TEST-TIME FEATURE ALIGNMENT UNDER UNCERTAIN CORRESPONDENCE

Anonymous authors
Paper under double-blind review

Anonymous code and artifacts

ABSTRACT

Feature alignment can improve a frozen classifier's accuracy under test-time distribution shift without establishing whether shifted examples recover the predictions made on their clean counterparts. This work asks whether prediction recovery can be assessed without knowing which shifted example corresponds to which clean example. It introduces an assignment framework that considers possible matchings subject to a cost constraint. With unpaired features, the framework bounds prediction disagreement relative to a baseline matching. Independent paired calibration provides simultaneous recovery-error bounds for a new batch drawn under the same conditions as the calibration batches. These bounds can certify that one method better recovers clean predictions than another without identifying the true matching. Across natural ImageNet shifts and ImageNet-C, substantial ambiguity remains despite accuracy gains. In paired ImageNet-C experiments, every certified ordering agreed with the hidden pairs. A randomized-cost control preserved observed coverage but could no longer distinguish methods, showing that coverage alone does not guarantee informative bounds.

1 INTRODUCTION

The setting is test-time feature alignment with a frozen encoder f and classifier h. Normalization- and covariance-based corrections can improve performance under distribution shift (Nado et al., 2020; Schneider et al., 2020; Sun et al., 2016). However, matching reference and target feature distributions does not determine which shifted example corresponds to which clean example (Moriakov et al., 2020; Gulrajani & Hashimoto, 2022). Marginal recovery, examplewise feature recovery, and prediction recovery are therefore distinct. A shifted example exhibits prediction recovery when the corrected classifier reproduces the frozen classifier's prediction on its clean counterpart. The central questions are what can be established about prediction recovery, and whether alignment methods can be compared, without observing the clean–target matching.

A possible correspondence is a one-to-one assignment between a clean reference batch and a shifted target batch. A feature-space cost and budget select the allowed assignments. Each assignment preserves the two observed endpoint multisets and changes only their correspondence. The available guarantees depend on whether paired calibration data are available.

With unpaired features, the minimum-cost assignment is an operational baseline, rather than ground truth. The analysis maximizes the fraction of its assigned clean predictions that can change under another allowed assignment. The resulting *task-conditioned ambiguity profile* (subsequently, ambiguity profile) quantifies what the endpoints, readout, cost, and budget leave unresolved. It does not estimate the true matching or target error.

With independent paired calibration, true-pair costs from calibration batches set one common threshold for a future batch whose pair identities remain hidden during bound construction. A split-conformal rank argument (Shafer & Vovk, 2008; Johnstone & Cox, 2021) makes inclusion of the future pairing a single simultaneous event for a fixed family of corrections. Each method's recovery error can then be bounded, and the difference between two methods' errors can be maximized over the same correspondence. This common optimization can certify an ordering even when separately optimized error ranges overlap.

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.

Each worst-case objective is bracketed computationally. A feasible assignment gives a lower bound, and weak Lagrangian duality gives an upper bound. Standard LP duality and Birkhoff–von Neumann integrality yield the relaxed optimum; fixed-multiplier subproblems are linear assignments (Birkhoff, 1946; Kuhn, 1955). Their gap measures numerical optimization uncertainty rather than statistical uncertainty.

Prior work establishes that distribution-matched maps can be non-unique and can induce different downstream predictions (Moriakov et al., 2020; Gulrajani & Hashimoto, 2022). The analysis instead concerns a finite, cost-constrained assignment set on an observed batch and provides computable prediction-level bounds without claiming population-map identification.

Across the three natural shifts and ImageNet-C (Hendrycks et al., 2021; Wang et al., 2019; Recht et al., 2019; Hendrycks & Dietterich, 2019), the shift-level mean ambiguity at the primary budget ranges from 70.49% to 78.73% (Section 4.2). These ambiguity levels coexist with Bures accuracy gains of 3.40 points on the natural shifts and 4.83 points on ImageNet-C. In the paired ImageNet-C study, all 263 issued orderings agree with the hidden pairs. Separate methodwise ranges issue 220 orderings. Common signed upper bounds certify 33 additional orderings on 32 batches. Feasible extrema witnesses establish that separate ranges cannot certify those cases. Ten cases in the raw 43-ordering difference remain numerically unresolved. Those 32 batches retain a mean anchored-ambiguity bracket of 16.54–16.99%. A method comparison can therefore be resolved while the correspondence and individual errors remain uncertain. Randomly reassigning costs to reference endpoints and recalibrating retains observed coverage, but makes every absolute upper bound 100% and issues no ordering. Coverage alone is therefore insufficient for an informative comparison.

The contributions are:

- **Recovery under unknown correspondence.** A finite, cost-constrained assignment model yields an unpaired prediction-ambiguity profile relative to a minimum-cost baseline.
- **Calibrated common-correspondence comparisons.** Independent paired batches calibrate one assignment set for a future exchangeable batch. Signed bounds compare methods on the same correspondence and can be strictly more informative than separate error ranges.
- **Computational certificates and controls.** Feasible assignments and Lagrangian bounds bracket each worst case. Experiments separate correspondence ambiguity, accuracy gains, observed coverage, and informativeness.

2 One correspondence set for ambiguity and recovery

Finite endpoint model. Consider clean reference features $r_1, \dots, r_n$, shifted target features $t_1, \dots, t_n$, and a frozen readout $h : \mathbb{R}^d \to \{1, \dots, K\}$. A permutation $\pi \in S_n$ represents the correspondence $t_i \leftrightarrow r_{\pi(i)}$. Under the latent true permutation $\pi^\star$, prediction recovery compares $h(T(t_i))$ with $h(r_{\pi^\star(i)})$. Ground-truth accuracy is a separate estimand. The population-map motivation, moment-matching identities, and decision-theoretic results are given in Sections A.1 to A.3.

Each indexed atom $\bar{r}_j = (j, r_j)$ and $\bar{t}_i = (i, t_i)$ receives mass $1/n$. Fix an anchor $\pi_0 \in S_n$ and finite assignment costs $c_{ij} \geq 0$. Define

$$C(\pi) = \sum_i c_{i,\pi(i)}, \qquad D_{h,\pi_0}(\pi) = \frac{1}{n} \sum_i \mathbf{1}\{h(r_{\pi_0(i)}) \neq h(r_{\pi(i)})\}, \tag{1}$$

$$\Gamma_{h,\pi_0}(B) = \max_{\pi \in S_n :\, C(\pi) \leq B} D_{h,\pi_0}(\pi), \qquad B \geq C(\pi_0). \tag{2}$$

The quantity Γ_{h,π_0} is called the *anchored task-conditioned correspondence-ambiguity profile* (subsequently, ambiguity profile). It depends on the endpoint batch, anchor, readout, cost, and budget. It is therefore not a property of the marginals alone.

This finite profile is motivated by a two-world decision bound. If two observationally indistinguishable correspondences imply predictions that disagree on a fraction δ of examples, no world-independent prediction rule can guarantee agreement above $1 - \delta/2$ in both worlds (Gulrajani & Hashimoto, 2022; Le Cam, 1986). Admissibility of both inverse corrections is sufficient for the attainment construction. The upper bound does not require it. The formal result, Gaussian example, and anchor-sensitivity bounds appear in Sections A.3, A.4 and A.11.

Proposition 1 (Finite-instance ambiguity certificate). *For any feasible $\widehat{\pi}$ and any $\lambda \geq 0$, let*

$$U_\lambda(B) = \lambda B + \max_{\pi \in S_n} \{D_{h,\pi_0}(\pi) - \lambda C(\pi)\}. \tag{3}$$

Then

$$D_{h,\pi_0}(\widehat{\pi}) \leq \Gamma_{h,\pi_0}(B) \leq \Gamma^{\mathrm{LP}}_{h,\pi_0}(B) = \min_{\eta \geq 0} U_\eta(B) \leq \min\{1, U_\lambda(B)\}, \tag{4}$$

where Γ^{LP} is the budgeted doubly-stochastic relaxation. For fixed λ, the maximum in (3) is a linear-assignment problem.

A feasible assignment gives the lower bound, weak Lagrangian duality gives the upper bound, and standard LP strong duality with Birkhoff–von Neumann integrality gives the equality. For fixed multipliers, the remaining subproblem is a linear assignment (Birkhoff, 1946; Kuhn, 1955). The exact proof, explicit LP dual, and solver-level verification protocol for the empirical bounds appear in Sections A.4 and G.

Paired calibration of method-specific recovery error. The assignment formulation also yields recovery bounds when paired calibration data are available. In this paired setting, each reference batch contains the clean counterpart of every target example. Pair identities are observed in calibration batches and withheld while evaluation bounds are constructed. The unpaired ambiguity analysis does not require counterpart containment. In batch b, let $\pi^\star_b$ be the true assignment and define the normalized score

$$S_b(\pi) = \frac{C_b(\pi) - C_b^{\min}}{\overline{C}_b - C_b^{\min}}, \qquad \mathcal{P}_b(q) = \{\pi : S_b(\pi) \leq q\}, \tag{5}$$

where $C_b^{\min} = \min_\pi C_b(\pi)$ and $\overline{C}_b = \mathbb{E}_{\pi \sim \mathrm{Unif}(S_n)} C_b(\pi)$. The score is defined when $\overline{C}_b > C_b^{\min}$, and the protocol verifies this condition for each calibration and future batch before issuing a certificate. For correction a, define its recovery error under assignment π by

$$E_{a,b}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{h(T_a(t_{b,i})) \neq h(r_{b,\pi(i)})\}. \tag{6}$$

Then $G_{a,b}(q) = \max_{\pi \in \mathcal{P}_b(q)} E_{a,b}(\pi)$. Every method in a fixed finite family shares this one feasible set. This finite endpoint condition does not assume an invertible population corruption map.

Proposition 2 (Paired-calibrated recovery). *Suppose the corrections and common cost geometry are fitted on data disjoint from m calibration batches and one future batch. Conditional on the fitted objects and locked protocol, assume that the $m+1$ true-pair scores are well defined and exchangeable. For $\alpha \in (0,1)$, set $k = \lceil (m+1)(1-\alpha) \rceil$. Let $\widehat{q}$ be the kth calibration score when $k \leq m$, and set $\widehat{q} = +\infty$ otherwise. Then, with probability at least $1-\alpha$, every method in any fixed, finite, predeclared family simultaneously satisfies*

$$E_{a,m+1}(\pi^\star_{m+1}) \leq G_{a,m+1}(\widehat{q}) \leq U_{a,m+1}, \tag{7}$$

where $U_{a,m+1}$ is any valid Lagrangian upper bound on $G_{a,m+1}(\widehat{q})$. All methods share the same true-pair feasibility event, so no methodwise union bound is required.

Let L_a denote any feasible lower bound. Then $[L_a, U_a]$ brackets the optimized worst-case error G_a. On the calibrated feasibility event, every realized method error obeys $E_a(\pi^\star) \leq G_a \leq U_a$. The probability statement is marginal over the calibration batches and one future batch, conditional on the fit. The formal statement and proof are in Proposition 3. The rank argument follows standard conformal principles (Shafer & Vovk, 2008; Johnstone & Cox, 2021). The experiments apply this result separately within each corruption condition. For a predeclared ordered pair (a, c), define the signed worst case over the same feasible set,

$$H_{a,c,b}(q) = \max_{\pi : S_b(\pi) \leq q} \{E_{a,b}(\pi) - E_{c,b}(\pi)\}. \tag{8}$$

Unlike two separate absolute maxima, this objective compares the methods at one common correspondence.

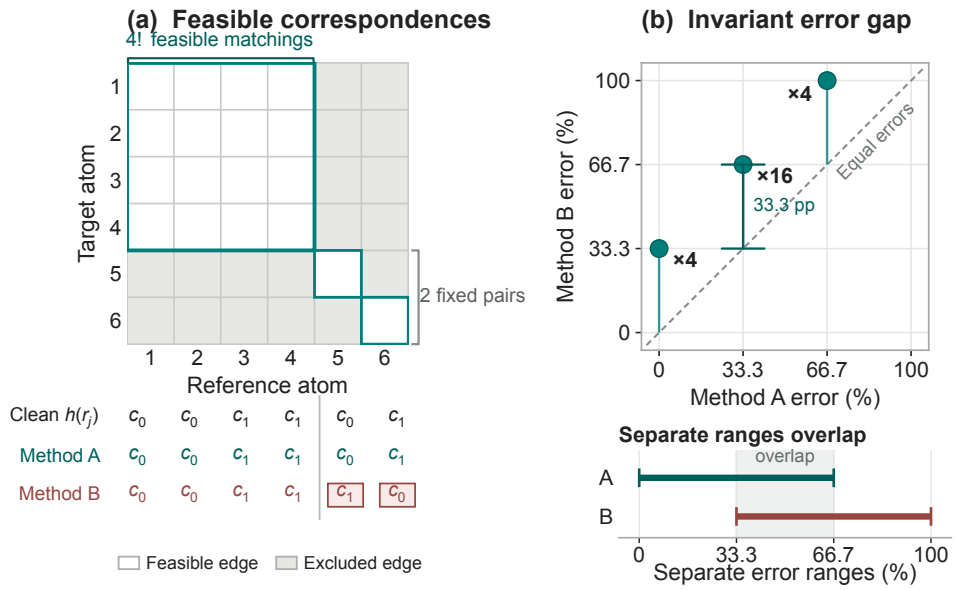


Figure 1: **A common correspondence identifies an ordering that separate ranges do not.** This exact $m = 4, k = 2$ instance of Proposition 4 is illustrative, not empirical. **(a)** At zero assignment-cost budget, the first four target atoms may match any permutation of the first four reference atoms, while the last two pairs are fixed. The methods make identical predictions on the uncertain block; Method B reverses the two fixed predictions, one from each class. This balanced illustrative choice creates a constant two-error offset while the uncertain block retains 24 possible correspondences. Symbols c_0 and c_1 denote the two classifier outputs. **(b)** Exact enumeration gives three error pairs with counts 4, 16, and 4. Every common matching has $E_B - E_A = 1/3$, although the separate ranges $E_A \in [0, 66.7]\%$ and $E_B \in [33.3, 100]\%$ overlap. Multiplicities are enumeration counts, not probabilities.

To make that distinction explicit, write $\mathcal{P}_b(q) = \{\pi : S_b(\pi) \le q\}$ and define the exact methodwise extrema

$$\ell_{a,b}(q) = \min_{\pi \in \mathcal{P}_b(q)} E_{a,b}(\pi), \qquad u_{a,b}(q) = \max_{\pi \in \mathcal{P}_b(q)} E_{a,b}(\pi). \tag{9}$$

For every $\pi \in \mathcal{P}_b(q)$, $E_{a,b}(\pi) \le u_{a,b}(q)$ and $E_{c,b}(\pi) \ge \ell_{c,b}(q)$. Consequently,

$$H_{a,c,b}(q) \le u_{a,b}(q) - \ell_{c,b}(q). \tag{10}$$

Equality holds exactly when one feasible assignment jointly attains the maximum error of a and the minimum error of c. Thus separate methodwise ranges can lose information by combining extrema from different correspondences. Here $\ell_{a,b}$ is an exact minimum and is unrelated to the symbol L_a used above for a feasible lower certificate on the maximum G_a.

Figure 1 makes the loss in (10) explicit. Separate optimization permits overlapping error ranges, whereas evaluating both methods under each common matching fixes their signed difference. The construction is the $m = 4, k = 2$ case of Proposition 4.

Corollary 1 (Signed common-correspondence comparison)**.** *Under Proposition 2, fix before calibration a finite set $\mathcal{C}$ of ordered method pairs. If every $\overline{H}_{a,c,m+1}$ is a valid upper bound on $H_{a,c,m+1}(\widehat{q})$ over that same common set, then, with probability at least $1 - \alpha$, every $(a, c) \in \mathcal{C}$ simultaneously satisfies*

$$E_{a,m+1}(\pi^\star_{m+1}) - E_{c,m+1}(\pi^\star_{m+1}) \le \overline{H}_{a,c,m+1}. \tag{11}$$

Thus a negative upper bound certifies strictly lower error for a, whereas a nonpositive bound certifies no larger recovery error.

The corollary inherits the calibration event and one-future-batch scope of Proposition 2. Its proof, numerical-verification details, and an unevaluated batchwise identity-fallback consequence appear in

Sections A.5 and G. The result does not claim joint coverage over repeated batches or an unpaired deployment guarantee.

Appendix A.4 gives a worked ambiguity example, and Proposition 4 constructs a case in which a strict method ordering is identified although the correspondence is not. Multiworld, output-mass, and approximately distinguishable extensions are also deferred to Sections A.6, A.7, A.9 and A.10.

3 Experimental design

Evidence map and frozen models. Three studies answer different questions. The marginal benchmark measures accuracy and clean-prediction agreement. The unpaired profiles measure correspondence ambiguity. The paired study uses separate fitting, calibration, and evaluation data to assess recovery bounds. Each guarantee is limited to its corresponding data setting. Features are extracted from CLIP ViT-L/14 ($d = 768$), DINOv3 ViT-L/16 ($d = 1{,}024$), and AugReg ViT-B/16 ($d = 768$) (Radford et al., 2021; Siméoni et al., 2025; Dosovitskiy et al., 2021; Steiner et al., 2022). For each encoder–dataset pair, a multinomial logistic classifier is fitted once on clean data and then frozen. The custom ImageNet head uses 25 clean validation images per class. Those identities are excluded from all later data. Exact checkpoints, splits, and sample counts are in Section B.1.

Alignment information. The corrections use different feature statistics. Identity uses no target statistic. Mean, diagonal, and coordinate-quantile alignment use the corresponding marginal statistics. CORAL and Bures use joint means and OAS-shrunk covariances (Sun et al., 2016; Chen et al., 2010; Givens & Shortt, 1984; Gelbrich, 1990; Bhatia et al., 2019). Marginal methods receive independently sampled reference and target sets without assignments. Matched and deliberately shuffled assignments are reserved for paired diagnostics.

Marginal alignment and unpaired profiles. The primary marginal-alignment study uses five split seeds and disjoint calibration and evaluation sets. It reports top-1 accuracy and agreement with the frozen classifier's clean prediction, with corruption as the inference unit. The benchmark spans 19 CIFAR-100-C and 15 ImageNet-C corruptions at severities 1, 3, and 5. Clean–corrupt identities are used only for diagnostic evaluation. At severity 5, split and encoder measurements are first averaged within corruption. Inference uses 20,000 percentile-bootstrap draws for 95% intervals, exact cluster sign-flip tests, and Holm correction over ten contrasts (Davison & Hinkley, 1997; Lehmann & Romano, 2005; Holm, 1979). The intervals have a population interpretation under the working assumption that the observed corruptions represent draws from a conceptual corruption population. Without that assumption, they are fixed-panel resampling summaries. Exact sign-flip validity requires clusterwise sign invariance. Additional frozen-alignment benchmarks and streaming diagnostics are reported in Sections B and C.

The unpaired panel covers ImageNet-R, ImageNet-Sketch, ImageNet-V2, and severity-5 ImageNet-C (Hendrycks et al., 2021; Wang et al., 2019; Recht et al., 2019; Hendrycks & Dietterich, 2019). For CLIP and DINOv3, each split's reference and target endpoints each contribute 3,840 alignment-fitting examples and 1,024 profile examples. After Bures fitting, the anchor is the minimum-cost matching. A high-disagreement feasible assignment is obtained using greedy swaps and Lagrangian assignment subproblems. All assignments act on indexed atoms and preserve both endpoint multisets. Frozen predictions define the disagreement reward. External labels, row-level WNIDs, semantic identities, and clean–target pairs are excluded from alignment fitting and assignment optimization. For profile unit u with feature dimension d_u, the primary post-Bures cost is

$$c_{ij}^{(u)} = d_u^{-1} \|\widehat{\Sigma}_{r,u}^{-1/2}(\widehat{T}_{B,u}(t_i) - r_j)\|_2^2, \qquad B_{q,u} = C_{\min,u} + q(C_{\mathrm{rand},u} - C_{\min,u}).$$

Here $C_u(\pi) = \sum_i c_{i,\pi(i)}^{(u)}$, $C_{\min,u} = C_u(\pi_{0,u})$, and $C_{\mathrm{rand},u}$ is the expected cost of a uniformly random assignment. The anchor $\pi_{0,u}$ minimizes this cost and is not ground truth. The dimensionless q scale interpolates between minimum and expected random cost. It is invariant to a common transformation $c'_{ij} = ac_{ij} + b$, $a > 0$, but does not equate different geometries. At each budget, a feasible assignment and a Lagrangian relaxation give $L_u(q) \le \Gamma_u(q) \le U_u(q)$, and seed averaging preserves this bracket. The primary value $q = .05$ was fixed before its outcomes were examined. The natural shifts use $q \in \{.01, .05, .10, .20, .40\}$. All 15 ImageNet-C corruptions use $q = .05$, and six focus corruptions use the full grid with disjoint and paired endpoints. Pair identities remain hidden

until the bounds are frozen. The nested profile-batch audit at $n \in \{1{,}024, 4{,}096, 8{,}192\}$ and the optimization checks appear in Sections E.3, E.4 and G.

Paired-calibrated recovery. The paired study evaluates severity-5 Gaussian noise and motion blur using identity, diagonal, and Bures corrections, three CLIP stability partitions, and one ViT-B/16 partition matched to the first CLIP split. Each of the eight encoder–corruption–partition cells uses a 4,096-pair fitting split, 29 disjoint calibration batches, and 20 disjoint evaluation batches of 256 examples. Every calibration and evaluation reference batch contains the clean counterparts of its targets. Calibration pairings set the threshold, while evaluation pairings remain hidden until the bounds are frozen. All methods share squared Mahalanobis costs after Bures, computed using the OAS covariance of the full reference fitting set (Chen et al., 2010; Givens & Shortt, 1984; Gelbrich, 1990; Bhatia et al., 2019). At nominal 90% coverage, the guarantee in Proposition 2 assumes that batch scores are exchangeable within a cell conditional on the fit and locked protocol. The study also evaluates direct split-conformal calibration of paired absolute errors and a separately recalibrated cost-column-permutation control (Shafer & Vovk, 2008; Johnstone & Cox, 2021). All calibration and evaluation batches pass the score-denominator check. Full protocol and verification details are in Section E.5.

Post-hoc paired audits. Three explicitly labeled audits reuse the fitted maps and 160 held-out batches. First, the primary feasible sets and thresholds are held fixed while common signed bounds are compared with separate methodwise-extrema bounds. A comparison is classified as *proven common-only* only when a negative signed upper bound certifies the ordering and feasible assignments establish that the separate ranges overlap. Second, a direct signed audit compares sum-of-SDs, paired-difference-SD, and centered-unscaled calibration at three confidence levels. The paired-difference-SD rule at $\alpha = .10$ was predeclared for that audit. These direct procedures use fitting-split pairs only after the correction maps are frozen. The geometric procedure does not use these pairs for fitting. Third, a sensitivity audit replaces the Bures cost generator with the frozen identity or diagonal correction and recalibrates each geometry at $\alpha = .10$. This audit was specified before its own outcomes were examined. Guarantees are separate across geometries, calibration families, confidence levels, and cells. Pooled repeated-batch counts are descriptive.

4 Results

Paired calibration is considered first because it supports recovery comparisons. The unpaired analysis then quantifies the remaining ambiguity.

4.1 Paired calibration resolves common-correspondence comparisons

Table 1 summarizes the paired recovery-comparison results.

The withheld correspondence lies in the primary Bures-generated set for 149/160 batches. The repeated-batch coverage and ordering counts in Table 1 are descriptive: Proposition 2 concerns one future exchangeable batch within a cell, not joint coverage over these 160 batches.

Table 1: **Recovery comparisons at 90% nominal coverage.** Joint coverage requires all three signed contrasts to cover within a batch. Counts pool repeated batches descriptively and are not independent trials. Direct procedures are post hoc. Slack is the mean signed upper bound minus the observed contrast in percentage points.

Procedure	Joint coverage (/160)	Covered contrasts (/480)	Issued orderings (/480)	Incorrect orderings	Mean slack (pp)
Bures-generated correspondence set	160	480	263	0	3.91
Paired-difference-SD direct	143	457	240	8	4.14
Sum-of-SDs direct	141	454	280	2	4.05
Centered, unscaled direct	143	457	280	2	4.03
Scrambled Bures costs	160	480	0	0	11.94

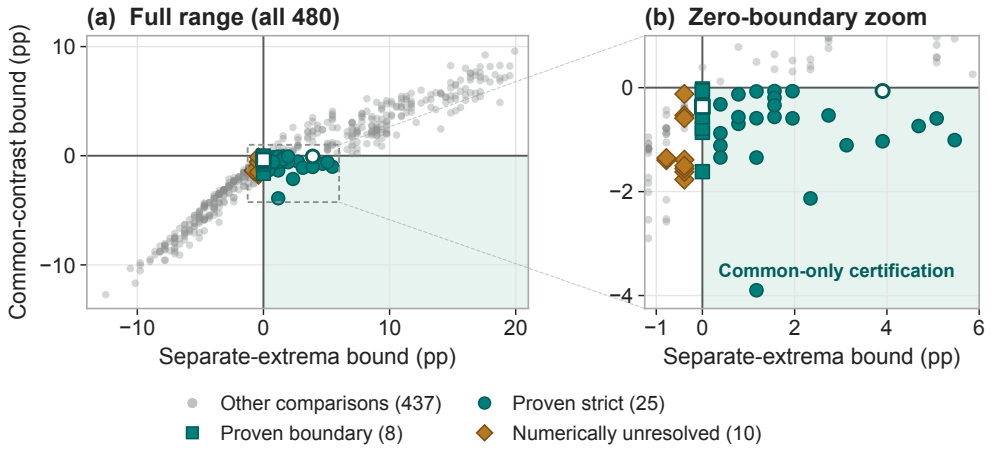


Figure 2: **Common correspondence certifies additional orderings.** Each point is one of 480 batch–contrast comparisons, pooled descriptively. Over the same correspondence set, x is a feasible lower bound on $\max E_a - \min E_c$, while y is an upper bound on $\max(E_a - E_c)$. Here x is not a lower bound on the realized contrast. Panel (b) enlarges the marked region in (a). The shaded region $x \geq 0,\ y < 0$ identifies orderings certified by the shared objective but not by separate ranges. Teal markers show 33 proven cases on 32 batches. Hollow markers flag two whose true pairing is infeasible. Amber diamonds mark ten numerically unresolved cases. All comparisons use Bures-generated costs. These are optimization bounds, not confidence intervals.

Why a common correspondence helps. The common-correspondence objective issues 263 strict orderings, compared with 220 from separate methodwise extrema (Table 2 and Figure 2). Ten cases in the raw 43-count difference remain numerically unresolved. Common signed upper bounds certify 33 additional orderings on 32 batches, with feasible witnesses establishing that separate error ranges cannot certify them: 25 are strict range crossings and eight are boundary overlaps. All 32 batches have positive anchored ambiguity, with mean bracket 16.54–16.99%, and 31 of the 33 certificates occur when the withheld assignment is feasible.

Thus a comparison can be identified while correspondence and individual errors remain unresolved. Two of the 33 proven common-only cases lie outside the sufficient feasibility event in Proposition 2. Because contrasts in a 256-example batch lie on a $1/256$ grid, the analysis reports the sharpened advantage in percentage points, $-100\lfloor 256\overline{H}_{a,c}\rfloor/256$. Its minimum and median are 0.391 and 3.516 points across all 263 certificates, and 0.391 and 0.781 points among the 33 common-only cases. Full margin-threshold counts are reported in Section E.5.

Calibration and geometry controls. Permuting the Bures cost columns and recalibrating preserves observed joint coverage, but makes every absolute bound 100% and issues no ordering (Table 1). Identity- and diagonal-generated costs also retain observed joint coverage with no wrong-sign

Table 2: **The common matching resolves additional comparisons.** Each contrast is evaluated on 160 batches. Thirty-three cases are proven common-only on 32 distinct batches: a negative signed upper bound certifies the ordering, while feasible witnesses show that separate error ranges overlap. Ten further cases are numerically unresolved.

Contrast	Common signed	Separate extrema	Proven common-only	True pairing feasible among common-only cases
Diagonal − identity	100	91	7	6
Bures − identity	104	94	9	8
Bures − diagonal	59	35	17	17
Total (480 comparisons)	**263**	**220**	**33**	**31**

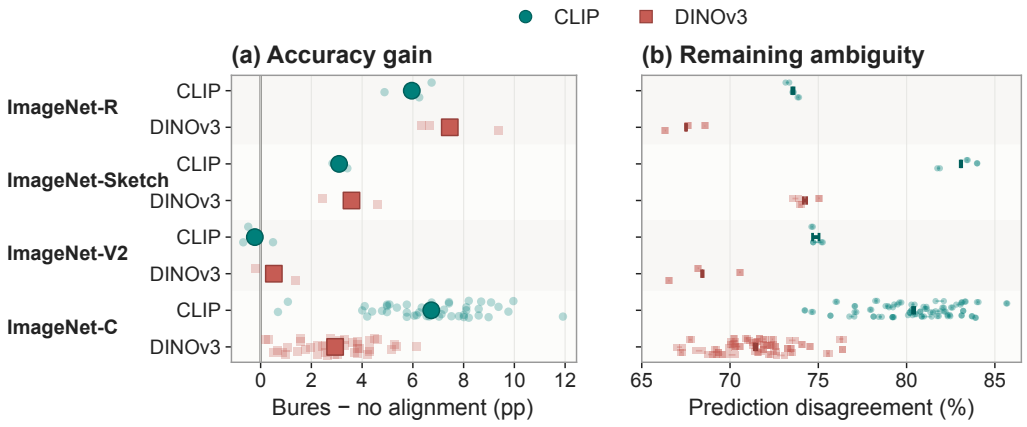


Figure 3: **Accuracy gains coexist with correspondence ambiguity.** Unpaired batches use $q = 0.05$ and $n = 1{,}024$. (a) Bures accuracy gains over no alignment. (b) Prediction-disagreement bounds relative to the minimum-cost matching. Small markers represent split-level results for natural shifts and corruption-by-split results for ImageNet-C. Large markers in (a) show means. Capped segments in (b) show $[\text{mean } L, \text{mean } U]$, not confidence intervals or classification-error estimates.

decisions. Each issues 201 orderings and has roughly twice the primary geometry's mean absolute slack. The fixed comparison tests sensitivity to the cost generator without selecting a geometry from the evaluation data (Table 22).

Across the three post-hoc direct procedures, joint coverage is 141–143/160 and 2–8 issued orderings have the wrong sign, compared with 160/160 and zero for the primary Bures-generated set (Table 1). The direct procedures use true fitting pairs after map fitting, whereas the geometric method does not. Absolute-error tightness is condition-dependent: direct calibration is tighter for CLIP Gaussian noise, while the Bures-generated set is tighter for motion blur and ViT-B/16. Full results are in Tables 19 and 21.

4.2 Unpaired ambiguity remains high at larger profile batch sizes

Under the common normalized cost scale, Table 15 shows substantial ambiguity for natural shifts and focus corruptions. At primary $q = .05$, the three natural-shift means and the all-15 ImageNet-C macro lie between 70.49% and 78.73% disagreement. On the six-corruption paired-endpoint control, the mean bracket contracts to $[28.02, 28.15]\%$. Holding the cost definition fixed, this contraction shows that the profile responds strongly to endpoint structure. Full budget grids and all-15 encoder brackets appear in Section E.3.

Increasing the profile batch size from 1,024 to 8,192 raises mean same-class candidate availability from 63.16% to 99.94%. Nevertheless, all 96 feasible lower bounds at the largest size remain above 60.99%. The direction of change depends on the encoder, so this fixed-panel analysis supports no single monotone scaling trend. Complete cell counts, brackets, and post-freeze class summaries are in Tables 17 and 18.

Endpoint structure changes the profile under the same cost definition. At $q = .05$, all 36 paired-versus-disjoint corruption–encoder–seed brackets are separated, with a mean paired-minus-disjoint contrast of $[-47.34, -47.07]$ points. All withheld clean–corrupt pairings are feasible. Their excess costs average 0.67% and never exceed 4.71% of the random-minus-minimum gap. Thus the budget contains the withheld pairings without exposing them to the optimizer. Because the paired and disjoint arms use different endpoint sets, this comparison tests informativeness of endpoint structure rather than a causal effect of pairing (Section E.3).

4.3 USEFUL ALIGNMENT CAN COEXIST WITH HIGH AMBIGUITY

High ambiguity under the declared correspondence model does not preclude accuracy gains. At $q = .05$, shift-level mean ambiguity is roughly 70–80%, yet Bures improves macro accuracy by 3.40 points on the natural shifts and 4.83 points on ImageNet-C. Across the 18 shift and corruption clusters, the Spearman association between ambiguity and Bures gain is $\rho = .278$ (within-family permutation $p = .365$). The estimate is inconclusive, so the profile is treated as an ambiguity diagnostic rather than a calibrated predictor of alignment benefit (Figure 3 and Sections E.3 and E.4).

Optimization uncertainty is small for the original ambiguity estimand. Across the 720 common-cost certificates at 1,024 rows, $U - L$ averages 0.103 points and never exceeds 0.952. No numerical width among the 1,170 ambiguity-profile certificates exceeds 0.993 points. Exact-check and reconstruction diagnostics are reported in Table 14 and Section G.

5 RELATED WORK

Domain adaptation and identifiability. Unsupervised domain adaptation requires assumptions linking source and target domains: classical lower bounds construct observationally indistinguishable tasks, and invariant representations can preserve domain invariance while increasing target error (Ben-David et al., 2010; Zhao et al., 2019). Exact representation matching can also be incompatible with low target error when label masses differ (Wu et al., 2019). At the level of transport maps, distribution-matched translators are generally non-unique under source automorphisms (Galanti et al., 2018; Moriakov et al., 2020). Source symmetries therefore induce spurious maps with different downstream behavior (Gulrajani & Hashimoto, 2022). Multiple corresponding distributions, or a single paired anchor combined with structural restrictions, can restore identification (Shrestha & Fu, 2024; Shrestha et al., 2026). The analysis quantifies finite-batch prediction ambiguity among cost-constrained one-to-one assignments of observed endpoints. It does not claim population identification.

Feature alignment. Coordinate-statistic matching and CORAL are established feature-alignment methods (Nado et al., 2020; Schneider et al., 2020; Sun et al., 2016). Gaussian/Bures geometry supplies second-moment transport (Givens & Shortt, 1984; Gelbrich, 1990; Bhatia et al., 2019) and has also been used in domain adaptation (Yair et al., 2019; Luo & Ren, 2021). Parameter-updating test-time adaptation and dedicated modules also align representations (Adachi et al., 2023; Jung et al., 2023; Mirza et al., 2023; Thopalli et al., 2023). Post-hoc alignment can fail under realistic shifts (Burns & Steinhardt, 2021), while matched marginals need not identify an examplewise correction (Moriakov et al., 2020). Here the encoder and readout remain fixed, separating marginal, sample, and clean-prediction recovery. The certificates apply only to the declared correspondence model and neither bound target-label accuracy nor evaluate parameter-updating test-time adaptation.

Robust prediction under map uncertainty. Worst-case Over Maps formalizes population-level uncertainty over spurious maps and downstream predictions (Gulrajani & Hashimoto, 2022); the two-world randomized zero–one recovery value is a finite specialization. Related work also studies posterior-mode domain adaptation (Dong et al., 2026), transport entanglement (Koç et al., 2025), and classification under output-distribution constraints (Xian et al., 2023; Gaucher et al., 2023). Unlike Worst-case Over Maps, this work brackets the finite worst case over one-to-one assignments of indexed batch atoms, which need not define a pointwise feature map. With disjoint paired batches, exchangeable-rank calibration converts the same feasible set into simultaneous method-specific and signed recovery bounds for one future batch (Shafer & Vovk, 2008; Johnstone & Cox, 2021).

6 CONCLUSION

When clean–shifted correspondences are unknown, the framework separates ambiguity, method comparison, and accuracy. Unpaired ambiguity coexists with Bures accuracy gains. In the paired study, all 263 issued orderings agree with the hidden pairs. Common signed bounds certify 33 orderings on 32 batches that separate ranges cannot, despite uncertain matching and absolute errors. Reassigned-cost controls retain coverage but eliminate every ordering, locating informativeness in cost–endpoint association. Shared uncertainty can thus certify a method ordering without identifying correspondence or absolute errors.

AI USE STATEMENT

Generative AI tools were used to aid or polish writing, support retrieval and discovery, including finding related work, and assist research ideation or execution. The authors independently verified the manuscript and reported results and take responsibility for the work.

REPRODUCIBILITY STATEMENT

The complete assumptions and proofs are in Section A. The models, datasets, preprocessing, splits, methods, calibration budgets, metrics, and inference procedures are specified in Section B. Complete result panels and additional experiments are reported in Sections C to E, including the complete common-cost, endpoint-structure, and realized-profile follow-up in Section E.3, and the nested profile-batch audit in Section E.4. Paired method-specific recovery-bound validation and the common-correspondence, direct-calibration, and cost-generator audits are documented in Section E.5. The anonymous repository contains the resolved configurations, implementation, tests, and verification code. Source and execution evidence are kept as separate, checksum-verified bundles so that a source-only archive cannot be mistaken for a results release. Third-party data and checkpoints are not redistributed. The artifact documents the exact variants, official sources, locked model revisions, and acquisition and validation procedures.

REFERENCES

Kazuki Adachi, Shin'ya Yamaguchi, and Atsutoshi Kumagai. Covariance-aware feature alignment with pre-computed source statistics for test-time adaptation to multiple image corruptions. In *IEEE International Conference on Image Processing*, pp. 800–804, 2023. doi: 10.1109/ICIP49359.2023.10222901. URL https://arxiv.org/abs/2204.13263.

Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pp. 129–136, 2010.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019. doi: 10.1016/j.exmath.2018.01.002.

Garrett Birkhoff. Tres observaciones sobre el álgebra lineal. *Universidad Nacional de Tucumán. Revista A*, 5:147–151, 1946.

Benjamin M. Bolstad, Rafael A. Irizarry, Magnus Åstrand, and Terence P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003. doi: 10.1093/bioinformatics/19.2.185.

Collin Burns and Jacob Steinhardt. Limitations of post-hoc feature alignment for robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2525–2533, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Burns_Limitations_of_Post-Hoc_Feature_Alignment_for_Robustness_CVPR_2021_paper.html.

Yilun Chen, Ami Wiesel, Yonina C. Eldar, and Alfred O. Hero, III. Shrinkage algorithms for MMSE covariance estimation. *IEEE Transactions on Signal Processing*, 58(10):5016–5029, 2010. doi: 10.1109/TSP.2010.2053029.

Anthony C. Davison and David V. Hinkley. *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge, 1997. doi: 10.1017/CBO9780511802843.

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

Zhiyi Dong, Zixuan Liu, and Yongyi Mao. On the hardness of unsupervised domain adaptation: Optimal learners and information-theoretic perspective. In *Proceedings of the 4th Conference on*

Lifelong Learning Agents, volume 330 of *Proceedings of Machine Learning Research*, pp. 89–111. PMLR, 2026. URL https://proceedings.mlr.press/v330/dong26a.html.

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

Aryeh Dvoretzky, Abraham Wald, and Jacob Wolfowitz. Elimination of randomization in certain statistical decision procedures and zero-sum two-person games. *The Annals of Mathematical Statistics*, 22(1):1–21, 1951. doi: 10.1214/aoms/1177729689.

Morris L. Eaton and Edward I. George. Charles Stein and invariance: Beginning with the Hunt–Stein theorem. *The Annals of Statistics*, 49(4):1815–1822, 2021. doi: 10.1214/21-AOS2075. URL https://doi.org/10.1214/21-AOS2075.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. doi: 10.1007/BF02288367.

Tomer Galanti, Lior Wolf, and Sagie Benaim. The role of minimal complexity functions in unsupervised learning of semantic mappings. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1709.00074.

Solenne Gaucher, Nicolas Schreuder, and Evgenii Chzhen. Fair learning with Wasserstein barycenters for non-decomposable performance measures. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 2436–2459. PMLR, 2023. URL https://proceedings.mlr.press/v206/gaucher23a.html.

Matthias Gelbrich. On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990. doi: 10.1002/mana.19901470121.

Clark R. Givens and Rae Michael Shortt. A class of Wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231–240, 1984. doi: 10.1307/mmj/1029003026.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://www.jmlr.org/papers/v13/gretton12a.html.

Ishaan Gulrajani and Tatsunori Hashimoto. Identifiability conditions for domain adaptation. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 7982–7997, 2022. URL https://proceedings.mlr.press/v162/gulrajani22a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Alfred Haar. Der massbegriff in der theorie der kontinuierlichen gruppen. *Annals of Mathematics*, 34 (1):147–169, 1933. doi: 10.2307/1968346.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020. doi: 10.1038/s41586-020-2649-2.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Hendrycks_The_Many_Faces_of_Robustness_A_Critical_Analysis_of_Out-of-Distribution_ICCV_2021_paper.html.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Qi Huangfu and J. A. Julian Hall. Parallelizing the dual revised simplex method. *Mathematical Programming Computation*, 10(1):119–142, 2018. doi: 10.1007/s12532-017-0130-5.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975. doi: 10.1016/0047-259X(75)90042-1.

Chancellor Johnstone and Bruce Cox. Conformal uncertainty sets for robust optimization. In *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 152 of *Proceedings of Machine Learning Research*, pp. 72–90. PMLR, 2021. URL https://proceedings.mlr.press/v152/johnstone21a.html.

Sanghun Jung, Jungsoo Lee, Nanhee Kim, Amirreza Shaban, Byron Boots, and Jaegul Choo. CAFA: Class-aware feature alignment for test-time adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19060–19071, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Jung_CAFA_Class-Aware_Feature_Alignment_for_Test-Time_Adaptation_ICCV_2023_paper.html.

Olav Kallenberg. *Foundations of Modern Probability*, volume 99 of *Probability Theory and Stochastic Modelling*. Springer, Cham, 3 edition, 2021. doi: 10.1007/978-3-030-61871-1.

Okan Koç, Alexander Soen, Chao-Kai Chiang, and Masashi Sugiyama. Domain adaptation and entanglement: an optimal transport perspective. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3034–3042. PMLR, 2025. URL https://proceedings.mlr.press/v258/koc25a.html.

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

Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955. doi: 10.1002/nav.3800020109.

Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer, New York, 1986. doi: 10.1007/978-1-4612-4946-7.

Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, New York, 3 edition, 2005. doi: 10.1007/0-387-27605-X.

You-Wei Luo and Chuan-Xian Ren. Conditional Bures metric for domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13989–13998, 2021. doi: 10.1109/CVPR46437.2021.01377.

Leon Mirsky. A trace inequality of John von Neumann. *Monatshefte für Mathematik*, 79(4):303–306, 1975. doi: 10.1007/BF01647331.

Muhammad Jehanzeb Mirza, Pol Jané Soneira, Wei Lin, Mateusz Kozinski, Horst Possegger, and Horst Bischof. ActMAD: Activation matching to align distributions for test-time-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24152–24161, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Mirza_ActMAD_Activation_Matching_To_Align_Distributions_for_Test-Time-Training_CVPR_2023_paper.html.

Nikita Moriakov, Jonas Adler, and Jonas Teuwen. Kernel of CycleGAN as a principal homogeneous space. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1eWOJHKvB.

Zachary Nado, Shreyas Padhy, D. Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL https://www.jmlr.org/papers/v12/pedregosa11a.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, pp. 1177–1184, 2007. URL https://proceedings.neurips.cc/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400, 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Siegfried M. Rump. Verification methods: Rigorous results using floating-point arithmetic. *Acta Numerica*, 19:287–449, 2010. doi: 10.1017/S096249291000005X.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11539–11551, 2020.

Peter H. Schönemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10, 1966. doi: 10.1007/BF02289451.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008. URL https://jmlr.org/papers/v9/shafer08a.html.

William Fleetwood Sheppard. On the application of the theory of error to cases of normal distribution and normal correlation. *Philosophical Transactions of the Royal Society of London. Series A*, 192: 101–167, 1899. doi: 10.1098/rsta.1899.0003. URL https://doi.org/10.1098/rsta.1899.0003.

Sagar Shrestha and Xiao Fu. Towards identifiable unsupervised domain translation: A diversified distribution matching approach. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8bb5a934785817f752e7f9322f9b4d54-Abstract-Conference.html.

Sagar Shrestha, Subash Timilsina, Hoang-Son Nguyen, and Xiao Fu. Domain transfer becomes identifiable via a single alignment. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://arxiv.org/abs/2605.17918.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231, 1959.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? Data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022.

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, pp. 2058–2065, 2016. doi: 10.1609/aaai.v30i1.10306. URL https://ojs.aaai.org/index.php/AAAI/article/view/10306.

Kowshik Thopalli, Pavan Turaga, and Jayaraman J. Thiagarajan. Domain alignment meets fully test-time adaptation. In *Proceedings of the 14th Asian Conference on Machine Learning*, volume 189 of *Proceedings of Machine Learning Research*, pp. 1006–1021. PMLR, 2023. URL https://proceedings.mlr.press/v189/thopalli23a.html.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, 2009. doi: 10.1007/978-3-540-71050-9. URL https://doi.org/10.1007/978-3-540-71050-9.

John von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928. doi: 10.1007/BF01448847. URL https://doi.org/10.1007/BF01448847.

Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/3eefceb8087e964f89c2d59e8a249915-Abstract.html.

Yifan Wu, Ezra Winston, Divyansh Kaushik, and Zachary Lipton. Domain adaptation with asymmetrically-relaxed distribution alignment. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6872–6881. PMLR, 2019. URL https://proceedings.mlr.press/v97/wu19f.html.

Ruicheng Xian, Lang Yin, and Han Zhao. Fair and optimal classification via post-processing. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 37977–38012. PMLR, 2023. URL https://proceedings.mlr.press/v202/xian23b.html.

Or Yair, Mirela Ben-Chen, and Ronen Talmon. Parallel transport on the cone manifold of SPD matrices for domain adaptation. *IEEE Transactions on Signal Processing*, 67(7):1797–1811, 2019. doi: 10.1109/TSP.2019.2894801.

Han Zhao, Rémi Tachet des Combes, Kun Zhang, and Geoffrey J. Gordon. On learning invariant representations for domain adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 7523–7532, 2019.

APPENDIX

A Proofs and precise scope of correspondence ambiguity and recovery **16**
A.1 Why the map supremum is not an information ceiling . . . 16
A.2 Diagonal alignment and orthogonal ambiguity . . . 16
A.3 A sharp two-world ceiling . . . 17
A.4 Finite-instance ambiguity profile and certificate . . . 18
A.5 Paired method-specific recovery certificate . . . 19
A.6 Approximately distinguishable observation laws . . . 22
A.7 Posterior-mode baseline and exact-alignment capacity . . . 22
A.8 Finite exact-quota control . . . 23
A.9 Finite-world constrained minimax dual . . . 24
A.10 Exact compact-group value . . . 24
A.11 Gaussian binary closed form . . . 25
A.12 Readout-weighted value of paired correspondence . . . 26
B Complete experimental protocol **28**
B.1 Models, data, and frozen classifiers . . . 28
B.2 Alignment methods and information available . . . 28
B.3 Suites, metrics, and inference . . . 29
C Complete real-data results **30**
C.1 Primary marginal hierarchy . . . 30
C.2 Direct numerical check of coordinate invariance . . . 30
C.3 Severity and calibration size . . . 30
C.4 Pair assignments and estimator choice . . . 31
C.5 Strict predict-before-update streaming . . . 32
D Controlled identifiability experiments **33**
E Additional ambiguity and recovery results **34**
E.1 Held-out orientation selector . . . 34
E.2 Optimization-certified natural-shift profiles . . . 34
E.3 Common-cost and endpoint-structure follow-up . . . 35
E.4 Nested profile-batch and class-structure audit . . . 37
E.5 Paired method-specific recovery-bound validation . . . 39
F Practical guidance from the evaluated setting **45**
G Integrity, amendments, and reproducibility **46**
H Scope of the empirical conclusions **47**

A PROOFS AND PRECISE SCOPE OF CORRESPONDENCE AMBIGUITY AND RECOVERY

Throughout, the task label set $\mathcal{Y}$ is finite, the fixed classifier and desired-label functions are measurable, and all maps called bijections are bimeasurable. A *marginal-only procedure* may use the reference and target laws P_r, P_t, the fixed classifier h, a declared ambiguity model, and fresh randomization, but not the realized latent correspondence. Consequently, its distribution over corrections is identical in all worlds with the same observable (P_r, P_t, h). All algorithmic randomness below is independent of the evaluation sample and the hidden world. For compactness, $g = h$ denotes this fixed downstream classifier, while τ is reserved for a possibly different external task.

Several ingredients below are classical: the copula consequence of Sklar's theorem for increasing coordinate transformations (Sklar, 1959), Gaussian/Bures covariance transport (Givens & Shortt, 1984; Gelbrich, 1990; Bhatia et al., 2019), best low-rank approximation (Eckart & Young, 1936), Gaussian sign disagreement (Sheppard, 1899), and the SVD solution of orthogonal Procrustes (Schönemann, 1966). Prior work also connects admissible-map ambiguity to downstream disagreement and worst-case prediction error (Gulrajani & Hashimoto, 2022). Related decision baselines include posterior-mode domain adaptation (Dong et al., 2026), transport entanglement (Koç et al., 2025), and classification under output-distribution constraints (Xian et al., 2023; Gaucher et al., 2023). This appendix develops the paper's finite, cost-constrained correspondence model. The model supports an unpaired ambiguity audit and, with paired calibration, method-specific recovery assessment. A further result shows how exact marginal transport imposes an output-label-mass constraint in correspondence worlds. The two-world and compact-group results specialize the Hunt–Stein robust-decision and Haar-symmetrization perspective (Eaton & George, 2021).

A.1 WHY THE MAP SUPREMUM IS NOT AN INFORMATION CEILING

Let $P_r = P_t = P$ and write $\pi_y = P\{h(Z) = y\}$. Under the identity coupling $Z_t = Z_r$, the feasible map $T = I$ has agreement one. Now consider a coupling outside the deterministic bijective-drift model, in which Z_t and Z_r are independent. Any (possibly randomized) T whose output marginal is P obeys $h(T(Z_t)) \sim (\pi_y)_y$ and remains independent of $h(Z_r)$. Therefore

$$\Pr\{h(T(Z_t)) = h(Z_r)\} = \sum_{y \in \mathcal{Y}} \pi_y^2.$$

The reference and target laws and fixed downstream classifier are identical in these two examples, while the oracle coupling-dependent supremum differs. Moreover, whenever $Z_t = F(Z_r)$ for an invertible F, the true inverse F^{-1} is feasible and attains one even if it is unidentifiable from the two marginals. Thus a push-forward constraint restricts the chosen map but does not restrict the information used to choose it. A marginal-information ceiling must instead evaluate one decision rule across observationally equivalent worlds.

A.2 DIAGONAL ALIGNMENT AND ORTHOGONAL AMBIGUITY

Standardized-vector invariance. Let

$$T_{\mathrm{diag}}(z) = \mu_r + D_r D_t^{-1}(z - \mu_t),$$

where D_r, D_t are diagonal with strictly positive entries. Direct substitution gives, pointwise,

$$D_r^{-1}\{T_{\mathrm{diag}}(z) - \mu_r\} = D_t^{-1}(z - \mu_t).$$

If the supplied parameters are population moments, taking expectations gives the reference mean and taking coordinatewise second moments gives the reference coordinate variances. If they are empirical moments, this exact moment matching holds on the same fitted sample. Because the two standardized random vectors are equal almost surely, every measurable functional of that vector is preserved. In particular, this includes all finite standardized mixed moments, correlations, coordinatewise ranks, and, for continuous coordinate marginals, the copula. No Gaussian assumption is used. When moments are estimated, the identity holds for the moments supplied to the map. It is not a statement that those estimates equal their population values.

All covariance-matching linear maps. Suppose $\Sigma_r, \Sigma_t \succ 0$. If $A\Sigma_t A^\top = \Sigma_r$, then

$$R = \Sigma_r^{-1/2} A \Sigma_t^{1/2} \quad \text{satisfies} \quad RR^\top = I,$$

so $R \in O(d)$ and

$$A = \Sigma_r^{1/2} R \Sigma_t^{-1/2}.$$

The converse follows by substitution. If A is additionally symmetric positive definite, set $B = \Sigma_t^{1/2} A \Sigma_t^{1/2}$. Then $B \succ 0$ and

$$B^2 = \Sigma_t^{1/2} \Sigma_r \Sigma_t^{1/2}.$$

The positive-definite square root is unique, which yields

$$A = \Sigma_t^{-1/2} (\Sigma_t^{1/2} \Sigma_r \Sigma_t^{1/2})^{1/2} \Sigma_t^{-1/2}.$$

This proves both the unidentified orthogonal factor and uniqueness of the Gaussian Bures representative. The orthogonal factor is an ambiguity of second moments. It is an ambiguity of the complete reference and target laws only when the full law has the relevant symmetry, as for a whitened standard Gaussian law, or when the procedure is explicitly limited to second moments.

Exact inversion of a linear drift. Suppose centered target and reference variables obey $Y = MX$ with $M \in GL(d)$ and $\mathrm{Cov}(X) = \Sigma_r \succ 0$. Exact recovery of every centered feature by Bures alignment is equivalent to

$$A_{\mathrm{BW}} M = I,$$
$$A_{\mathrm{BW}} = M^{-1}.$$

Because A_{BW} is symmetric positive definite, this can hold only if M^{-1}, equivalently M, is symmetric positive definite. Here $\Sigma_r \succ 0$ implies that the support is not contained in a proper linear subspace, which justifies the matrix identity from almost-sure recovery. Conversely, if M is symmetric positive definite, then M^{-1} is a symmetric-positive-definite solution of $A(M\Sigma_r M^\top)A^\top = \Sigma_r$. Uniqueness of the symmetric-positive-definite congruence solution implies $A_{\mathrm{BW}} = M^{-1}$, proving the exact-inversion result.

A.3 A SHARP TWO-WORLD CEILING

Theorem 1 (Sharp two-world ceiling)**.** *Let F_0, F_1 be bimeasurable bijections satisfying $(F_0)_\# P_r = (F_1)_\# P_r = P_t$. For $X \sim P_t$, set $\ell_i(X) = \tau(F_i^{-1}X)$, $\delta = \Pr\{\ell_0(X) \neq \ell_1(X)\}$, and $A_i = \Pr\{\widehat{Y}(X) = \ell_i(X)\}$ for any world-independent randomized prediction rule. Then*

$$\min(A_0, A_1) \leq 1 - \frac{\delta}{2}. \tag{12}$$

If $\tau = g$ and both inverses are admissible corrections, choosing either inverse uniformly at random attains equality.

Proof. Pointwise,

$$\mathbf{1}\{\widehat{Y} = \ell_0\} + \mathbf{1}\{\widehat{Y} = \ell_1\} \leq 2 - \mathbf{1}\{\ell_0 \neq \ell_1\}.$$

Taking expectations yields

$$A_0 + A_1 \leq 2 - \delta,$$
$$\min(A_0, A_1) \leq \frac{A_0 + A_1}{2},$$
$$\frac{A_0 + A_1}{2} \leq 1 - \frac{\delta}{2}.$$

The last two lines give (12). This upper bound does not require $\widehat{Y}$ to arise from a transport.

If $\tau = g$ and both F_0^{-1}, F_1^{-1} are admissible corrections, choose each inverse with probability $1/2$. Its two world-specific agreements are

$$A_0 = \frac{1}{2}\{1 + (1 - \delta)\} = 1 - \frac{\delta}{2},$$
$$A_1 = \frac{1}{2}\{(1 - \delta) + 1\} = 1 - \frac{\delta}{2}.$$

Thus both world-specific expected agreements equal $1-\delta/2$, proving that (12) is the exact randomized minimax value under these assumptions. For an imperfect fixed classifier evaluated against an external task $\tau \neq g$, or for a correction class excluding an inverse, the upper bound remains valid but the attainability argument does not. □

A.4 Finite-instance ambiguity profile and optimization certificate

Worked three-row example. Let the anchor be the identity, let the three reference predictions be (A, A, B), and take $c = \left(\begin{smallmatrix} 0 & 3 & 3 \\ 3 & 0 & 1 \\ 3 & 1 & 0 \end{smallmatrix}\right)$ with budget $B = 2$. Swapping rows 2 and 3 costs 2 and changes both implied predictions, so it certifies $L = 2/3$. Every other reward-improving permutation exceeds the budget, and the Lagrangian bound at $\lambda = 1/3$ gives $U = 2/3$. Hence the declared instance has ambiguity exactly $2/3$, and the corresponding two-world agreement upper bound is $1 - (2/3)/2 = 2/3$. If the two indexed inverse corrections are admissible, uniform randomization attains this value. The real-data audit repeats these steps at larger scale, with a small nonzero numerical gap.

Anchor-sensitivity bounds. On a fixed feasible set $\mathcal{P}_B$, prediction Hamming distance is a pseudometric. For anchors $a, b \in \mathcal{P}_B$, write $\Gamma_a(B) = \max_{\pi \in \mathcal{P}_B} d_h(a, \pi)$ and $\Delta_h(B) = \max_{\pi,\sigma \in \mathcal{P}_B} d_h(\pi, \sigma)$. Hence $\Gamma_a(B) \leq \Delta_h(B)$, while for every $\pi, \sigma \in \mathcal{P}_B$,

$$\begin{aligned} d_h(\pi, \sigma) &\leq d_h(\pi, a) + d_h(a, \sigma) \\ &\leq 2\Gamma_a(B). \end{aligned}$$

Therefore

$$\begin{aligned} \Gamma_a(B) &\leq \Delta_h(B), \\ \Delta_h(B) &\leq 2\Gamma_a(B). \end{aligned}$$

Also,

$$\begin{aligned} d_h(a, \pi) &\leq d_h(a, b) + d_h(b, \pi), \\ \Gamma_a(B) &\leq d_h(a, b) + \Gamma_b(B), \\ |\Gamma_a(B) - \Gamma_b(B)| &\leq d_h(a, b), \end{aligned}$$

where the last line follows by exchanging a and b. These comparisons require the same feasible set and make no claim that its cost model contains the true correspondence.

Proof of Proposition 1. An explicit form of the relaxation is obtained by setting

$$\begin{aligned} d_{ij} &= \mathbf{1}\{h(r_{\pi_0(i)}) \neq h(r_j)\}, \\ a_{ij} &= d_{ij}/n, \end{aligned}$$

and letting $\mathfrak{B}_n$ be the Birkhoff polytope of nonnegative matrices whose row and column sums are one. The relaxed profile is the linear program

$$\Gamma^{\mathrm{LP}}_{h,\pi_0}(B) = \max_{\substack{x \in \mathfrak{B}_n \\ \sum_{ij} c_{ij} x_{ij} \leq B}} \sum_{ij} a_{ij} x_{ij}. \tag{13}$$

Because $B \geq C(\pi_0)$, its feasible set is nonempty, and compactness gives attainment. Every feasible permutation matrix is feasible for (13), proving $\Gamma_{h,\pi_0}(B) \leq \Gamma^{\mathrm{LP}}_{h,\pi_0}(B)$.

The linear-programming dual of (13) is

$$\min_{\substack{\lambda \geq 0,\ \alpha, \beta \in \mathbb{R}^n \\ \alpha_i + \beta_j + \lambda c_{ij} \geq a_{ij}}} \left\{ \lambda B + \sum_i \alpha_i + \sum_j \beta_j \right\}. \tag{14}$$

The row and column potentials are unrestricted because their primal constraints are equalities. Finite-dimensional LP strong duality applies because the primal is feasible and bounded. For fixed λ,

standard LP duality and Birkhoff–von Neumann integrality (Birkhoff, 1946) imply

$$\min_{\substack{\alpha,\beta\in\mathbb{R}^n \\ \alpha_i+\beta_j\geq a_{ij}-\lambda c_{ij}}} \left(\sum_i \alpha_i + \sum_j \beta_j\right) = \max_{x\in\mathfrak{B}_n} \sum_{ij}(a_{ij}-\lambda c_{ij})x_{ij}$$
$$= \max_{\pi\in S_n}\{D_{h,\pi_0}(\pi) - \lambda C(\pi)\}.$$

Minimizing over $\lambda \geq 0$ proves the equality in (4). Alternatively, for every feasible π and $\lambda \geq 0$, weak duality follows directly from

$$D_{h,\pi_0}(\pi) \leq D_{h,\pi_0}(\pi) + \lambda\{B - C(\pi)\}$$
$$\leq U_\lambda(B).$$

Thus any feasible permutation supplies a lower bound and any correctly solved Lagrangian assignment subproblem supplies an upper bound. Each fixed-λ subproblem is a standard linear assignment (Kuhn, 1955). The extra bound by one follows because disagreement is at most one. □

Two-world consequence. Because S_n is finite, let $\pi^\star$ attain $\Gamma_{h,\pi_0}(B)$. Apply Theorem 1 to the indexed worlds π_0 and $\pi^\star$. Their desired predictions disagree on exactly $\Gamma_{h,\pi_0}(B)$ of the atoms, so every world-independent randomized batch rule has worst-world agreement at most $R_{h,\pi_0}(B) = 1 - \Gamma_{h,\pi_0}(B)/2$. Its worst agreement over the full feasible world class cannot exceed its worst agreement over this pair. If both indexed inverses for a selected feasible $\widehat{\pi}$ are admissible, uniform randomization between them attains $1 - D_{h,\pi_0}(\widehat{\pi})/2$ in each selected world, establishing the corresponding pairwise equality. Finally, applying the decreasing affine map $d \mapsto 1 - d/2$ to $L \leq \Gamma_{h,\pi_0}(B) \leq U$ gives

$$\begin{aligned} 1 - U/2 &\leq R_{h,\pi_0}(B), \\ R_{h,\pi_0}(B) &\leq 1 - L/2, \\ R_{h,\pi_0}(B) &= 1 - \Gamma_{h,\pi_0}(B)/2. \end{aligned} \tag{15}$$

The resulting interval has width at most $(U - L)/2$.

Three quantities must be distinguished. A serialized feasible assignment proves a lower bound on the integer profile. The optimized Lagrangian value equals the LP relaxation and may exceed the integer optimum. The third quantity, $R_{h,\pi_0}(B)$, is the strongest *anchored pairwise* ceiling, not the full-world minimax value. Implementations using floating-point assignment solvers provide numerical computational bounds under their recorded solver model. A formally verified real-arithmetic upper bound would additionally require, for example, outward-rounded or exact verification of all dual inequalities in (14) (Rump, 2010).

A.5 Paired conformal calibration of method-specific recovery

The following result applies the standard split-conformal exchangeable-rank argument to the paper's correspondence score (Shafer & Vovk, 2008; Johnstone & Cox, 2021).

Proposition 3 (Paired split-conformal recovery certificate)**.** *Fit the corrections and cost geometry on data disjoint from m calibration batches and one future evaluation batch. Suppose the $m+1$ true-pair scores $S_b(\pi_b^\star)$ have positive normalizing denominators and are exchangeable conditional on that fit and on the locked protocol. For $\alpha \in (0,1)$, let $k = \lceil (m+1)(1-\alpha) \rceil$ and let $\widehat{q}$ be the kth calibration score when $k \leq m$, and set $\widehat{q} = +\infty$ otherwise. Then, for every finite predeclared method family $\mathcal{A}$,*

$$\mathbb{P}\big\{\ \forall a \in \mathcal{A}: \quad E_{a,m+1}(\pi^\star_{m+1}) \leq G_{a,m+1}(\widehat{q})\ \big\} \geq 1 - \alpha. \tag{16}$$

If a serialized feasible assignment and a valid Lagrangian relaxation give $L_{a,b} \leq G_{a,b}(\widehat{q}) \leq U_{a,b}$, then $U_{a,m+1}$ is a simultaneous upper certificate in (16), while $U_{a,b} - L_{a,b}$ records its optimization width.

Proof of Proposition 3. The result applies separately within each predeclared corruption. For a batch of n paired clean and corrupt endpoints, let

$$C_b(\pi) = \sum_i c_{b,i,\pi(i)},$$
$$C_b^{\min} = \min_{\pi\in S_n} C_b(\pi),$$
$$\overline{C}_b = \mathbb{E}_{\pi\sim\mathrm{Unif}(S_n)} C_b(\pi).$$

When $\overline{C}_b > C_b^{\min}$, the locked score used by the prospective protocol is

$$S_b(\pi) = \frac{C_b(\pi) - C_b^{\min}}{\overline{C}_b - C_b^{\min}}, \qquad \mathcal{P}_b(q) = \{\pi \in S_n : S_b(\pi) \le q\}. \tag{17}$$

The costs, normalization, and hence $\mathcal{P}_b(q)$ do not depend on the method a. The positive-denominator check is part of the numerical protocol. A batch failing it receives no certificate rather than a silently modified score. For each method, $G_{a,b}(q) = \max_{\pi \in \mathcal{P}_b(q)} E_{a,b}(\pi)$ is therefore a supremum over the same feasible correspondence set. In particular,

$$S_b(\pi_b^\star) \le q \quad \Longrightarrow \quad E_{a,b}(\pi_b^\star) \le G_{a,b}(q) \quad \text{for every } a \in \mathcal{A}. \tag{18}$$

Accordingly, inclusion of the true pairing in the common cost set ensures that each method's realized clean-prediction recovery error is bounded by its own worst case over that set.

For the conformal step, abbreviate the true-pair score by $Z_b = S_b(\pi_b^\star)$. Suppose first that $k \le m$, so $\widehat{q} = Z_{(k)}$, the kth order statistic of the m calibration scores. Conditional on the maps fitted on the disjoint fitting split and the locked protocol, exchangeability of $Z_1, \dots, Z_m, Z_{m+1}$ gives

$$\begin{aligned} \mathbb{P}\{Z_{m+1} \le \widehat{q}\} &\ge \frac{k}{m+1} \\ &\ge 1 - \alpha. \end{aligned} \tag{19}$$

Random tie-breaking proves the rank statement, and retaining ties with the weak inequality can only increase coverage. If $k > m$, setting $\widehat{q} = +\infty$ instead gives $\mathbb{P}\{Z_{m+1} \le \widehat{q}\} = 1$, so the same coverage conclusion holds.

Every finite normalized assignment score is at most n. Fix any $\pi = \pi_1$ and extend it to permutations $\pi_1, \dots, \pi_n$ whose edges partition the complete bipartite graph. Because $C_b(\pi_s) \ge C_b^{\min}$ for $s \ge 2$,

$$\begin{aligned} \sum_{s=1}^{n} C_b(\pi_s) &= \sum_{i,j} c_{b,ij} \\ &= n\overline{C}_b, \\ C_b(\pi) + (n-1)C_b^{\min} &\le n\overline{C}_b, \\ C_b(\pi) - C_b^{\min} &\le n(\overline{C}_b - C_b^{\min}), \\ 0 \le S_b(\pi) &\le n. \end{aligned}$$

Thus the finite threshold n would also admit every assignment. The value $+\infty$ is the protocol's explicit all-assignments convention.

On the event $Z_{m+1} \le \widehat{q}$, the true evaluation correspondence is in $\mathcal{P}_{m+1}(\widehat{q})$. Applying (18) to every fixed $a \in \mathcal{A}$ proves all inequalities in (16) on that single event. No union bound over methods is needed. Finally, an integer-feasible assignment and the same weak-Lagrangian argument as in (3) give, respectively,

$$\begin{aligned} L_{a,b} &\le G_{a,b}(\widehat{q}), \\ G_{a,b}(\widehat{q}) &\le U_{a,b}. \end{aligned}$$

Intersecting these deterministic inequalities with the conformal event proves the stated simultaneous upper certificate. □

Reporting both endpoints, their width, the realized true-pair slack, and tightening relative to the exact unconstrained label-count transport bound distinguishes statistical coverage from numerical informativeness. None of these statements transfers across corruptions or to unpaired natural shifts without a separate exchangeability and observation argument.

Proof of Corollary 1. The edge reward is

$$r_{ij}^{a,c} = \frac{1}{n}\left[\mathbf{1}\{h(T_a(t_i)) \ne h(r_j)\} - \mathbf{1}\{h(T_c(t_i)) \ne h(r_j)\}\right],$$

which is additive and signed, so the same assignment and Lagrangian arguments apply without clipping negative rewards. On the single calibrated feasibility event, every fixed contrast satisfies

$$E_a(\pi^\star) - E_c(\pi^\star) \le H_{a,c}(\widehat{q}),$$
$$H_{a,c}(\widehat{q}) \le \overline{H}_{a,c}.$$

This proves (11) without an additional pairwise union bound. □

Separate absolute-error maxima generally use different correspondences, so comparing their upper bounds alone cannot certify a signed method advantage. On that same event, the batchwise-fallback rule either chooses identity or chooses a with

$$E_a(\pi^\star) - E_0(\pi^\star) \le \overline{H}_{a,0},$$
$$\overline{H}_{a,0} \le 0.$$

Hence harm is possible only off the calibrated event, proving

$$\mathbb{P}\{E_{\widehat{a},m+1}(\pi^\star_{m+1}) > E_{0,m+1}(\pi^\star_{m+1})\} \le \alpha, \tag{20}$$

where $\widehat{a}$ is any fixed-rule choice among methods with $\overline{H}_{a,0} \le 0$, and identity is chosen if that set is empty. This fallback policy is a theoretical consequence and is not evaluated. If both ordered pairs are bounded, then

$$-\overline{H}_{c,a} \le E_a(\pi^\star) - E_c(\pi^\star),$$
$$E_a(\pi^\star) - E_c(\pi^\star) \le \overline{H}_{a,c}.$$

If valid bounds are available for every ordered pair, set $\overline{H}_{a,a} = 0$ and choose $\widehat{a} \in \arg\min_a \max_c \overline{H}_{a,c}$. On the calibrated event,

$$E_{\widehat{a}}(\pi^\star) - \min_c E_c(\pi^\star) = \max_c \{E_{\widehat{a}}(\pi^\star) - E_c(\pi^\star)\}$$
$$\le \min_a \max_c \overline{H}_{a,c}.$$

The inequality is a batchwise robust-regret statement. No corresponding policy was evaluated. The accompanying probability statement has the same marginal, one-future-batch scope as Proposition 3. In particular, it does not condition on which action was selected.

Proposition 4 (Comparison without correspondence identification)**.** *For $n = m + k$ with even $m \ge 2$ and $k \ge 1$, there is a finite binary-readout assignment problem, with identity anchor $\pi_0 = \mathrm{id}$ and budget $B = 0$, having exactly $m!$ feasible correspondences, anchored prediction ambiguity $\Gamma_{h,\pi_0}(0) = m/n$, and two fixed prediction rules a, c whose absolute errors both vary across the feasible correspondences, yet*

$$E_a(\pi) - E_c(\pi) = -k/n \qquad \textit{for every feasible } \pi.$$

Thus a strict recovery ordering may be identified while multiple correspondences and individual desired predictions remain unidentified.

Proof. Partition the n target and reference atoms into an uncertain block of size m and a fixed block of size k. Assign cost zero to every edge within the uncertain block and to each identity edge in the fixed block, cost one to all other edges, and set $B = 0$. Exactly the $m!$ permutations of the uncertain block are feasible. Set $h(r_j) = 0$ for $m/2$ uncertain reference atoms and $h(r_j) = 1$ for the remainder. Relative to the identity anchor π_0, a permutation exchanging these two readout groups changes every desired prediction in that block, while the fixed block cannot change. Hence $\Gamma_{h,\pi_0}(0) = m/n$.

On the uncertain block, let both rules predict the identity-indexed readout, $a(t_i) = c(t_i) = h(r_i)$. Their common uncertain-block error count, say $R(\pi)$, depends on the correspondence: it is zero for the identity and m for a permutation exchanging the two readout groups. On the fixed block, let $a(t_i) = h(r_i)$ and $c(t_i) = 1 - h(r_i)$. Thus

$$E_a(\pi) = R(\pi)/n,$$
$$E_c(\pi) = \{R(\pi) + k\}/n.$$

Both absolute errors vary, but their contrast is $-k/n$ for every feasible assignment. □

This construction separates identification of a correspondence from identification of a comparison. It neither equates the high-ambiguity unpaired batches with the paired empirical panels nor establishes a new general impossibility theorem.

A.6 APPROXIMATELY DISTINGUISHABLE OBSERVATION LAWS

The exact result above assumes that the complete observation available to the rule has the same law in both worlds. Let those complete laws instead be Q_0, Q_1 on a common measurable space. An observation includes every calibration sample, the query when it is available to the rule, and all random side information. Using the total variation distance of one marginal draw would generally be incorrect. Let $\ell_j(o)$ be the desired prediction in world j and $D = \{o : \ell_0(o) \neq \ell_1(o)\}$. Let $\mu = Q_0 + Q_1$ and $q_j = dQ_j/d\mu$. Under an equal world prior, the unrestricted Bayes value is

$$B_{\text{Bayes}} = 1 - \frac{1}{2}\int_D \min\{q_0(o), q_1(o)\}\, d\mu(o). \tag{21}$$

Where the desired labels agree, the rule can satisfy both worlds; where they differ, it chooses the label with greater local world mass. Consequently, for any rule with world agreements A_0, A_1,

$$\begin{aligned} \min(A_0, A_1) &\leq B_{\text{Bayes}}, \\ B_{\text{Bayes}} &\leq 1 - \frac{1}{2}\left(\max\{\delta_0, \delta_1\} - \text{TV}(Q_0, Q_1)\right)_+ . \end{aligned} \tag{22}$$

Here $\delta_j = Q_j(D)$ and $\text{TV}(Q_0, Q_1) = \sup_A |Q_0(A) - Q_1(A)|$. The quantity B_{Bayes} is the exact unrestricted equal-prior Bayes value, but it is only an upper bound on worst-case performance across worlds unless a separate equalizer rule proves minimax attainability. Equation (22) is the task-aware form of a two-point testing argument (Le Cam, 1986). The discrete positive control computes (21) exactly for calibration sizes $0, 1, 4, 16, 64, 256$, and $1{,}024$. In this control, as the complete calibration law becomes distinguishable, the finite-grid equal-prior Bayes value rises from 50.00% at $n = 0$ to 73.89% at $n = 1{,}024$. For the Bernoulli control specified in Section D, its asymptotic limit is 100%. When the complete laws coincide, its value is $1 - \delta/2$.

A.7 POSTERIOR-MODE BASELINE AND EXACT-ALIGNMENT CAPACITY

Proposition 5 (Posterior-mode baseline and exact-alignment capacity)**.** *Let $\theta \sim \nu$ index worlds with a common target marginal P_t, and set*

$$\ell_\theta(x) = \tau(F_\theta^{-1}x), \qquad q_y^\nu(x) = \Pr_{\theta\sim\nu}\{\ell_\theta(x) = y\}.$$

For finite $\mathcal{Y}$, the optimal prior-average agreement over all world-independent randomized predictors is

$$C_{\text{pred}}(\nu) = \int \max_{y\in\mathcal{Y}} q_y^\nu(x)\, dP_t(x). \tag{23}$$

Now let $\tau = g$, require $T_\# P_t = P_r$, and write $\pi_y = P_r\{g(Z) = y\}$. If all measurable transports are admissible, P_t and every positive-mass conditional law $P_r(\cdot \mid g(Z) = y)$ are nonatomic on standard Borel spaces, then

$$C_{\text{align}}(\nu) = \sup_{\substack{(B_y)_y \text{ measurable partition} \\ P_t(B_y)=\pi_y}} \sum_y \int_{B_y} q_y^\nu(x)\, dP_t(x) \leq C_{\text{pred}}(\nu). \tag{24}$$

For affine or otherwise restricted corrections, the partition value is only an upper bound.

Proof. For an arbitrary randomized prediction kernel $p(y \mid x)$, its prior-average agreement is

$$\int \sum_y p(y \mid x) q_y^\nu(x)\, dP_t(x) \leq \int \max_y q_y^\nu(x)\, dP_t(x).$$

The posterior-mode prediction attains equality, proving the unrestricted predictor ceiling. The ceiling is the common-observation specialization of Dong et al. (2026, Theorem 1 and Corollary 1) and serves here as a baseline.

There is a tighter characterization when $\tau = g$ and each realized correction satisfies $T_\# P_t = P_r$. Put $\pi_y = P_r\{g(Z) = y\}$. Every such T induces cells $B_y = \{x : g(Tx) = y\}$ satisfying $P_t(B_y) = \pi_y$, and its prior-average agreement is $\sum_y \int_{B_y} q_y^\nu\, dP_t$. Conversely, assume P_t is nonatomic on a

standard Borel space and, for every y with $\pi_y > 0$, the conditional law $P_r(\cdot \mid C_y)$ on the standard-Borel subspace $C_y = \{z : g(z) = y\}$ is nonatomic. For every measurable partition $(B_y)_y$ with $P_t(B_y) = P_r(C_y) = \pi_y$, the normalized restrictions on each pair of positive-mass cells (B_y, C_y) are measure-isomorphic modulo null sets. Zero-mass cells may be ignored. Combining these maps across y constructs a measurable T up to null sets with $T_\# P_t = P_r$ and $g(Tx) = y$ on B_y. This use of nonatomic standard-Borel isomorphism is standard. See, for example, Kallenberg (2021). Thus, when all measurable transports are admissible,

$$C_{\text{align}}(\nu) = \sup_{(B_y):P_t(B_y)=\pi_y} \sum_y \int_{B_y} q_y^\nu(x)\, dP_t(x). \tag{25}$$

For an affine or otherwise restricted correction class, (25) is only an upper bound.

Introduce multipliers λ_y for the cell-mass constraints. Maximizing the Lagrangian pointwise over the assigned label gives

$$\sum_y \lambda_y \pi_y + \int \max_y \{q_y^\nu(x) - \lambda_y\}\, dP_t(x).$$

Kantorovich duality gives the relaxed finite-label transport dual (Villani, 2009). Because P_t is nonatomic and the label set is finite, the Dvoretzky–Wald–Wolfowitz purification theorem realizes the same cell masses and objective with a measurable partition (Dvoretzky et al., 1951). Thus the following strong dual holds:

$$C_{\text{align}}(\nu) = \inf_{\lambda \in \mathbb{R}^{|\mathcal{Y}|}} \left[\sum_y \lambda_y \pi_y + \int \max_y \{q_y^\nu(x) - \lambda_y\}\, dP_t(x) \right]. \tag{26}$$

Neither the ambiguity set nor its prior is determined by the reference and target marginals. They are assumptions of the decision problem. □

A.8 Finite exact-quota control

The following indexed-atom example makes the distinction between unrestricted prediction and output-mass-constrained alignment strict. It is a finite integer-quota analogue of Proposition 5, not evidence for its nonatomic population assumptions. Give four target atoms equal mass, let the reference readout have label counts $(2, 2)$, and consider three equally likely correspondence worlds with desired-label vectors

$$(0, 0, 1, 1), \qquad (0, 1, 0, 1), \qquad (1, 0, 0, 1).$$

Every world respects the reference output masses. The posterior probability of label zero is $(2/3, 2/3, 2/3, 0)$, so unconstrained posterior-mode prediction has value

$$C_{\text{pred}} = \frac{1}{4}\left(\frac{2}{3} + \frac{2}{3} + \frac{2}{3} + 1\right) = \frac{3}{4}.$$

An exactly aligned correction must instead predict each label on exactly two atoms. Its best prior-average assignment labels any two of the first three atoms zero and the remaining two atoms one, giving

$$C_{\text{align}} = \frac{1}{4}\left(\frac{2}{3} + \frac{2}{3} + \frac{1}{3} + 1\right) = \frac{2}{3}.$$

The strict gap reflects the output-mass conservation mechanism underlying the binary obstruction of Wu et al. (2019, Proposition 3.1). The additional object here is the optimal prior/world agreement under the full quota.

The same example checks the finite-world minimax dual below. Use as actions the six labelings with counts $(2, 2)$. Uniformly mixing the three displayed world labelings gives agreement $(1 + 1/2 + 1/2)/3 = 2/3$ against each world. Conversely, the uniform world prior has quota-constrained Bayes value $2/3$, so it is a matching least-favorable-prior upper bound. Thus both matrix-game programs attain $2/3$. The companion executable enumerates all six actions, solves both LPs, and verifies a numerical primal–dual gap below 10^{-10}.

A.9 FINITE-WORLD CONSTRAINED MINIMAX DUAL

Corollary 2 (Finite-world constrained minimax dual)**.** *Under the exact-alignment assumptions of Proposition 5, let* $\Theta = \{1, \dots, m\}$ *and let* $\mathcal{T}_{\text{align}}$ *be the measurable exact transports. Define* $a_\theta(T) = \int \mathbf{1}\{h(Tx) = \ell_\theta(x)\}\, dP_t(x)$*, and let* $\Delta_f(\mathcal{T}_{\text{align}})$ *denote finite mixtures of such transports. Then*

$$\sup_{Q \in \Delta_f(\mathcal{T}_{\text{align}})} \min_{\theta \in \Theta} \mathbb{E}_{T \sim Q} a_\theta(T) = \min_{\nu \in \Delta(\Theta)} C_{\text{align}}(\nu). \tag{27}$$

The prior-side minimum is attained. The correction-side supremum is attained if and only if at least one maximizing point of the closed convex hull is realizable by an admissible finite mixture.

Proof. Let $\Theta = \{1, \dots, m\}$ and write

$$a(T) = (a_1(T), \dots, a_m(T)) \in [0,1]^m, \qquad a_\theta(T) = \int \mathbf{1}\{h(Tx) = \ell_\theta(x)\}\, dP_t(x).$$

Let $S = \{a(T) : T \in \mathcal{T}_{\text{align}}\}$ and $K = \overline{\text{conv}}(S)$. The set K is compact and convex in the finite-dimensional payoff space. Finite mixtures of corrections realize $\text{conv}(S)$, whose closure has the same supremum for the continuous function $v \mapsto \min_\theta v_\theta$. Consequently,

$$\begin{aligned}
\sup_{Q \in \Delta_f(\mathcal{T}_{\text{align}})} \min_{\theta} \mathbb{E}_{T \sim Q} a_\theta(T) &= \max_{v \in K} \min_{\theta} v_\theta \\
&= \min_{\nu \in \Delta(\Theta)} \max_{v \in K} \nu^\top v \\
&= \min_{\nu \in \Delta(\Theta)} \sup_{T \in \mathcal{T}_{\text{align}}} \sum_\theta \nu_\theta a_\theta(T) \\
&= \min_{\nu \in \Delta(\Theta)} C_{\text{align}}(\nu).
\end{aligned}$$

The middle equality is the finite-dimensional minimax theorem (von Neumann, 1928). The last equality follows because the prior-weighted payoff is $\int q^\nu_{g(Tx)}(x)\, dP_t(x)$ and the assumptions of Proposition 5 identify its supremum over exact transports with the mass-constrained partition value.

The function C_{align} is Lipschitz on the finite simplex: because every payoff coordinate lies in $[0,1]$, $|C_{\text{align}}(\nu) - C_{\text{align}}(\nu')| \le \|\nu - \nu'\|_1$. Hence the least-favorable prior is attained. The maximizing point in K need only be a limit of finite-mixture payoff vectors, so the correction-side expression is a supremum. It becomes a maximum when that point is realized by an admissible finite mixture, as it is for a finite correction class.

For a finite correction set $\mathcal{A}$, put $M_{T\theta} = a_\theta(T)$. The same identity is the following primal–dual pair:

$$\begin{aligned}
&\underset{w, v}{\text{maximize}} \quad v \qquad \text{subject to} \quad \sum_{T \in \mathcal{A}} w_T M_{T\theta} \ge v \quad (\theta \in \Theta), \quad \sum_{T \in \mathcal{A}} w_T = 1, \quad w_T \ge 0, \\
&\underset{\nu, z}{\text{minimize}} \quad z \qquad \text{subject to} \quad \sum_{\theta} \nu_\theta M_{T\theta} \le z \quad (T \in \mathcal{A}), \quad \sum_{\theta} \nu_\theta = 1, \quad \nu_\theta \ge 0.
\end{aligned}$$

Both programs attain their common value. With an incomplete candidate set of corrections, the primal is a lower bound for the full game. A full-game upper bound requires certifying every dual best-response constraint. □

A.10 EXACT COMPACT-GROUP VALUE

Let G be a second-countable compact Hausdorff group with normalized Haar Borel probability μ_G, acting jointly Borel measurably on a standard-Borel feature space and preserving P_r (Haar, 1933; Eaton & George, 2021). Let the possible worlds be $X = HZ$ for $H \in G$. Assume that $\tau = g$, that all inverse group actions are admissible corrections, and that the finite-label task is equivariant:

$$g(Hz) = \rho(H) g(z),$$

where $\rho : G \to \text{Sym}(\mathcal{Y})$ is a measurable action. Write $\mathcal{O}_y$ for the label orbit of y.

Let $\mathcal{A}$ range over randomized correction rules whose law is common to all observationally indistinguishable worlds, and define their randomized minimax agreement by

$$V_G := \sup_{\mathcal{A}} \inf_{H \in G} \Pr_{Z \sim P_r,\, T \sim \mathcal{A}} \{g(T(HZ)) = g(Z)\},$$

where the probability also includes the rule's internal randomization.

To justify conditioning on the observed target, let H be Haar distributed independently of Z. For measurable $B \subseteq G$ and measurable A in the feature space, invariance of P_r gives

$$\begin{aligned}\Pr\{H \in B,\, HZ \in A\} &= \int_B P_r(H^{-1}A)\, d\mu_G(H) \\ &= \mu_G(B) P_r(A).\end{aligned}$$

Thus H and $X = HZ$ are independent. Hence, for fixed observed x, Haar-random H makes $g(H^{-1}x)$ uniform on $\mathcal{O}_{g(x)}$: invariance of Haar measure makes all elements of a finite orbit equiprobable. Any world-independent prediction therefore has Haar-average agreement at most $1/|\mathcal{O}_{g(x)}|$. Integrating gives

$$V_G \le \mathbb{E}_{Z \sim P_r} \frac{1}{|\mathcal{O}_{g(Z)}|}. \tag{28}$$

For the reverse inequality, draw R from Haar measure and apply $T_R(x) = R^{-1}x$. In every fixed world H, the relative action $R^{-1}H$ is Haar distributed. Conditional on $g(Z) = y$, its prediction agrees with y with probability $1/|\mathcal{O}_y|$. Thus the right-hand side of (28) is attained in every fixed world. It is therefore both the Haar-Bayes value and the randomized minimax value. A transitive action on K labels gives $V_G = 1/K$.

Notice that a fixed representative need not be an equalizer. In the cyclic experiment, identity has mean agreement $1/K$ under a uniform prior on worlds, but its worst-case agreement is zero. The minimax attainment is witnessed by the uniform random mixture of group inverses, whose expected agreement is $1/K$ separately in every world.

A.11 Gaussian binary closed form

Let $Z \sim \mathcal{N}(0, \Sigma)$ with $w^\top \Sigma w > 0$, let $g_w(z) = \operatorname{sign}(w^\top z)$, and suppose $H \Sigma H^\top = \Sigma$. Assume the identity and H^{-1} corrections are admissible. The scores

$$U = w^\top Z, \qquad V = w^\top H^{-1} Z$$

are centered jointly Gaussian. Symmetry implies $H^{-1}\Sigma H^{-\top} = \Sigma$, so both variances equal $w^\top \Sigma w$, while

$$\operatorname{Cov}(U, V) = w^\top \Sigma H^{-\top} w.$$

Their correlation is consequently

$$\begin{aligned}\rho_H &= \frac{w^\top \Sigma H^{-\top} w}{\sqrt{(w^\top \Sigma w)(w^\top H^{-1} \Sigma H^{-\top} w)}} \\ &= \frac{w^\top \Sigma H^{-\top} w}{w^\top \Sigma w}.\end{aligned}$$

The final equality uses $H^{-1}\Sigma H^{-\top} = \Sigma$. After whitening this bivariate Gaussian, rotational symmetry shows that the two signs differ on wedges of total angular measure $2\arccos(\rho_H)$ out of 2π. Hence

$$\begin{aligned}\delta_g(H) &= \frac{\arccos(\rho_H)}{\pi}, \\ V_2(H) &= 1 - \frac{\arccos(\rho_H)}{2\pi}.\end{aligned}$$

The resulting expression is the classical Gaussian sign-disagreement identity (Sheppard, 1899). For $\Sigma = I$ and angle $\theta \in [0, \pi]$ between the two effective readout normals, this reduces to $V_2 = 1 - \theta/(2\pi)$.

A.12 READOUT-WEIGHTED VALUE OF PAIRED CORRESPONDENCE

The final result is a paired diagnostic, not a marginal-only ceiling. Let X, Y be centered paired reference and target features with exact positive-definite population covariances Σ_r, Σ_t. Here specialize the readout to the deterministically tie-broken affine-logit classifier $h(z) = \arg\max_{k\in[K]}(Wz+b)_k$, with $W \in \mathbb{R}^{K\times d}$ and $b \in \mathbb{R}^K$. The bias cancels from clean–corrected logit differences, and centering by $I - \mathbf{1}\mathbf{1}^\top/K$ preserves every class-score difference. Define

$$u = \Sigma_r^{-1/2}X, \qquad v = \Sigma_t^{-1/2}Y, \qquad C = \mathbb{E}[vu^\top],$$

so $\mathbb{E}[uu^\top] = \mathbb{E}[vv^\top] = I$. Let

$$\widetilde{W} = \left(I - \frac{\mathbf{1}\mathbf{1}^\top}{K}\right)W, \qquad S = \Sigma_r^{1/2}\widetilde{W}^\top\widetilde{W}\Sigma_r^{1/2}.$$

The covariance-matching orthogonal factor $R \in O(d)$ maps Y to $\Sigma_r^{1/2}Rv$. Its expected squared Euclidean centered-logit error is

$$\begin{aligned} L_g(R) &= \mathbb{E}\left\|\widetilde{W}\Sigma_r^{1/2}(Rv-u)\right\|_2^2 \\ &= \operatorname{tr}(R^\top SR) + \operatorname{tr}(S) - 2\operatorname{tr}(RCS) \\ &= 2\operatorname{tr}(S) - 2\operatorname{tr}(RCS), \end{aligned}$$

where orthogonality gives $\operatorname{tr}(R^\top SR) = \operatorname{tr}(S)$. By the classical orthogonal Procrustes solution (Schönemann, 1966), if $CS = U\operatorname{diag}(s)V^\top$, von Neumann's trace inequality (Mirsky, 1975) gives

$$\max_{R\in O(d)} \operatorname{tr}(RCS) = \|CS\|_*,$$

$$R^* = VU^\top.$$

Therefore

$$L_g^* = 2\operatorname{tr}(S) - 2\|CS\|_*.$$

The displayed L_g sums squared error over the K centered logits. A per-logit mean squared error is L_g/K. If the implementation whitens using shrunk or eigenvalue-clipped covariances, its empirical whitened covariances are not exactly identity. In that case this formula is exact for the fitted covariance surrogate, not necessarily for the raw finite-sample empirical loss.

Rank constraint on empirical covariance. A centered covariance from n samples in $\mathbb{R}^d$ has rank at most $n-1$. Consequently, in whitened coordinates,

$$\inf_{\operatorname{rank}(B)\le n-1} \frac{\|B - I_d\|_F}{\|I_d\|_F} = \sqrt{\frac{d-n+1}{d}}, \qquad n \le d. \tag{29}$$

The Eckart–Young argument (Eckart & Young, 1936) leaves $d-n+1$ unit residuals. Shrinkage can make a surrogate invertible but cannot identify unsupported directions. The next result gives the complementary readout-weighted finite-pair bound.

Proposition 6 (Finite-pair stability of readout-weighted Procrustes)**.** *Retain the population-whitened coordinates above. Let $(u_i, v_i)_{i=1}^n$ be i.i.d. copies of (u, v), let*

$$\widehat{C} = \frac{1}{n}\sum_{i=1}^n v_i u_i^\top,$$

and treat S as fixed. Choose $R^\star$ arbitrarily and $\widehat{R}$ via any measurable maximizer selection:

$$R^\star \in \arg\max_{R\in O(d)} \operatorname{tr}(RCS), \qquad \widehat{R} \in \arg\max_{R\in O(d)} \operatorname{tr}(R\widehat{C}S).$$

If $r = \operatorname{rank}(S) \le \min\{d, K-1\}$, then

$$0 \le L_g(\widehat{R}) - L_g(R^\star) \le 4\|(\widehat{C}-C)S\|_* \le 4\sqrt{r}\,\|(\widehat{C}-C)S\|_F. \tag{30}$$

If, in addition,

$$\sigma_S^2 = \mathbb{E}\|(vu^\top - C)S\|_F^2 < \infty, \tag{31}$$

then

$$\mathbb{E}\Big[L_g(\widehat{R}) - L_g(R^\star)\Big] \le 4\sigma_S\sqrt{\frac{r}{n}}, \tag{32}$$

and, for every $\alpha \in (0,1)$*, with probability at least* $1-\alpha$*,*

$$L_g(\widehat{R}) - L_g(R^\star) \le 4\sigma_S\sqrt{\frac{r}{n\alpha}}. \tag{33}$$

Proof. Put $A = CS$, $\widehat{A} = \widehat{C}S$, and $E = \widehat{A} - A$. Population optimality gives the first inequality in (30), while empirical optimality gives $\operatorname{tr}(\widehat{R}\widehat{A}) \ge \operatorname{tr}(R^\star\widehat{A})$. Hence

$$\begin{aligned}\operatorname{tr}(R^\star A) - \operatorname{tr}(\widehat{R}A) &\le \operatorname{tr}\{(\widehat{R} - R^\star)E\}\\ &\le |\operatorname{tr}(\widehat{R}E)| + |\operatorname{tr}(R^\star E)|\\ &\le 2\|E\|_*.\end{aligned}$$

The last step is nuclear/operator norm duality and $\|R\|_{\mathrm{op}} = 1$ for $R \in O(d)$. Multiplying by two proves the nuclear-norm bound. Since $\operatorname{rank}(E) \le r$, $\|E\|_* \le \sqrt{r}\|E\|_F$.

For $X_i = (v_i u_i^\top - C)S$, independence, centering, and (31) give

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n X_i\right\|_F^2 = \frac{\sigma_S^2}{n}.$$

Jensen's inequality gives

$$\begin{aligned}\mathbb{E}\|E\|_F &\le \{\mathbb{E}\|E\|_F^2\}^{1/2}\\ &\le \frac{\sigma_S}{\sqrt{n}},\end{aligned}$$

which yields (32). Markov's inequality gives

$$\mathbb{P}\left\{\|E\|_F > \frac{\sigma_S}{\sqrt{n\alpha}}\right\} \le \alpha,$$

$$\|E\|_F \le \frac{\sigma_S}{\sqrt{n\alpha}} \quad \text{with probability at least } 1-\alpha.$$

Combining the last line with the deterministic bound proves (33). □

This proposition controls excess squared centered-logit loss over the best population orthogonal correction. It does not control $L_g(\widehat{R})$ itself or top-1 disagreement without an additional margin condition. It is an oracle-whitening result: the population centering, Σ_r, Σ_t, and S are fixed, and only C is estimated from the pairs. It therefore does not directly describe the implementation, which estimates means and OAS-shrunk covariances from the same sample (Chen et al., 2010). Extending the rate to that estimator would also require covariance-square-root perturbation bounds, lower spectral assumptions, and control of the resulting dependence. The elementary proposition is included to isolate the paired estimator's task-weighted sampling error, not as a new general Procrustes theorem.

B Complete experimental protocol

B.1 Models, data, and frozen classifiers

Table 3 records the exact representations used in the completed study. Features were produced by fresh forward passes in float32, with batch size 128 and eight data-loader workers. Each model's specified native preprocessing transform was applied before feature extraction. No encoder was updated in the frozen-feature campaign. The three encoder families are CLIP, DINOv3, and AugReg Vision Transformers (Radford et al., 2021; Siméoni et al., 2025; Dosovitskiy et al., 2021; Steiner et al., 2022).

Table 3: Frozen feature extractors. The artifact records the locked model revisions, feature definitions, and executed checkpoint and preprocessing hashes.

Encoder	Pretraining setting	Feature	d
CLIP ViT-L/14	image–text pretraining	projected class token	768
DINOv3 ViT-L/16	self-supervised LVD-1689M	final class token	1,024
AugReg ViT-B/16	ImageNet-21K $\rightarrow$ ImageNet-1K	pre-logits	768

For every encoder–dataset pair, one multinomial logistic regression classifier was fitted on clean features and then frozen. Fitting used scikit-learn's LBFGS solver (Pedregosa et al., 2011), $C = 1$, an intercept, at most 2,000 iterations, tolerance 10^{-3}, and a fail-on-nonconvergence policy. Clean-fit feature means and population standard deviations were folded into the final classifier. CIFAR-100 used all 50,000 training images (Krizhevsky, 2009). For ImageNet (Deng et al., 2009), this study used a class-stratified subset of 25 validation images per class. These identities were excluded from all subsequent reference, calibration, and evaluation sets.

The resulting benchmark sizes and disjoint evaluation protocol are summarized in Table 4. Both corruption benchmarks follow the standard CIFAR-100-C/ImageNet-C construction (Hendrycks & Dietterich, 2019).

Table 4: Original marginal-alignment benchmark data and split protocol. Task metrics use every evaluation row. Distribution metrics use all 3,000 CIFAR-100 rows and the first 4,096 ImageNet rows.

Dataset	Classes	Corruptions	Raw examples /condition	Evaluation examples	Severities
CIFAR-100-C	100	19	10,000	3,000	1, 3, 5
ImageNet-C	1,000	15	50,000	10,000	1, 3, 5

The split generator enforces four invariants. Evaluation identities are disjoint from all fitting subsets; the evaluation set is fixed across calibration sizes; calibration prefixes are nested; and unpaired reference and target calibration sets are sampled independently. Identity-matched subsets are constructed only for paired diagnostics. A shuffled-pair control permutes the assignment while preserving both sample multisets exactly.

B.2 Alignment methods and information available

Covariance eigensolvers used a relative floor of $\max(10^{-6}\bar{\lambda}, \epsilon_{\mathrm{mach}})$, where $\bar{\lambda}$ is the mean eigenvalue scale. The amended Bures middle square root uses a positive semidefinite factor SVD. OAS denotes oracle-approximating shrinkage (Chen et al., 2010). Ridge penalties and retained ranks were fixed before the reported run.

No target joint structure. The identity baseline uses no target statistic. Mean alignment uses coordinate means; diagonal alignment adds coordinate variances with a 10^{-6} variance floor; coordinate quantile alignment uses at most 257 knots per coordinate.

Unpaired distribution structure. The CORAL and Bures implementations use joint first and second moments with OAS covariance estimates (Sun et al., 2016; Chen et al., 2010; Bhatia et al., 2019). Empirical Bures omits shrinkage. The low-rank variants use rank 4; rank 64 is evaluated only when $n \geq 1{,}024$.

Assignments. Paired ridge and paired low-rank maps receive either true or deliberately shuffled assignments; both use the relative ridge factor $r = 0.01$, and the latter uses rank 64. The unpaired Bures control uses the same reference and target multisets without the assignments.

Quantile implementation. Coordinate-wise quantile matching follows the standard quantile-normalization idea (Bolstad et al., 2003). For a calibration set of size n and cap $K_q = 257$, each coordinate retains all ordered observations at empirical mid-ranks when $n \leq K_q$. When $n > K_q$, it instead evaluates the `linear` empirical-quantile method in NumPy 2.5.1 (Harris et al., 2020) at $p_j = (j + 1/2)/K_q$, $j = 0, \ldots, K_q - 1$. Repeated target values or quantiles are collapsed and assigned the mean of their tied probabilities. At application time, linear interpolation maps a held-out target coordinate to its fitted target probability and then to the fitted reference quantile. Values beyond the fitted target range are clipped to endpoint probabilities. The map stores no rows or assignments and is evaluated on disjoint data. The cap K_q was fixed before the run.

Comparator scope. The real-data grid compares frozen-feature alignment rules. Parameter-updating test-time adaptation is outside this setting. The retained real-feature paired suite tests affine ridge and rank-64 reduced-rank ridge (Izenman, 1975) with relative ridge factor $r = 0.01$, giving $\lambda = r\|X_{sc}\|_F^2/d$; ranks and the relative ridge factor are fixed. The covariance-rank suite compares full and rank-restricted OAS Bures with unshrunk empirical Bures. Procrustes appears only in controlled full-rank simulations because its real-feature version failed the prespecified identification checks.

B.3 Suites, metrics, and inference

The primary marginal suite used all severity-5 corruptions, methods from no alignment through Bures, $n = 4{,}096$, and split seeds 0–4. Severity breadth used all corruptions at severities 1, 3, and 5 with seeds 40–42. The finite-sample suite used five corruptions (Gaussian noise, motion blur, fog, contrast, and JPEG compression), all three severities, $n \in \{64, 256, 1{,}024, 4{,}096\}$, and seeds 10–14. The rank suite used those five corruptions at severity 5, the same four sample sizes, and seeds 20–22. The assignment-noise sweep used the same five severity-5 corruptions, pair budgets $\{4, 16, 64, 256, 1{,}024, 4{,}096\}$, assignment-noise rates $\{0, 0.10, 0.25, 0.50, 1\}$, and seeds 30–34. The confirmatory correspondence suite instead used every severity-5 corruption, $4{,}096$ pairs, and seeds 35–37. Exact row counts were 9,000 finite-sample, 8,250 assignment-noise, 4,590 severity, 3,060 primary, 1,620 rank, 918 confirmatory, and 324 representative higher-order rows, totaling 27,762 real-data records.

The two task metrics are ground-truth top-1 accuracy and agreement with the frozen classifier's prediction on the clean version of the same evaluation identity. Distribution diagnostics include coordinate-variance and full-covariance relative errors, squared Gaussian-kernel maximum mean discrepancy (MMD), approximated with 512 random Fourier features and a median-heuristic bandwidth from up to 512 pooled rows (Rahimi & Recht, 2007; Gretton et al., 2012), negative log-likelihood, and expected calibration error (ECE) with 15 equal-width confidence bins (Guo et al., 2017). Distribution diagnostics use at most 4,096 evaluation rows. All accuracies and agreements below are percentages. Differences are percentage points (pp).

At severity 5, inference is separate by dataset after averaging split and encoder repetitions within corruption. Percentile intervals use 20,000 bootstrap resamples of the 19 CIFAR-100-C or 15 ImageNet-C corruption means (Davison & Hinkley, 1997). Exact two-sided sign-flip tests enumerate all 2^m sign patterns (Lehmann & Romano, 2005), with Holm correction over ten contrasts (Holm, 1979). The intervals interpret the corruptions as i.i.d. draws from a conceptual population; otherwise, they are fixed-panel resampling intervals. Exact sign-flip validity requires coordinatewise sign invariance, as under independent symmetric cluster effects. Neither procedure covers arbitrary future shifts. Cell counts for marginal/correspondence contrasts are 285/171 for CIFAR-100-C and 225/135 for ImageNet-C. Split-level standard deviations are descriptive, not confidence intervals.

C Complete real-data results

C.1 Primary marginal hierarchy

Table 5: **Severity-5 agreement/accuracy (%).** Each cell is clean-prediction agreement / ground-truth top-1 accuracy at $n = 4{,}096$, macro-averaged over corruptions and split seeds 0–4. Agreement is the confirmatory endpoint. The best agreement and accuracy in each row are bold.

Encoder	No align.	Mean	Diagonal	Quantile	CORAL	Bures
CIFAR-100-C						
CLIP	55.46/54.45	60.31/59.16	60.44/59.30	60.42/59.28	59.27/58.14	**61.13/59.90**
DINOv3	70.74/69.21	73.17/71.58	73.24/71.65	73.29/71.70	72.68/71.12	**74.09/72.45**
ViT-B/16	56.20/54.87	60.64/59.12	60.69/59.18	60.73/59.21	60.33/58.90	**62.21/60.71**
ImageNet-C						
CLIP	40.49/38.22	47.19/44.28	47.31/44.40	47.32/44.40	47.61/44.58	**48.24/45.08**
DINOv3	69.64/64.36	71.94/66.33	72.09/66.48	72.12/66.50	73.14/67.58	**73.32/67.66**
ViT-B/16	51.50/48.35	53.42/50.00	53.58/50.17	53.59/50.17	55.45/51.77	**55.69/51.92**

Table 6: All prespecified primary contrasts. Effects are percentage points in clean-prediction agreement. Intervals are pointwise 95% corruption-cluster bootstrap CIs. The p-values are Holm-adjusted over the ten rows.

Dataset	Contrast	Effect	95% bootstrap CI	Holm p	Positive clusters
CIFAR	Diagonal − no alignment	+3.99	[3.11, 4.91]	< 0.001	19/19
ImageNet	Diagonal − no alignment	+3.79	[3.26, 4.29]	< 0.001	15/15
CIFAR	Quantile − diagonal	+0.023	[0.002, 0.043]	0.070	13/19
ImageNet	Quantile − diagonal	+0.013	[0.003, 0.026]	0.070	9/15
CIFAR	Bures − diagonal	+1.02	[0.75, 1.32]	< 0.001	18/19
ImageNet	Bures − diagonal	+1.42	[1.10, 1.76]	< 0.001	15/15
CIFAR	Paired rank-64 ridge − shuffled	+59.10	[55.23, 63.02]	< 0.001	19/19
ImageNet	Paired rank-64 ridge − shuffled	+31.55	[28.25, 35.17]	< 0.001	15/15
CIFAR	Paired rank-64 ridge − unpaired Bures	−5.98	[−7.28, −4.63]	< 0.001	1/19
ImageNet	Paired rank-64 ridge − unpaired Bures	−27.71	[−29.46, −26.04]	< 0.001	0/15

C.2 Direct numerical check of coordinate invariance

Table 7: Dataset-macro dimensionless distribution residuals at the primary evaluation condition. Lower is better. Diagonal alignment reduces the coordinate-variance residual but can increase the full-covariance residual; Bures substantially reduces the latter.

Dataset	Coordinate variance: no align. → diagonal	Covariance: no align. → diagonal	Covariance: Bures
CIFAR-100-C	0.242 → 0.037	0.653 → 0.714	0.296
ImageNet-C	0.265 → 0.037	0.754 → 0.855	0.405

For diagonal alignment, the largest standardized-vector MSE over all primary summaries is 4.45×10^{-30}, the largest standardized correlation change is 3.55×10^{-16}, and coordinate-rank MSE is exactly zero. These are direct numerical checks of the algebraic identity in Section A.2, not claims about population-moment estimation.

C.3 Severity and calibration size

Table 8 reports severity breadth, while Tables 9 and 10 isolate calibration-size and covariance-rank effects under their separately seeded descriptive suites.

Table 8: Bures-minus-no-alignment agreement gain (pp) across severities. Each entry averages all corruptions in the dataset and split seeds 40–42 at the fixed fitting size $n = 4{,}096$. These are descriptive summaries from the severity-breadth suite.

Dataset	Encoder	Severity 1	Severity 3	Severity 5
CIFAR-100-C	CLIP	+1.62	+3.19	+5.67
	DINOv3	+0.39	+1.57	+3.39
	ViT-B/16	+1.36	+3.14	+6.10
ImageNet-C	CLIP	+3.51	+5.89	+7.74
	DINOv3	+0.55	+1.72	+3.66
	ViT-B/16	+1.67	+2.46	+4.07

Bures gives the highest agreement among evaluated methods in 16 of 18 cells. At severity 1, diagonal is narrowly highest for CIFAR-100-C/DINOv3 and quantile is narrowly highest for ImageNet-C/DINOv3.

Table 9: Bures-minus-diagonal agreement (pp) in the finite-sample suite. Each entry averages five corruptions, severities 1, 3, and 5, and split seeds 10–14 at the displayed fitting size. Negative entries favor diagonal alignment.

Dataset	Encoder	$n = 64$	$n = 256$	$n = 1{,}024$	$n = 4{,}096$
CIFAR-100-C	CLIP	-10.51	-6.04	-1.65	+0.48
	DINOv3	-4.88	-4.16	-1.59	+0.27
	ViT-B/16	-11.21	-6.89	-1.99	+0.72
ImageNet-C	CLIP	-7.53	-4.63	-1.26	+0.32
	DINOv3	-0.56	-0.82	-0.57	+0.35
	ViT-B/16	-3.81	-3.06	-0.66	+1.04

Table 10: Unweighted six-cell macro agreement (%) in the severity-5 rank suite. Rank-4 and full columns are Bures variants.

n	Diagonal	Rank-4 OAS Bures	Full OAS Bures	Full empirical Bures
64	59.12	56.76	54.10	5.49
256	59.50	58.62	56.61	10.13
1,024	59.58	59.22	59.22	56.38
4,096	59.61	59.25	60.79	60.71

The separately evaluated rank-64 OAS Bures setting attains 58.89 at $n = 1{,}024$ and 59.41 at $n = 4{,}096$.

C.4 Pair assignments and estimator choice

Table 11: **Full-output paired-ridge comparison.** Method entries are agreement (%). Here $\pm$ is the SD across three split-level corruption averages. Bures uses unpaired samples, while ridge methods use true pairs. The shuffled control uses the same estimator with permuted assignments. Differences are computed before rounding.

Dataset	Encoder	Unpaired Bures	Paired rank-64 ridge	Paired full ridge	Shuffled full ridge	Full ridge − Bures (pp)
CIFAR-100-C	CLIP	61.44 ± 0.20	57.49 ± 0.29	64.34 ± 0.22	1.03 ± 0.03	+2.89
	DINOv3	74.30 ± 0.25	66.77 ± 0.25	76.36 ± 0.16	0.99 ± 0.02	+2.05
	ViT-B/16	62.51 ± 0.05	56.04 ± 0.30	66.38 ± 0.08	1.02 ± 0.05	+3.87
ImageNet-C	CLIP	48.52 ± 0.07	26.37 ± 0.11	46.36 ± 0.08	0.11 ± 0.01	-2.15
	DINOv3	73.62 ± 0.05	40.97 ± 0.26	73.35 ± 0.06	0.10 ± 0.00	-0.27
	ViT-B/16	55.94 ± 0.07	27.63 ± 0.12	55.26 ± 0.14	0.11 ± 0.01	-0.68

Table 12: **Post-hoc paired-ridge comparisons.** Effects are agreement differences in percentage points. Intervals are pointwise 95% corruption-cluster bootstrap CIs. The p-values use exhaustive sign-flip tests with Holm correction across these four post-hoc comparisons. Encoders and splits are repeated measurements within corruption.

Dataset	Contrast	Effect	95% bootstrap CI	Holm p	Corruption clusters
CIFAR-100-C	Full ridge − Bures	+2.94	$[2.21, 3.72]$	1.53×10^{-5}	19
	Full ridge − shuffled	+68.01	$[63.56, 72.50]$	1.53×10^{-5}	19
ImageNet-C	Full ridge − Bures	-1.03	$[-1.33, -0.74]$	1.22×10^{-4}	15
	Full ridge − shuffled	+58.22	$[53.08, 63.87]$	1.22×10^{-4}	15

The post-hoc full-output follow-up covers all 19 CIFAR-100-C and 15 ImageNet-C corruptions, three encoders, and split seeds 35–37. It adds 612 result records across 102 experimental conditions at 4,096 pairs. The relative ridge factor is fixed at 0.01, using the scaling rule in Section B.2. No explicit rank truncation is used. The reduced-rank rows in Table 11 are verified comparison results from the earlier study and do not enter the four-test family.

The paired-versus-shuffled effects show that correspondence is useful under the same estimator in both datasets. The paired-versus-Bures reversal shows that information availability does not determine estimator ranking: full ridge is better on CIFAR-100-C and worse on ImageNet-C under this fixed relative ridge factor and pair budget.

C.5 Strict predict-before-update streaming

The streaming evaluation uses five severity-5 corruptions per dataset, 2,048 ordered target examples, blocks of 256, a separate 2,048-example reference calibration set, and three order seeds. Each block is predicted before its features update cumulative or fixed-window statistics. It covers CLIP and DINOv3 only. Table 13 reports the resulting task endpoints.

Table 13: Strict streaming accuracy and agreement (%). "Best tested" is selected descriptively within each cell and must not be interpreted as a confirmatory deployment policy. For CIFAR-100-C/DINOv3, the diagonal 256- and 1,024-example windows tie in accuracy; the displayed 1,024-example window has higher agreement.

Dataset	Encoder	No align.	Best tested strategy	Accuracy	Gain (pp)	Agreement
CIFAR-100-C	CLIP	49.60	Diagonal, window 1,024	54.62	+5.02	55.62
	DINOv3	65.17	Diagonal, window 1,024	67.75	+2.58	68.87
ImageNet-C	CLIP	37.09	Diagonal, window 1,024	43.72	+6.63	46.04
	DINOv3	66.04	Cumulative Bures	68.30	+2.27	73.70

Cumulative diagonal exceeds cumulative Bures in agreement in three of four cells. With a 64-example window, Bures trails diagonal by 0.76–3.41 pp. Across the 52 dataset–encoder–strategy macro summaries, the largest predict-before-update relative-Frobenius covariance residual, averaged over five corruptions, is 3.51 (SD 0.03 across the three order-seed means). It is attained by the 64-example Bures window on CIFAR-100-C/DINOv3. Diagonal alignment gives 0.93 in the same cell. These observations support diagonal alignment for short histories, not a universal online ranking.

Calibration diagnostic. Bures is strongest in primary agreement and accuracy but not in every probabilistic metric. Mean alignment versus Bures yields NLL 1.832 versus 1.998 and ECE 16.64% versus 18.60% on CIFAR-100-C. The ImageNet-C values are 2.282 versus 2.418 and 8.96% versus 14.83%. Alignment quality and probability calibration are therefore distinct.

D CONTROLLED IDENTIFIABILITY EXPERIMENTS

The exact Gaussian rotation simulation uses 20 seeds and 250,000 samples per seed at angles $0, 30, \ldots, 180$ degrees. Its aggregate Monte Carlo error from $1 - \theta/(2\pi)$ is below 0.01 pp. Cyclic group experiments give the exact values 50%, 25%, and 12.5% for orbit sizes 2, 4, and 8.

The approximate two-world experiment uses calibration sizes $0, 1, 4, 16, 64, 256, 1{,}024$, 20 seeds, and 50,000 trials per seed. For full task disagreement and calibration probabilities $0.5 \pm \epsilon$ with $\epsilon = 0.01$, the exact equal-prior Bayes value rises from 50.00% at $n = 0$ to 73.89% at $n = 1{,}024$ as the total variation distance between the complete observation laws rises from 0 to 0.478. This experiment evaluates (21). It does not replace the complete observation law with a single-draw distance.

In the controlled four-world PCA-orbit stress test on real embeddings, the generic frozen classifiers are not label-equivariant, so the transitive $1/K$ theorem does not apply. These are constructed endpoint-identical worlds, not ceilings for natural corruptions. The gap between the achievable randomized group-map minimax value and the unrestricted equal-prior Bayes upper bound instead measures restriction to the tested correction family. In a separate paired-rank phase diagram, a controlled experiment using the classical full Procrustes solution (Schönemann, 1966) reaches 100% at $n = d + 1$ for $d \in \{8, 32, 128\}$ and $K \in \{2, 10, 100\}$. Ten-percent assignment noise removes this clean transition.

E ADDITIONAL AMBIGUITY AND RECOVERY RESULTS

This appendix reports the held-out orientation selector, natural-shift and endpoint-structure profiles, the nested profile-batch-size audit, and paired recovery-bound validations. These analyses support Sections 4.1 and 4.2. Complete run inventories and verification records are in the anonymous repository.

E.1 HELD-OUT ORIENTATION SELECTOR

The selected panel crosses CIFAR-100-C and ImageNet-C, CLIP and DINOv3, five severity-5 corruptions, and split seeds 101–103. The estimator uses unpaired features and the frozen readout but neither target labels nor clean–target assignments. Its held-out selection rule chooses an orientation rank or returns the Bures map.

All 60 runs select rank zero and identity shrinkage, for a fallback rate of one. The selector minus Bures is therefore exactly zero for every reported task, feature, logit, likelihood, and calibration metric. The selection rule is a conservative heuristic rather than a significance test. The null result establishes neither equivalence nor the absence of higher-order information. The separate calibrate-then-freeze TTA numbers are omitted because those implementations do not instantiate the published online protocols, and parameter-updating TTA is outside the frozen correction class studied here.

E.2 OPTIMIZATION-CERTIFIED NATURAL-SHIFT PROFILES

Initial study and subsequent evaluation. The initial ImageNet-Sketch outcomes were inspected before the additive panel was fixed. The ImageNet-V2 and ImageNet-R outcomes had not been inspected. The additive panel fixed the fitting set at 3,840 examples to accommodate ImageNet-R's 200-class reference intersection, prespecified three geometries, and required the bound gap to meet a prespecified numerical tolerance.

Additive panel. The 162 units cross ImageNet-R, ImageNet-Sketch, and ImageNet-V2 (Hendrycks et al., 2021; Wang et al., 2019; Recht et al., 2019); CLIP and DINOv3; three costs after Bures alignment (Mahalanobis using the OAS estimate of the full reference covariance, standardized squared Euclidean using its diagonal, and uncentered normalized cosine); seeds 201–203; and 1%/5%/10% relative budgets. Each unit uses 3,840 unpaired fitting examples, 1,024 disjoint profile rows, and a deterministic 64-row exact-MILP diagnostic. The target/reference/class inventories are respectively 30,000/5,000/200 for ImageNet-R, 50,000/25,000/1,000 for ImageNet-Sketch, and 10,000/25,000/1,000 for ImageNet-V2. Target labels, row-level WNIDs, semantic IDs, and pairing are unavailable to optimization. Frozen-readout predictions define the assignment reward. Anonymous identifiers index atoms and resolve only exact feature ties. The alternative is the best feasible greedy or Lagrangian assignment. Thus both serialized assignments preserve the exact endpoint multisets, $L_u \le \Gamma_u \le U_u$, and the selected pair has exact value $1 - L_u/2$ when both indexed inverses are admissible.

Table 14: **Ambiguity-profile numerical tightness.** Width $U - L$ in percentage points over the three ambiguity-profile certificate families below. The paired recovery-error certificate families are excluded. Quantiles summarize optimization-bound widths, not statistical uncertainty.

Certificate family	Certificates	Median	95th percentile	Maximum
Three-geometry panel	162	0.044	0.362	0.597
Common-cost follow-up	720	0.071	0.363	0.952
Profile-batch-size sweep	288	0.540	0.958	0.993
All ambiguity-profile certificates	**1,170**	**0.095**	**0.781**	**0.993**

Under the earlier minimum-relative convention, the same nominal budget spent 3.8–3.9 times more of the mean random gap on ImageNet-Sketch than on the paired ImageNet-C control. Equal nominal percentages were therefore not comparable cost severities across shifts. The common-cost follow-up in Section E.3 resolves this limitation directly.

The maximum numerical gap over the full panel is 0.596915 percentage points (mean 0.085682), far below the prespecified five-point threshold. At 5% cost, even the smallest of the 18 descriptive conditional lower bounds is 73.02%. Observed WNID clusters after pooling are 200 for ImageNet-R, 938–939 for ImageNet-Sketch, and 965–967 for ImageNet-V2. Four disjoint verification runs collectively reconstruct every unit's assignments, bijection checks, feature costs, frozen-readout rewards, exact diagnostics, and WNID statistics. These checks verify the computations for the declared finite instances. They do not turn a constructed correspondence into a claim about the realized natural shift.

E.3 Common-cost and endpoint-structure follow-up

Timing, protocol, and information boundary. The grid for this follow-up was specified after the earlier marginal-alignment results, the initial ImageNet-Sketch experiment in Section E.2, the cost-scale audit, and the paired-ridge results had been examined, but before this follow-up was evaluated. The analysis is post hoc and is not part of the original confirmatory family. Every group uses CLIP or DINOv3, split seed 201–203, 3,840 fitting examples, and 1,024 profile examples. Costs are squared Mahalanobis costs after Bures alignment, using the full reference covariance estimated with OAS shrinkage. For exact $C_{\mathrm{rand}} = \mathbb{E}_{\pi\sim\mathrm{Unif}(S_n)}C(\pi)$, the common budget is $C_{\min} + q(C_{\mathrm{rand}} - C_{\min})$ at $q \in \{.01, .05, .10, .20, .40\}$. The value $q = .05$ was designated primary. All five budget levels are run for the three natural shifts and six focus ImageNet-C corruptions. The other nine ImageNet-C corruptions are run only at $q = .05$. The six focus corruptions additionally use endpoint sets constructed from administrative clean–corrupt pairs, but those identities are withheld from map fitting, costs, rewards, and assignment optimization and are revealed only after the assignments are frozen. Target-label files are verified separately. Their labels are used to compute accuracy only after the assignments are fixed.

The complete inventory is 18 natural, 90 disjoint-unpaired ImageNet-C, and 36 paired-endpoint geometry groups. The five normalized common-cost budget levels contribute 504 certificate rows; three legacy direct-budget levels on both six-corruption arms contribute 216 more, for 720. The realized $q = .05$ profile contains the 18 natural and 90 disjoint groups only.

Table 15: **Complete common-cost profile by encoder.** Each cell is the numerical bracket $[\mathrm{mean}\, L, \mathrm{mean}\, U]$, expressed in percent, over three split seeds and, for ImageNet-C, the specified corruption set. Numbers in parentheses give the number of geometry groups.

Panel	Encoder (groups)	$q = .01$	$q = .05$	$q = .10$	$q = .20$	$q = .40$
ImageNet-R	CLIP (3)	46.19–46.35	73.50–73.62	87.57–87.74	99.93–99.95	100–100
	DINOv3 (3)	43.95–44.21	67.48–67.54	79.69–79.74	92.51–92.61	100–100
ImageNet-Sketch	CLIP (3)	53.87–54.10	83.04–83.12	94.92–94.97	100–100	100–100
	DINOv3 (3)	47.79–48.40	74.15–74.34	85.58–85.74	95.90–95.99	100–100
ImageNet-V2	CLIP (3)	46.00–46.32	74.67–75.04	89.00–89.16	100–100	100–100
	DINOv3 (3)	42.68–42.96	68.39–68.47	80.86–80.93	92.71–92.81	100–100
ImageNet-C focus, disjoint	CLIP (18)	49.82–50.06	79.33–79.47	92.49–92.59	99.97–99.97	100–100
	DINOv3 (18)	44.85–45.06	71.10–71.24	83.38–83.51	94.78–94.83	100–100
ImageNet-C focus, paired	CLIP (18)	21.27–21.39	39.02–39.14	52.03–52.13	69.72–69.81	90.99–91.04
	DINOv3 (18)	6.15–6.25	17.01–17.15	27.18–27.30	43.65–43.76	69.53–69.63

The all-15 ImageNet-C panel was evaluated only at $q = .05$. Its mean bracket is 80.34–80.46 for CLIP and 71.38–71.52 for DINOv3, with 45 geometry groups per encoder.

The complete natural-shift budget curves are shown in Figure 4.

Endpoint structure. At the primary $q = .05$, every paired endpoint bracket lies strictly below its matched disjoint-unpaired bracket. Consequently, the two exact maxima are ordered in all 36 cells under the numerical solver model. This conclusion does not rest only on comparing two feasible lower bounds. The selected feasible-disagreement contrast remains the prespecified inferential estimand in Table 16. The withheld true pairing itself is feasible at this budget in all 36 paired groups: its excess cost averages 0.67% of the random gap and never exceeds 4.71%.

Using an exhaustive two-sided sign-flip test and a fixed-panel cluster bootstrap (Lehmann & Romano, 2005; Davison & Hinkley, 1997), the six corruption-level means are all negative, the test gives $p =$

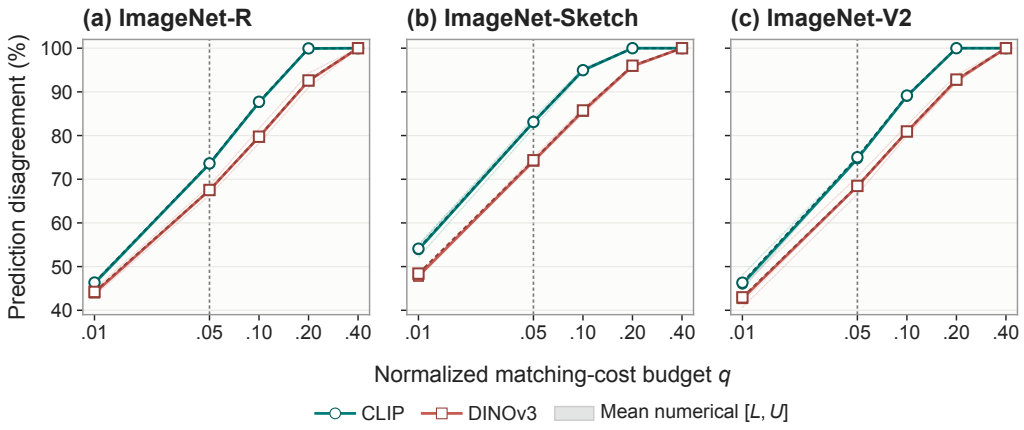


Figure 4: **Natural-shift correspondence-ambiguity profiles.** The normalized budget q is the fraction of the random-minus-minimum assignment-cost gap, not the fraction of reassigned examples. The horizontal axis is logarithmic. The dashed line marks $q = 0.05$. Faint curves show individual data splits, and bands show mean numerical $[L, U]$ bounds, not confidence intervals.

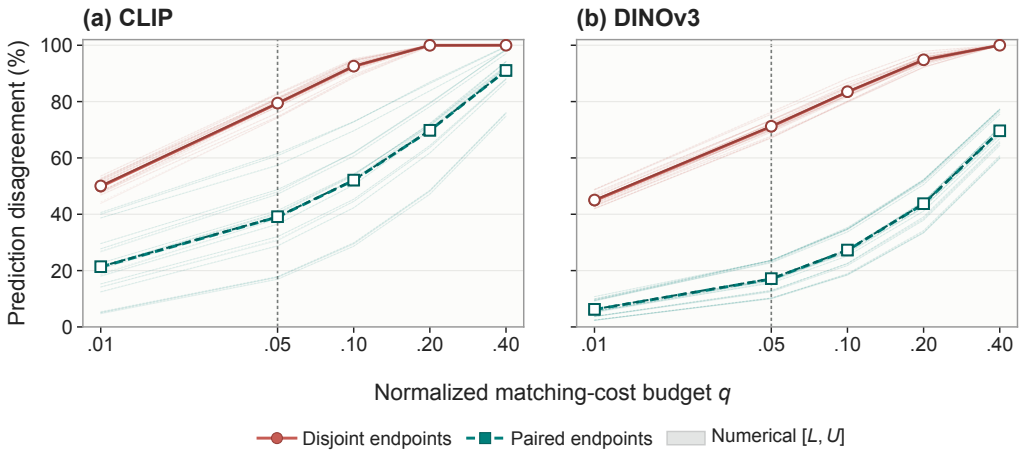


Figure 5: **Endpoint structure changes the ambiguity profile.** On six ImageNet-C corruptions, disjoint reference and target sets are compared with sets containing the targets' clean counterparts. Pairing is hidden from fitting, cost construction, and optimization. Because the endpoint sets differ, this is not the effect of revealing their correspondence. Faint curves show corruption-by-split profiles; bands show mean numerical $[L, U]$ bounds, not confidence intervals. The dashed line marks $q = 0.05$ on the logarithmic budget axis.

.03125 (the minimum possible with six clusters), and the 10,000-draw interval is $[-52.30, -41.75]$ points. Averaging the 36 pairwise numerical difference intervals gives $[-47.34, -47.07]$ points. All 288 contrasts across the five normalized common-cost and three direct-budget levels are negative and interval-separated, although only $q = .05$ is the designated statistical contrast. The comparison is a controlled endpoint-structure diagnostic, not a causal pairing effect, because the two arms use different endpoint sets.

ImageNet-Sketch also strictly exceeds the paired ImageNet-C control in all 180 matched encoder–seed–corruption comparisons over the common grid. At $q = .05$, the average separation interval is $[50.45, 50.71]$ points and the smallest lower-bound separation is 20.91. Across all five budget values, the smallest lower-bound separation is 0.36. The contraction therefore establishes non-vacuity under

Table 16: **Primary endpoint-structure contrast at $q = .05$.** Disjoint and paired columns give mean numerical profile brackets over both encoders and three seeds. Δ_{sel} is paired minus disjoint selected feasible disagreement. Profile brackets are numerical bounds, not confidence intervals. Corruptions are the fixed focus panel.

Corruption	Disjoint $[L, U]$ (%)	Paired $[L, U]$ (%)	Δ_{sel} CLIP (pp)	Δ_{sel} DINOv3 (pp)	Pooled Δ_{sel} (pp)
Brightness	71.11–71.24	13.77–13.89	−57.52	−57.16	−57.34
Defocus blur	77.56–77.68	35.64–35.75	−33.20	−50.62	−41.91
Fog	74.74–74.91	23.37–23.48	−48.31	−54.43	−51.37
Gaussian noise	77.83–77.94	41.59–41.69	−22.20	−50.29	−36.25
JPEG compression	74.30–74.44	26.43–26.54	−38.83	−56.90	−47.87
Motion blur	75.78–75.93	27.29–27.54	−41.80	−55.18	−48.49
Macro	**75.22–75.36**	**28.02–28.15**	−40.31	−54.10	−47.20

a different endpoint construction, not a natural-versus-corruption distinction: disjoint ImageNet-C occupies the same high-ambiguity band as the natural-shift panels (Figure 5).

Realized accuracy and the profile. After the $q = .05$ assignments were frozen, the same 108 unpaired groups were scored against external target labels. Bures improves over no alignment in all 90 ImageNet-C groups, averaging 4.83 points. It improves 15/18 natural groups and averages 3.40 points. Bures slightly reduces mean accuracy in the ImageNet-V2/CLIP row. Bures is uniquely best among the four evaluated methods in 67/90 ImageNet-C groups and best including ties in 70/90.

Across the fixed 15 corruption and three natural-shift clusters, Spearman correlation between profile disagreement and Bures gain is .278. Using a family-stratified cluster bootstrap and a within-family permutation procedure (Davison & Hinkley, 1997; Lehmann & Romano, 2005), the 10,000-draw interval spans $[-.254, .768]$ and the 10,000-draw test gives $p = .365$; the 18 leave-one-cluster-out correlations range from .184 to .517. Encoders and overlapping split seeds are nested within clusters. This analysis finds no statistically supported association and does not calibrate a deployment rule: the profile is an ambiguity diagnostic, not a method-benefit predictor.

Numerical and reconstruction checks. All 144 geometry groups and 720 certificate rows are present. The mean 1,024-row primal–dual width is 0.103 points and the maximum is 0.952. Every width is below the prespecified five-point tolerance. All 720 deterministic 64-row MILPs report a proven optimum within the corresponding bracket, with maximum reported MIP gap 3.45×10^{-14}. Eight disjoint verification runs reconstruct all group identities, budgets, assignments, costs, predictions, scores computed after assignments were fixed, and aggregate rows. These are computational checks under the recorded floating-point solver model, not formal real-arithmetic proofs or population uncertainty intervals.

E.4 Nested profile-batch-size and class-structure audit

Timing, design, and information boundary. This post-hoc audit was designed after the earlier results were examined. Its complete protocol was fixed before its own scaling outcomes were inspected, and it provides a fixed-panel description, not population inference. The audit fixes squared Mahalanobis costs after Bures alignment, full-reference OAS covariance estimation, $q = .05$, 3,840 fitting examples, and the readout. It varies the encoder (CLIP or DINOv3), split seed (201–203), and dataset (ImageNet-Sketch or one of 15 severity-5 ImageNet-C corruptions) under the disjoint-unpaired construction. Within each of the 96 condition–encoder–seed groups, one 8,192-row endpoint selection was frozen first. The 4,096- and 1,024-row profile batches are literal prefixes after the same calibration segment. The largest power of two below the construction's 8,660-row disjoint-identity limit is 8,192.

The audit creates 288 fresh certificates. It does not splice the earlier 1,024-row outcomes into the sweep because those endpoints are not prefixes of the newly selected 8,192-row batches. Ground-truth class values are excluded from selection, map fitting, geometry, cost, reward, and assignment optimization. They are interpreted only after all three certificates are frozen to describe class structure. For sizes $n_a < n_b$, the numerical difference bracket is $[L(n_b) - U(n_a), U(n_b) - L(n_a)]$. A decrease is certified only when its upper endpoint is below zero, an increase only when its lower endpoint

is above zero, and otherwise the direction is inconclusive. The joint evolution of ambiguity and same-class candidate availability is summarized in Figure 6.

Table 17: **Ambiguity and class structure across nested profile-batch sizes.** Entries report the mean numerical bracket, the minimum feasible lower bound across cells, same-class candidate availability, and maximum class-preserving matching without a cost constraint. There are 45 cells per ImageNet-C encoder–size row and three per ImageNet-Sketch row. Class summaries are computed after optimization.

Panel	Encoder	n	Mean $[L, U]$ (%)	Minimum L (%)	Same-class candidate (%)	Maximum class-preserving match (%)
ImageNet-C, disjoint	CLIP	1,024	80.28–80.75	73.44	63.02	47.98
		4,096	81.56–82.07	73.66	98.48	71.92
		8,192	81.95–82.49	72.94	99.94	80.39
	DINOv3	1,024	70.24–70.81	65.14	63.02	47.98
		4,096	68.04–68.57	63.45	98.48	71.92
		8,192	66.42–66.94	61.00	99.94	80.39
ImageNet-Sketch	CLIP	1,024	82.71–83.12	81.45	66.44	48.05
		4,096	83.55–83.98	83.47	99.06	73.84
		8,192	84.09–84.47	83.62	100.00	82.97
	DINOv3	1,024	73.86–74.35	73.54	64.00	48.14
		4,096	72.83–73.10	72.14	98.99	73.42
		8,192	71.67–72.27	71.02	100.00	82.05

Table 18: **Within-cell scaling directions.** Changes and mean numerical difference brackets are percentage points for the right size minus the left size. Counts distinguish certified decreases, certified increases, and numerically unresolved cases. Split seeds and corruptions are nested fixed cells, not independent draws from a population.

Panel	Encoder	Sizes	Mean ΔL (pp)	Mean numerical difference bracket (pp)	Decreases	Increases	Unresolved
ImageNet-C	CLIP	1,024→4,096	$+1.28$	$[+0.81, +1.78]$	5	33	7
		4,096→8,192	$+0.39$	$[-0.11, +0.93]$	2	20	23
		1,024→8,192	$+1.67$	$[+1.20, +2.20]$	2	34	9
	DINOv3	1,024→4,096	-2.20	$[-2.77, -1.67]$	39	0	6
		4,096→8,192	-1.62	$[-2.15, -1.10]$	40	0	5
		1,024→8,192	-3.82	$[-4.39, -3.29]$	45	0	0
ImageNet-Sketch	CLIP	1,024→4,096	$+0.84$	$[+0.43, +1.26]$	0	2	1
		4,096→8,192	$+0.53$	$[+0.11, +0.92]$	0	1	2
		1,024→8,192	$+1.37$	$[+0.97, +1.76]$	0	3	0
	DINOv3	1,024→4,096	-1.03	$[-1.52, -0.76]$	2	0	1
		4,096→8,192	-1.16	$[-1.43, -0.56]$	2	1	0
		1,024→8,192	-2.19	$[-2.67, -1.59]$	3	0	0

Result. Same-class candidate availability becomes nearly complete without removing the ambiguity. At $n = 8{,}192$, every ImageNet-C cell contains all 1,000 target classes, 99.94% of target rows have a same-class reference candidate on average, and the mean maximum class-preserving matching fraction is 80.39%. The corresponding mean feasible lower bound remains 81.95% for CLIP and 66.42% for DINOv3. ImageNet-Sketch has 100% candidate coverage and 82.05–82.97% mean class-matching capacity, yet its mean lower bounds remain 84.09% and 71.67%. All 96 individual $8{,}192$-row lower bounds exceed 60.99%.

The direction varies by encoder in this fixed panel; there is no universal decay with rows per class. Over the long $1{,}024 \rightarrow 8{,}192$ comparison, CLIP increases in 37/48 cells, decreases in 2, and is inconclusive in 9. DINOv3 decreases in 48/48. ImageNet-C brightness is the only corruption whose CLIP mean falls, and one ImageNet-Sketch DINOv3 seed increases from $4{,}096$ to $8{,}192$, so the fixed-panel results do not establish monotonicity in every adjacent cell. Sparse class support is therefore not a sufficient deterministic explanation for these fixed-panel profiles, but the audit does not identify a population scaling law or isolate all size-dependent properties of the assignment geometry.

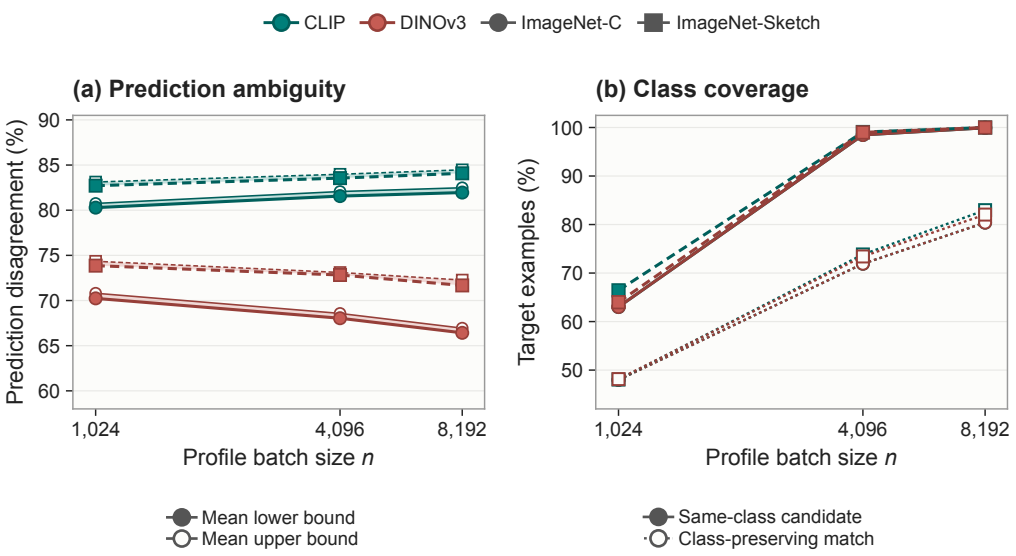


Figure 6: **Near-complete same-class candidate availability leaves ambiguity unresolved.** At $q = 0.05$, (a) shows mean lower (filled) and upper (open) ambiguity bounds, with shading between them. (b) compares same-class candidate availability with the maximum fraction matchable to same-class references, without a cost constraint. Class labels are used only after optimization. Smaller batches are prefixes of larger batches. The size axis is logarithmic. Numerical bounds are not confidence intervals.

Numerical and recovery checks. All 96 groups, 288 fresh certificate rows, and 288 pairwise-difference records are present. Every computational width is at most 0.993 percentage points. Each 64-row exact diagnostic reports a proven optimum inside its numerical bracket, and two disjoint verification runs reconstruct every frozen-input assignment and diagnostic.

E.5 Paired method-specific recovery-bound validation

Design and information boundary. An initial 40-batch study motivated this follow-up, whose design was fixed before its own outcomes were computed. The study crosses severity-5 Gaussian noise and motion blur with three partitions for CLIP and one matched partition for ViT-B/16. Repetitions reuse identities and assess stability rather than independent-dataset replication. Within each of the resulting eight cells, a 4,096-pair fitting split fits the diagonal and Bures maps (identity is fixed), 29 paired batches calibrate a 90% split-conformal threshold (Shafer & Vovk, 2008; Johnstone & Cox, 2021), and 20 held-out batches of 256 identities are evaluated.

Identity, diagonal, and Bures share a feasible set defined by squared Mahalanobis costs after Bures alignment, using the full reference covariance estimated with OAS shrinkage. A direct absolute-error comparator instead standardizes paired method errors using 16 batches from the fitting split and calibrates the maximum standardized error on the same 29 calibration batches. It therefore uses true pairs from the fitting split after the correction maps are frozen, whereas the geometric maps are fitted without pairing. A second control independently permutes the columns of each cost matrix, preserves its minimum and random expected assignment costs, and receives its own calibration threshold. Neither reference predictions nor true pairs are permuted. Evaluation pairs are withheld from all bound construction and revealed only for post-bound scoring. After these outcomes were reviewed, direct signed comparators were added for the same three ordered pairs. For paired batch b, define

$$D_{ac,b} = E_{a,b}(\pi_b^\star) - E_{c,b}(\pi_b^\star). \tag{34}$$

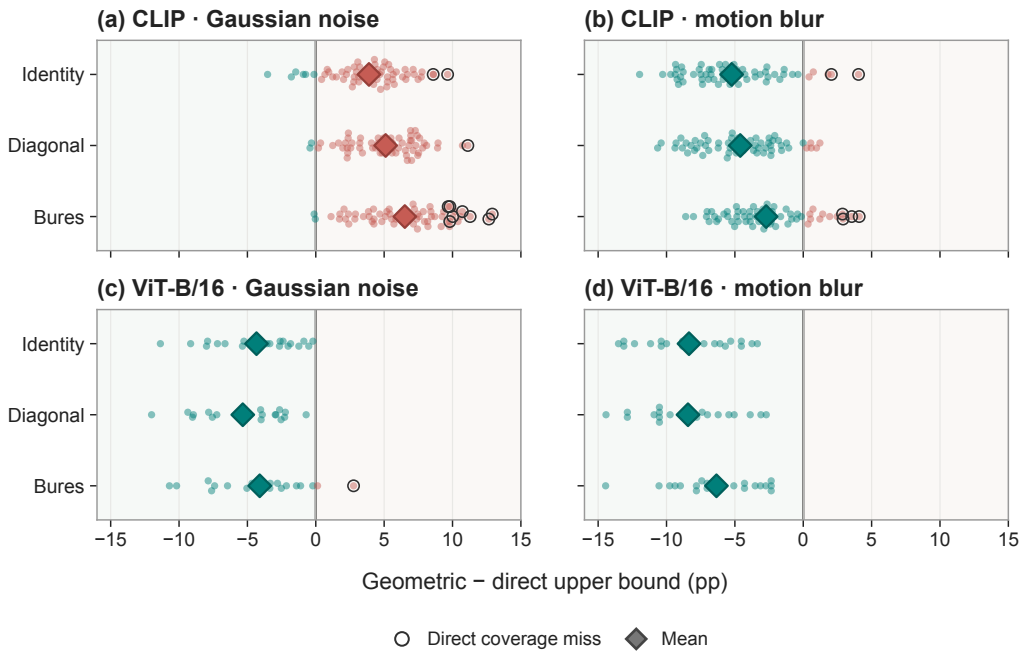


Figure 7: **Absolute-bound tightness depends on the condition.** Points show Bures-generated geometric upper bounds minus direct absolute-error calibration bounds for the same batch and method. Negative values indicate a tighter geometric bound. Positive values indicate a tighter direct bound. Diamonds show condition–method means. Open rings mark direct bounds below the observed recovery error. All batches are shown. Vertical offsets separate overlapping points.

The center μ_{ac} is the mean of $D_{ac,b}$ over 16 batches from the fitting split. The three scale rules are

$$s_{ac}^{\text{unscaled}} = 1, \tag{35}$$

$$s_{ac}^{\text{diff}} = \max\{\text{sd}_{\text{fit}}(D_{ac}), 1/256\}, \tag{36}$$

$$s_{ac}^{\text{sum}} = \max\{\text{sd}_{\text{fit}}(E_a), 1/256\} + \max\{\text{sd}_{\text{fit}}(E_c), 1/256\}. \tag{37}$$

The final rule uses sum-of-SDs normalization, not a variance identity. For any one of these fixed scales, the 29 calibration batches use

$$Z_b = \max_{(a,c)\in\mathcal{C}} \frac{D_{ac,b} - \mu_{ac}}{s_{ac}}, \tag{38}$$

$$q_{\text{dir}} = Z_{(k)}, \tag{39}$$

$$U_{ac}^{\text{dir}} = \text{clip}_{[-1,1]}\{\mu_{ac} + s_{ac} q_{\text{dir}}\}, \tag{40}$$

with $k = \lceil (m+1)(1-\alpha) \rceil$. The maximum over contrasts gives one simultaneous calibration event. These direct bounds are constant within a cell and contrast, whereas the geometric bound also uses the current evaluation batch's endpoint geometry.

Result. The primary Bures-generated feasible set contains the withheld true correspondence in 111/120 CLIP batches and 38/40 ViT-B/16 batches. Its absolute method bounds simultaneously cover all three observed method errors in 160/160 batches. The direct absolute-error calibration does so in 105/120 and 39/40, respectively. Every one of the 480 primary signed upper bounds also covers its realized contrast, jointly within all 160 batches, and none of the 263 issued strict orderings has the wrong sign. Conditional on observed true-pair feasibility, coverage is 447/447 contrasts with 248 certificates; for the 11 infeasible batches it is descriptively 33/33 with 15 certificates. These counts are distinct: feasibility is sufficient for the coverage event in Proposition 3, not necessary for empirical coverage. The primary geometry is also condition-dependent. For CLIP Gaussian noise,

Table 19: **Absolute recovery-error bounds.** Observed recovery errors and upper bounds are percentages. Optimization gaps and slack are percentage points. Bures-generated U uses the common correspondence set, and slack is U minus observed error. Direct columns use absolute-error calibration. CLIP pools three partitions, while ViT-B/16 pools two corruptions.

Panel	Method	Observed recovery error (%)	Bures U (%)	Gap (pp)	Bures slack (pp)	Direct U (%)	Direct slack (pp)
CLIP pooled (120)	Identity	65.85	71.04	0.265	5.19	71.71	5.86
	Diagonal	59.01	65.34	0.294	6.33	65.09	6.08
	Bures	58.13	64.64	0.293	6.51	62.74	4.61
ViT-B/16 pooled (40)	Identity	51.37	51.58	0.127	0.215	57.93	6.56
	Diagonal	49.66	49.87	0.143	0.211	56.75	7.10
	Bures	47.84	48.06	0.117	0.215	53.28	5.44

Table 20: **Signed common-correspondence certificates.** Observed contrasts are descriptive mean error differences in percentage points. Method columns count issued orderings and do not imply correctness. This table disaggregates the Bures-generated and Sum-of-SDs procedures. The recalibrated column control issues 0/480 orderings (Table 1). † Two of these 40 Sum-of-SDs orderings have the wrong observed sign.

Panel	Contrast	Observed contrast (pp)	Bures-generated issued	Sum-of-SDs issued
CLIP, Gaussian noise	Bures − identity	−6.95	6/60	60/60
	Diagonal − identity	−6.00	3/60	60/60
	Bures − diagonal	−0.95	0/60	0/60
CLIP, motion blur	Bures − identity	−8.50	60/60	60/60
	Diagonal − identity	−7.68	60/60	60/60
	Bures − diagonal	−0.82	24/60	0/60
ViT-B/16, pooled	Bures − identity	−3.53	38/40	40/40†
	Diagonal − identity	−1.71	37/40	0/40
	Bures − diagonal	−1.82	35/40	0/40

direct absolute-error calibration has 4.11–5.16 points of mean slack versus 8.43–10.62 for geometry. For motion blur, geometry has only 1.95–2.40 points versus 5.12–7.18 direct. In the matched ViT-B/16 panel its slack is 0.211–0.215 points versus 5.44–7.10 direct (Figure 7).

The post-hoc sum-of-SDs comparator directly calibrates signed recovery-error differences. It covers 454/480 realized contrasts and all three jointly in 141/160 batches. Two of its 280 negative bounds have the wrong realized sign. Both errors are ViT-B/16 Bures–identity comparisons. The primary geometry covers 480/480 contrasts jointly in 160/160 batches, with zero wrong signs among 263 negative bounds. Mean signed slack is 4.05 points direct versus 3.91 primary. This descriptive comparison favors the geometric set on coverage and correctness, not on the raw number of issued orderings. It was not prospectively specified.

The column-permutation control isolates what makes the geometric certificate useful. It retains separately calibrated coverage in 160/160 batches, yet all 480 absolute upper bounds are exactly 100% and all 480 signed comparisons fail to certify improvement. The primary geometry yields 263/480 signed certificates (Table 20). Recalibration therefore preserves observed coverage, but retaining cost-to-endpoint association provides the information lost under column permutation.

Common-correspondence audit. The methodwise-extrema audit revalidates all 263 primary signed certificates and issues 220 certificates from separate ranges. The 43-count difference is not itself a proof of 43 common-only cases: ten remain unresolved by the available numerical brackets. Common signed upper bounds certify 33 additional orderings on 32 batches, with feasible witnesses establishing that separate error ranges cannot certify them (7 diagonal–identity, 9 Bures–identity, and 17 Bures–diagonal). All 32 batches have positive anchored ambiguity, with a mean numerical bracket of 16.54–16.99%. Of the 33 certificates, 31 occur on the observed true-pair-feasibility event. Across all 160 sets, anchored ambiguity averages 20.71–21.10%, with a positive feasible witness in 140 batches. Figure 8 shows all 480 audited comparisons by condition and contrast.

On the grid-sharpened scale, 246, 228, and 179 of the 263 issued comparisons certify margins of at least 0.5, 1, and 2 percentage points, respectively. The corresponding counts among the 33 proven common-only cases are 21, 12, and 2.

Direct signed-calibration sensitivity. Table 21 reports all nine predeclared family–level combinations. The paired-difference-SD rule at $\alpha = .10$ was primary for this new audit and did not uniformly improve the sum-of-SDs rule. The counts pool repeated batches descriptively. Every family and confidence level has a separate one-future-batch coverage event.

Table 21: **Complete direct signed-calibration sensitivity.** α is the miscoverage level. Joint coverage requires all three contrasts to cover within a batch. Mean slack is the signed upper bound minus the observed contrast in percentage points. Counts pool repeated batches descriptively.

Family	α	Joint coverage (/160)	Covered contrasts (/480)	Issued orderings (/480)	Incorrect orderings	Mean slack (pp)
Centered, unscaled	.05	153	472	160	0	5.43
	.10	143	457	280	2	4.03
	.20	128	440	320	11	2.97
Sum-of-SDs	.05	149	467	140	0	5.49
	.10	141	454	280	2	4.05
	.20	129	440	300	9	3.04
Paired-difference-SD	.05	151	470	180	0	5.20
	.10	143	457	240	8	4.14
	.20	129	437	320	11	3.08

Cost-generator sensitivity. Identity-, diagonal-, and primary Bures-generated costs put the withheld true assignment in their separately calibrated sets for 145, 146, and 149 of 160 batches, respectively. Every geometry nevertheless has observed joint absolute-and-signed coverage in 160/160 batches and zero wrong-sign orderings. The alternatives remain informative, issuing 201 orderings each, but their methodwise absolute slacks are materially larger than the primary geometry's (Table 22).

Table 22: **Cost-generator sensitivity at $\alpha = .10$.** Joint coverage requires all three method errors and all three signed contrasts to cover within a batch. Mean absolute slack is upper bound minus observed error in percentage points. Each generator has a separate calibration event. Pooled counts are descriptive.

Cost generator	True pairing feasible (/160)	Joint coverage (/160)	Issued orderings (/480)	Incorrect orderings	Mean absolute slack (pp)		
					Identity	Diagonal	Bures
Identity	145	160	201	0	7.90	9.55	9.90
Diagonal	146	160	201	0	7.73	9.35	9.69
Bures (primary)	149	160	263	0	3.95	4.80	4.94

The audit contains 2,880 numerical objectives and 144 prespecified 16-row exact MILP checks. Every exact optimum falls inside its bracket, and the maximum alternative-geometry numerical width is 1.7484 points. These are fixed-panel sensitivity results, not a joint guarantee across geometries or a universal optimality claim for Bures costs.

Numerical precision and scope. Pooled-by-encoder method means of primary absolute optimization width are 0.265–0.294 points for CLIP and 0.117–0.143 for ViT-B/16. Corresponding contrast means are 0.249–0.321 and 0.108–0.162 points. The global maximum over the original 1,920 primary-plus-control certificates is 1.182 points. All 96 prespecified 16-row exact MILP optima fall inside their numerical brackets. Optimization uncertainty is therefore much smaller than the condition-dependent upper-bound slack. Before the $1/256$ grid sharpening used for the margin summary, the signed certificate closest to zero has upper bound -0.0245936 points, versus a 0.00000837-point numerical tolerance, a factor of 2,939.

The split-conformal statement in Proposition 3 concerns one future exchangeable batch within one encoder–corruption–partition cell conditional on its separately fitted maps. Geometry and its column-permuted control have separate coverage events. Empirical coverage is not a joint guarantee over

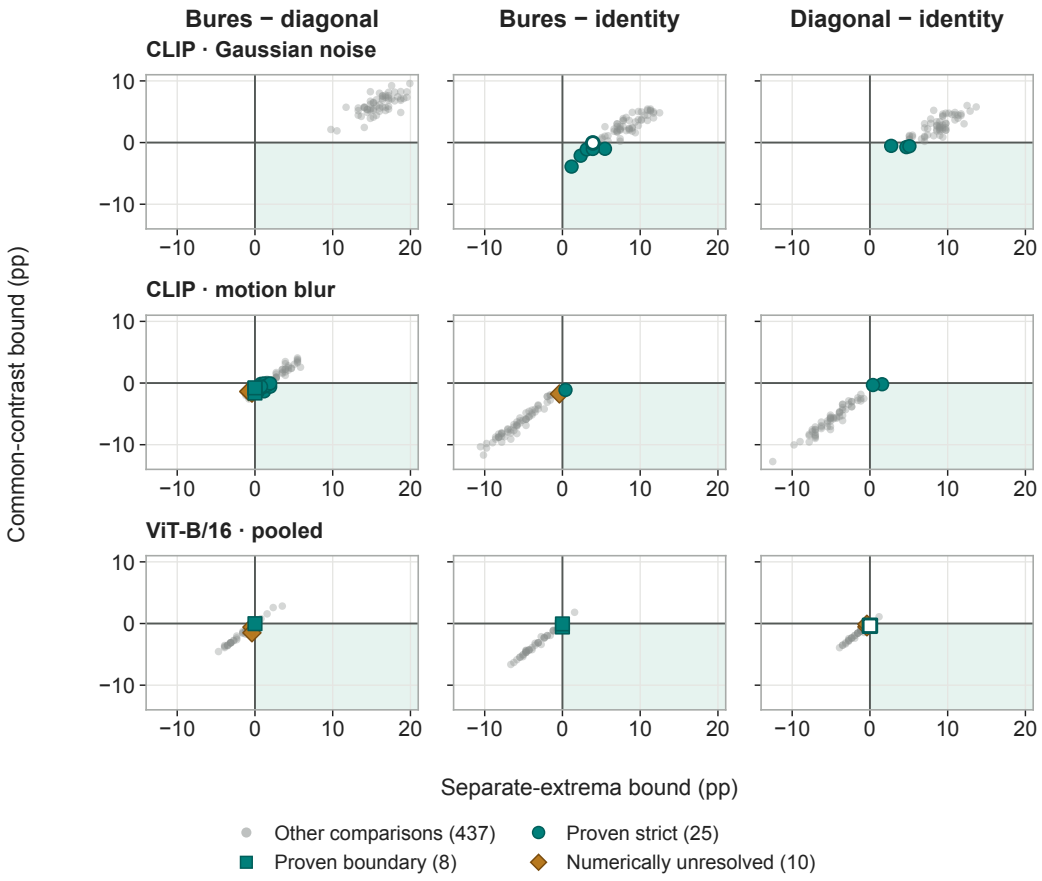


Figure 8: **Common-correspondence comparisons by condition.** All 480 comparisons from Figure 2 are shown once: 60 per CLIP panel and 40 per pooled ViT-B/16 panel. Over the same correspondence set, x is a feasible lower bound on $\max E_a - \min E_c$, and y upper-bounds $\max(E_a - E_c)$. Shading marks $x \geq 0,\ y < 0$. Teal markers indicate 33 proven common-only cases. Hollow markers flag two with infeasible true pairings. Amber diamonds mark ten numerically unresolved cases. Gray points show the remaining comparisons. All panels use Bures-generated costs and common axis limits. Counts are descriptive. These are optimization bounds, not confidence intervals.

batches, partitions, encoders, or corruptions. It also provides neither a population statement over corruptions nor guarantees for external target-label accuracy or unpaired natural shifts. The fixed two-corruption panel and matched single-partition ViT-B/16 check are descriptive.

Verification. The original paired-study verification reconstructed all 160 evaluation units and 1,920 absolute and signed certificates, recalculated eight calibration records, reran 96 exact 16-row MILP diagnostics, and deterministically rederived the aggregate. Every inventory and cross-binding check passed. The exact diagnostics are implementation checks rather than additional scientific evidence. The common-correspondence, direct-calibration, and cost-generator audits were subsequently verified from their separate locked exports as described in Section G.

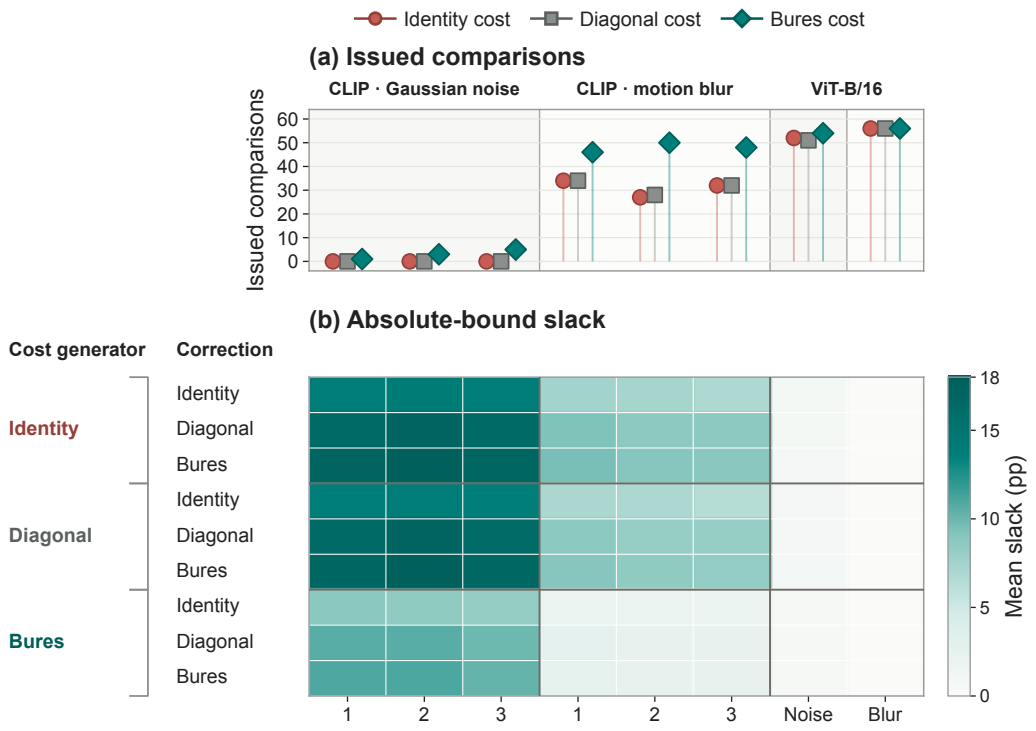


Figure 9: **Cost-generator sensitivity.** At $\alpha = 0.10$, (a) shows issued orderings out of 60 comparisons per condition (20 batches $\times$ three contrasts). CLIP columns show partitions 1–3 within each corruption. ViT-B/16 columns show Gaussian noise and motion blur. (b) shows mean absolute-bound slack, defined as upper bound minus observed recovery error, in percentage points. Lighter cells indicate tighter bounds. Each geometry is calibrated separately. These fixed-panel results provide no joint guarantee across geometries.

F PRACTICAL GUIDANCE FROM THE EVALUATED SETTING

The experiments support the following certification workflow for the evaluated setting.

Identify the observation model. First determine whether the reference set merely samples the clean distribution or contains the clean counterpart of every target example. The former supports only the unpaired ambiguity profile. The paired guarantee requires counterpart containment and independent paired calibration.

Separate fitting, calibration, and evaluation. Fit corrections and the cost geometry without using calibration or evaluation pairs. Use calibration pairs only to set the common correspondence threshold, and keep evaluation pairings hidden until every bound is frozen.

Compute absolute and signed bounds. For each evaluation batch, compute lower and upper optimization bounds over the declared feasible set. A negative signed upper bound certifies an ordering under one common correspondence. If its sign remains unresolved, abstain from ranking the methods.

Audit validity and informativeness separately. Check observed coverage after revealing evaluation pairs, but also report bound slack, issued orderings, optimization gaps, and a cost-association control. Coverage alone does not show that the correspondence geometry is useful.

The workflow applies only to the paper's paired-reference setting and does not cover an arbitrary reference bank, future shifts, or repeated deployment batches.

G INTEGRITY, AMENDMENTS, AND REPRODUCIBILITY

Artifact and verification. The accompanying anonymous repository contains resolved experiment configurations, analysis and verification code, compact frozen result artifacts, and machine-readable provenance manifests. Its documented verification command checks completion status, expected row counts, missing or duplicate conditions, and recorded SHA-256 hashes. Run inventories for the included evidence, documented implementation amendments, checksums, and reconstruction reports are provided with the artifacts.

Original-campaign outcome-blind numerical amendments. Before inspecting outcomes, the implementation replaced the ill-conditioned Bures Gram eigendecomposition with an algebraically equivalent factor SVD, used absolute-eigenvalue ordering for symmetric low-rank corrections, and withdrew estimators that failed their numerical-identification checks. An earlier outcome-blind amendment replaced a nonconvergent clean-classifier fit with deterministic L-BFGS. Datasets, evaluation conditions, split seeds, metrics, and all confirmatory contrasts were unchanged.

Extension selection and numerical validation. The design of the compute-limited experiments was selected using only the declared wall-clock and memory budget, and no failed or incomplete run enters a numerical summary. The additive profile panel was fixed after the initial ImageNet-Sketch outcomes were known but before the ImageNet-V2 and ImageNet-R profile outcomes were examined. The reproducibility log records a validator-tolerance correction for machine-precision HiGHS gaps (Huangfu & Hall, 2018). It neither changed nor recomputed a scientific artifact.

Timing and data reuse in later analyses. Later analyses left the original campaign and additive profile unchanged. The common-cost and nested batch-size follow-ups were specified after earlier outcomes had been examined but before their own outcomes. An initial 40-batch paired study preceded the 160-batch follow-up, whose design was fixed before its outcomes. Subsequent common-correspondence, direct-calibration, and cost-generator audits reused those frozen 160 batches. Their protocols were fixed before their own outcomes were examined. Their pooled counts are descriptive, and complete timing and verification records are provided in the anonymous repository.

H SCOPE OF THE EMPIRICAL CONCLUSIONS

Observation model. The unpaired numerical bracket concerns an anchor- and cost-dependent finite-instance profile, not a population estimand. Assignments act on indexed batch atoms so duplicate rows are well defined. Their inverses are not claimed to extend to pointwise feature maps. The paired result additionally requires reference batches containing the clean counterparts of all target examples and independent paired calibration. Direct calibration uses true pairs from the fitting split that the geometric procedure does not use for fitting. Parameter-updating online test-time adaptation is outside the frozen correction class studied here. Agreement is not accuracy, and paired data are usually unavailable.

Statistical scope. Empirical conclusions are limited to the reported vision models, corruption benchmarks, feature maps, and calibration budgets. No ranking is established for segmentation, sequence models, nonlinear transport, or future shifts. WNID resampling conditions on the selected assignment and does not provide post-selection-valid inference for the optimized profile. The paired split-conformal statement is conditional on the fitted maps and concerns one future exchangeable batch within one cell, not joint multi-batch, cross-partition, population, unpaired, or natural-shift coverage. Each geometry, direct-calibration family, confidence level, and cell has a separate coverage event. The theoretical batchwise fallback in (20) has the same one-batch scope.

The ten original contrasts in Table 6 receive prespecified confirmatory adjusted inference. The four full-output paired-ridge contrasts in Table 12 form a separate, post-hoc Holm-adjusted family. The $q = .05$ endpoint-structure sign flip is another disclosed post-hoc one-test family. All other follow-up comparisons are descriptive. Natural shifts are not exchangeable with ImageNet-C corruption types, and overlapping split seeds and encoders are nested rather than independent replicates.

Empirical coverage. The common-cost endpoint comparison changes the endpoint sets and is not a causal effect of revealing pairs. The primary cross-shift comparison fixes the profile batch size at $1{,}024$. The nested sweep reaches $8{,}192$ but covers only ImageNet-Sketch and disjoint ImageNet-C, CLIP and DINOv3, one geometry, and $q = .05$. Near-complete same-class candidate availability does not remove ambiguity in this panel, but the audit does not isolate the effects of feature dimension, class-histogram imbalance, duplicates, or other size-dependent features. It establishes no population scaling law.

The paired study covers two severity-5 ImageNet-C corruptions, three overlapping-identity CLIP stability partitions, and one matched ViT-B/16 partition. The common-correspondence audit pools 160 reused batches descriptively: 31 of 33 proven common-only certificates occur when the hidden true pairing is feasible, and two are descriptive successes outside that sufficient event. Observed coverage neither implies true-pair feasibility nor validates a cost geometry by itself. The cost-generator audit does not authorize post-evaluation geometry selection or establish universal optimality of the Bures-generated cost. Pooled wrong-sign fractions are not conditional error guarantees for issued decisions. The controls, common-correspondence comparison, and condition-specific slack are therefore essential.

The profile–gain correlation is exploratory, does not yield a calibrated deployment rule, and cannot turn ambiguity into a realized-accuracy bound. The best streaming strategy was selected after evaluation, and all follow-up intervals are descriptive.

Numerical implementation. The exact mass-constrained capacity uses nonatomic standard-Borel laws and an unrestricted measurable-transport class. It is only an upper bound for the implemented affine and otherwise restricted correction families. Lagrangian upper bounds depend on the recorded floating-point solver model. A formal real-arithmetic certificate would require independently verified dual inequalities (Rump, 2010). Controlled symmetries and the finite-batch assignment test instantiate the declared assumptions; they do not yield bounds for realized natural shifts. The strongest anchored pairwise bound can exceed the full-world minimax value.