.502	12.49	14.69	yes
128	2	55	R	491	0.273	11.55	13.72	yes
128	2	55	A	491	0.465	11.71	13.87	yes
256	0	90	R	810	1.184	18.99	23.19	yes
256	0	90	A	810	2.749	20.60	24.79	yes
256	1	100	R	896	1.414	21.88	26.05	yes
256	1	100	A	896	3.027	23.45	27.65	yes
256	2	100	R	900	1.404	21.70	25.87	yes
256	2	100	A	900	3.043	23.49	27.66	yes

Table 47: Complete paired timing records (seconds), part 2. R: report-sensitive; A: all-pair. Support sums every prescribed target. Total includes raw-row prediction and scoring; score-to-report also includes common preparation and all report summaries. Checks require every target refinement and all output-agreement conditions.

K	Rep	s	Method	Targets	Support	Score-to-report	Total	Checks
512	0	127	R	1143	4.328	30.84	39.04	yes
512	0	127	A	1143	14.947	41.57	49.79	yes
512	1	165	R	1481	6.978	41.28	49.55	yes
512	1	165	A	1481	19.442	53.87	62.06	yes
512	2	156	R	1404	6.346	38.39	46.59	yes
512	2	156	A	1404	18.414	50.70	58.91	yes
1024	0	180	R	1620	16.636	55.56	71.70	yes
1024	0	180	A	1620	84.532	123.31	139.52	yes
1024	1	234	R	2106	27.252	76.86	93.16	yes
1024	1	234	A	2106	110.254	160.34	176.43	yes
1024	2	190	R	1710	18.452	58.06	74.21	yes
1024	2	190	A	1710	89.270	128.94	145.08	yes