How to Reuse a Representation

Anonymous authors
Paper under double-blind review

Abstract

An invertible feature change can preserve every classifier logit yet change which inputs a newly fitted out-of-distribution (OOD) rule accepts. This raises an information question for representation reuse: what must remain available for the decisions and subsequent operations of a downstream procedure? We attach representations to the structures their consumers use and check how fitting, selection and calibration carry through to final decisions. In a fixed-penalty quadratic contract with the right-hand side retained, agreement of labels for all real targets and queries identifies a Hessian ray; requiring agreement both before and after one prescribed nonzero, zero-target row identifies the full Hessian ($d \geq 2$). In controlled same-logit OOD comparisons, mild feature rescalings change fully reselected and recalibrated ASH-B decisions by a median 6.42% across 24 model/OOD comparisons. Compatible constructions preserve their decisions on valid tested domains. What a representation can safely discard depends on its declared use.

1 Introduction

Reusing a learned representation means fitting new decision procedures to its features. An invertible feature change can still alter a nearest-centroid rule: stretching one coordinate and recomputing class means can change its distance comparisons. The centers move correctly, but the ruler changes. What must a representation interface preserve for its intended downstream operations?

We study *behavior-preserving reuse* under declared representation changes acting on corresponding training, selection, calibration and query data. The same procedure is fitted, selected and, when needed, calibrated at both interfaces. Preservation compares specified class labels or rejection decisions across these interfaces, whether right or wrong. A later learning step may change both answers; the question is whether corresponding uses still agree. Predictive utility is evaluated separately.

Interface compatibility matters in model stitching and modular world models (Bansal et al., 2021; Mai et al., 2026; Wang & Huang, 2025; Zhang et al., 2026). Function-preserving changes need not preserve learning (Kornblith et al., 2019), and same-function OOD diagnostics have found fixed-threshold differences (Karasan, 2026). The fitted-policy question here is what happens when the detector is independently fitted, selected and calibrated at each interface while the classifier logits remain fixed. Its accepted set can still change. This makes the complete decision procedure, rather than the original classifier alone, the object whose preservation we check. These experiments transform one existing interface; independently trained stitching and world-model recomposition have different contracts. OOD evaluation supplies the fitted-detector setting (Zhang et al., 2024); training choices can change OOD results despite similar classifier accuracy (Szyc et al., 2023).

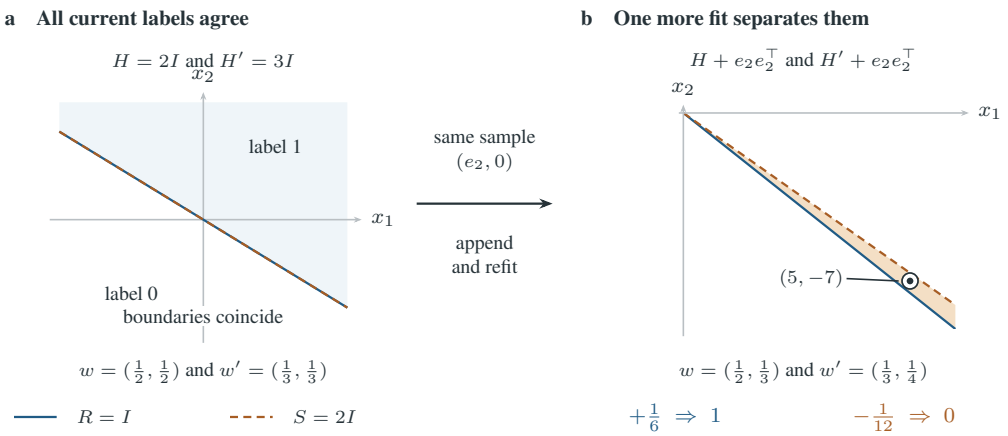
A downstream procedure, or *consumer*, reads particular structure: a distance rule uses a ruler, and a regularized readout uses a penalty and fitted coefficients. Its permitted next operation determines which details a reusable interface may discard. The Grothendieck construction organizes representations together with their consumer structures and admissible changes. This organization lets us identify the information each declared use needs and carry local fitting and scoring checks through selection, calibration and final decisions. It also gives compatible recipes across consumers.

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.

What the next operation needs. The declared ranking tails need score order, including ties. For quadratic fitting with fixed design X and penalties R, S, write $H = X^\top X + R$ and $H' = X^\top X + S$ for the two normal matrices. The states in Figure 1 give the same current labels but respond differently to one further fit. Under the all-target contract, current labels identify a Hessian ray; requiring agreement both before and after the specified nonzero row with target zero identifies the full Hessian (Section 4). For all scores and finite updates in a supplied subspace, a separate contract admits the exact interface $(w, H^{-1}U)$.

How compatibility reaches a decision. We check actual refitting and evaluation, then carry those checks through selection, calibration and final decisions (Sections 2–3). This yields preserving recipes for diagonal distances, diagonal penalties and RMS shaping.

What changes in fitted consumers. OOD and classification studies hold original classifier outputs fixed and refit the consumers (Section 5). They test whether the complete procedure preserves decisions and assess the utility of compatible recipes separately. The continued-fitting contrast above is a separate, exact information result.



Exact quadratic witness. Lines mark zero score; shading marks disagreement.

Figure 1: The same current labels need not survive another fit. For $X = I_2$, $y = (1, 1)^\top$ and no-intercept summed quadratic loss, penalties $R = I_2$ and $S = 2I_2$ give identical label boundaries (left). Appending $(e_2, 0)$ to both fits makes query $(5, -7)$ score $1/6$ versus $-1/12$ (right, zoomed in); a positive score has label 1. The two penalties also arise from one numerical ridge penalty I_2 under feature rescaling $g = I_2/\sqrt{2}$.

2 The representation and the structure it carries

Refitting the centers can still change the ruler. Fit class means $\mu_A = (0, 0)$ and $\mu_B = (2, 2)$, and classify query $h = (1.6, 0.6)$ by Euclidean distance. The squared distances $2.92, 2.12$ select B. Multiply every training and query feature's second coordinate by three and refit. The means become $(0, 0), (2, 6)$ and the query $(1.6, 1.8)$; distances $5.80, 17.80$ now select A. The means moved correctly. Giving the second squared difference weight $1/9$ restores both distances. This exact one-point-per-class example estimates no covariance; Figure 5 illustrates it.

Call a feature interface with its data roles a *presentation* p, supplying $h_p(x) \in \mathbb{R}^d$ for input identity x. A declared change $g : p \to gp$ gives $h_{gp}(x) = gh_p(x)$; p and gp are its two *endpoints*. We mainly use positive monomial g: positive coordinate rescalings and permutations. One fixed g acts on every training, selection, calibration and query role; input identities and labels correspond.

The consumer also uses a distance matrix K, coefficient vector w, or penalty R. To preserve $(h - \mu)^\top K(h - \mu)$, $w^\top h$, and $w^\top Rw$, their transformation laws, called *transports*, are

$$K' = g^{-\top} K g^{-1}, \qquad w' = g^{-\top} w, \qquad R' = gRg^\top. \tag{1}$$

The same feature rescaling thus divides a distance weight by nine but multiplies the corresponding penalty weight by nine. Reusing the numerical matrix I need not implement either law.

Knowing these laws is not enough: the fitting procedure must return the required structure. For each p, let $\mathsf{S}(p)$ be its admissible structures, with transports $g_* : \mathsf{S}(p) \to \mathsf{S}(gp)$ respecting identity and composition. A sufficient condition checks actual fitting F and evaluation E:

$$\underbrace{F(gp) = g_* F(p)}_{\text{actual refitting}}, \qquad \underbrace{E_{gp}(g_* s, x) = E_p(s, x)}_{\text{evaluation}}. \tag{2}$$

Then $E_{gp}(F(gp), x) = E_p(F(p), x)$. For means and a metric, recomputing means while returning $K = I$ again need not satisfy the first check for the full state. Merely making a transported metric available does not make the algorithm return it.

To organize these checks across interfaces, let G_d^+ denote the positive rescalings and permutations, and let $\mathcal{P} = P//G_d^+$ have presentations in a set P closed under these changes as objects and declared changes as arrows. The Grothendieck construction $\int_{\mathcal{P}} \mathsf{S}$ has

$$\text{objects } (p, s), \qquad \text{arrows } (p, s) \longrightarrow (gp, g_* s). \tag{3}$$

An object contains coordinates together with their ruler, penalty or fitted state; an arrow changes them compatibly (Vistoli, 2007). We use strict, set-valued families; Appendix A.1 gives the general convention.

The target is the *complete* output $T_p(x)$, including selection and calibration: $T_{gp}(x) = T_p(x)$ for all declared changes and queries. This is *descent*, the classical factorization through presentation orbits (Proposition 4). OOD outputs use $\mathbf{1}[q > \tau]$, comparing score q with cutoff τ; classification outputs a class ID. Data roles, candidate order, tie rules and any algorithmic randomness are fixed. Undefined cases return $\perp$; the transformation must preserve the domain where the procedure is defined. Undefined results remain in the comparison. Finite tests do not establish this universal property. We next construct a distance rule whose fitting and evaluation satisfy the two checks.

3 Constructing compatible fitting and evaluation procedures

We first fit a compatible ruler and follow its decision, then separate the local calculation from the proof steps and construction that other consumers can reuse.

3.1 A complete construction: fit the ruler and follow the decision

Estimate class means and pooled within-class covariance Σ from each interface's own training data. Assume all classes are represented and every diagonal entry of the pooled Σ is positive. Let D be the positive diagonal matrix with $D^2 = \operatorname{diag}\Sigma$. Standardize features by D^{-1}, shrink covariance toward I, and map it back:

$$M_\lambda = D\big[(1-\lambda)D^{-1}\Sigma D^{-1} + \lambda I\big]D = (1-\lambda)\Sigma + \lambda \operatorname{diag}\Sigma. \tag{4}$$

For $0 < \lambda \le 1$, M_λ is positive definite even if Σ is singular; its inverse is the ruler. This constructs a recipe whose *own* decisions we preserve.

Refitting under positive monomial g gives $\mu_c' = g\mu_c$ and $\Sigma' = g\Sigma g^\top$. Since taking a diagonal commutes with these rescalings and permutations, $M_\lambda' = gM_\lambda g^\top$. Evaluation then cancels the change:

$$(gh - g\mu_c)^\top (M_\lambda')^{-1}(gh - g\mu_c) = (h - \mu_c)^\top M_\lambda^{-1}(h - \mu_c). \tag{5}$$

Use confidence $q_\lambda(h) = -\min_c (h - \mu_c)^\top M_\lambda^{-1}(h - \mu_c)$. Each grid value of λ is a candidate distance rule. Selection chooses one using validation data; calibration sets its acceptance cutoff on calibration data. Every candidate score agrees on validation, calibration and query rows. Validation area under the receiver-operating-characteristic curve (AUROC)—the fraction of in-distribution (ID)/OOD pairs correctly ordered, with half credit for ties—therefore agrees. A fixed ordered grid and first-maximum rule select the same candidate; the same fixed calibration order-statistic index gives the same cutoff; and $q > \tau$ gives the same decision. Figure 2 shows the local checks and the downstream proof that another score producer can reuse.

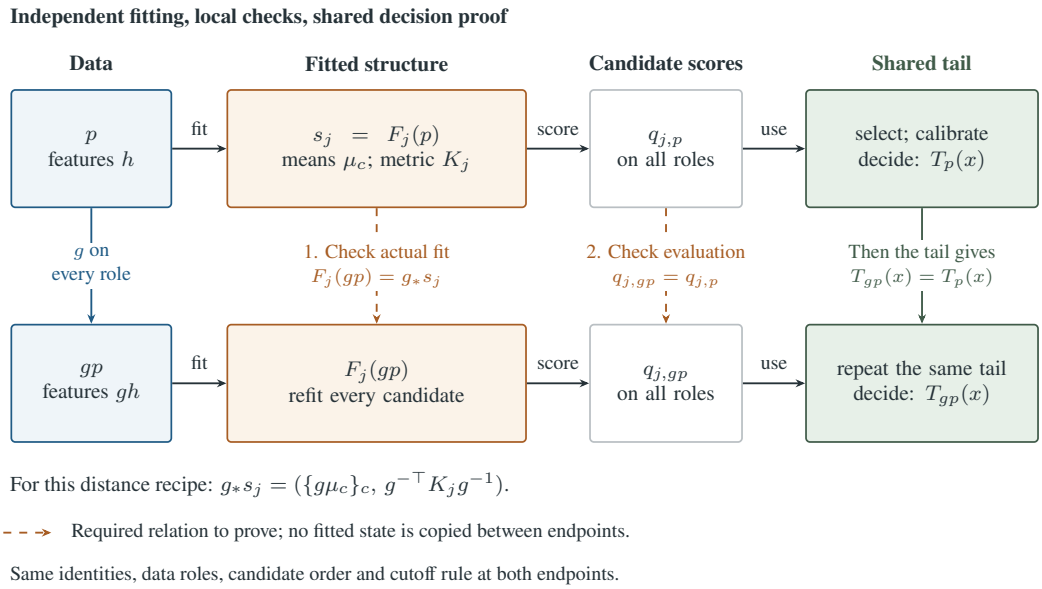


Figure 2: Local checks feed a shared complete-decision proof. Each row independently fits every candidate; dashed relations require fitted metrics to transport and scores to agree across validation, calibration and query roles. The shared tail then selects, calibrates and decides. No fitted state is copied. The positive-scale, nondegeneracy and fixed-rule domain is specified in Section 3.1. This diagram shows a proof mechanism, not data.

3.2 What another consumer can reuse

The distance calculation supplies the local work. The following classical composition principle supplies the common complete-program guarantee.

Proposition 1 (Composition of preservation conditions)**.** *For each presentation p, let a complete program have stages $a_{i,p} : X_i(p) \to X_{i+1}(p)$, $i = 0, \dots, m-1$. Suppose the stage states have transports $t_i(g)$ respecting identity and composition, and on the declared transport-stable domains*

$$a_{i,gp} \circ t_i(g) = t_{i+1}(g) \circ a_{i,p}. \tag{6}$$

Then $A_p = a_{m-1,p} \circ \cdots \circ a_{0,p}$ satisfies $A_{gp} \circ t_0(g) = t_m(g) \circ A_p$. If initial data correspond by $t_0(g)$ and $t_m(g)$ is identity on the terminal decision set, the complete decision descends.

Its proof substitutes the stage equations, cancelling internal transports (Appendix A.4). Every dependency a stage reads must be retained, including labels, optimizer choices and ties. Covariance of an objective transports its minimizers; the returned fit additionally needs uniqueness, a compatible selector, or a difference all later stages ignore. This is a sufficient certificate: equal terminal decisions need not make every raw intermediate state compatible.

The downstream selection, calibration and decision steps form the *tail*. They need less information than exact scores. Candidate order, data roles and the calibration order-statistic index are the same at both endpoints.

Lemma 1 (Selection and calibration from candidate score order)**.** *Fix an ordered candidate set, ID-positive validation AUROC with first-maximum selection, an order-statistic calibration cutoff, and the strict decision $q > \tau$. Suppose that, for every candidate j and declared arrow $g : p \to gp$,*

$$q_{j,gp}(x) = \phi_{j,g,p}(q_{j,p}(x)) \tag{7}$$

for a strictly increasing map common to all of that candidate's required validation, calibration, and query rows. Then the selected candidate and final decisions agree.

An attained order-statistic cutoff is an observed calibration score, not an interpolated value. For scores $(1, 2, 4)$, cutoff 2 and query 2.5, cubing gives cutoff 8 and query 15.625: the decision is unchanged. The proof preserves each candidate's validation orders and ties, then the selected candidate

index and cutoff comparison. It excludes row-dependent recodings, arbitrary selection metrics and interpolated quantiles. Classification analogously reuses preservation of candidate labels, validation correct counts and first-maximum selection. A new consumer thus supplies its candidate law and domain, then inherits the corresponding tail proof.

3.3 One normalization construction, different fitted structures

The distance recipe standardized coordinates before applying an ordinary recipe. Positive monomial changes leave only a permutation of these standardized features. Any complete consumer invariant to that remaining symmetry preserves its decisions. More generally, let ν normalize the interface and let H be the remaining symmetry group. If $\nu(gp) = h(g,p)\nu(p)$ with $h(g,p) \in H$, an H-invariant complete consumer C makes $C \circ \nu$ invariant to the original group G on domains preserved by the transformations (Proposition 5). Normalization acts on all consumed roles. It need not choose unique coordinates: whitening leaves an orthogonal freedom, which distances can ignore but coordinate ranking cannot.

For a linear probe, let X_c contain n centered training rows. Use total centered covariance $S = X_c^\top X_c/n$ and $D_0^2 = \operatorname{diag} S > 0$, rather than the distance recipe's within-class covariance. Ordinary ridge on $D_0^{-1}h$, with coefficients $\beta = D_0 w$, gives

$$\|\beta\|^2 = w^\top \operatorname{diag}(S) w. \tag{8}$$

The returned fit must still be checked. Let $\bar{h}, \bar{y}$ be the training feature and one-hot label means. With centered one-hot targets Y_c and $\lambda > 0$, minimizing $n^{-1}\|X_c B - Y_c\|_F^2 + \lambda \operatorname{tr}(B^\top R B)$, $R = \operatorname{diag} S$, gives

$$(S + \lambda R)B = X_c^\top Y_c/n, \qquad \beta_0 = \bar{y} - B^\top \bar{h}.$$

Positive variances and the positive grid make the solution unique. Feature rows transform as $X_c' = X_c g^\top$, with Y_c unchanged. Thus the normal matrix is $g(S + \lambda R)g^\top$ and its right-hand side is $g X_c^\top Y_c/n$. Thus refitting gives $B' = g^{-\top} B$ and the same unpenalized intercept $\beta_0' = \beta_0$, hence equal candidate class scores. This probe uses mean loss and a fitted intercept; Section 4's continuation theorem uses a separate summed-loss, no-intercept objective. The classification tail supplies selection and final labels. This is the new local calculation; the downstream proof is inherited.

Coordinate shaping exposes the same distinction. ASH retains large feature coordinates before the original head (Djurisic et al., 2023). Nonuniform rescaling can reverse a calibration/query ordering even after recalibration; Appendix B.5 gives a strict counterexample. Energy reads preserved raw logits (Liu et al., 2020); DICE's feature-mean/head-weight products cancel the scale (Sun & Li, 2022). Alternatively, RMS shaping fits uncentered second moments, removes coordinate scale before ASH ranking, and restores it before the compensated head. It has permutation residual symmetry, using positive fitted RMS and a permutation-compatible completion at boundary ties; ASH-S also requires a nonzero retained sum. The full formulas and domains are in Appendix B.5.

The same normalization principle gives diagonal distances, diagonal penalties and RMS shaping, but each consumer has its own fitted statistics, score law, domain and tie obligations (Table 2). Once those local checks hold, the selection, calibration and decision proof is shared. ASH supplies a strict counterexample when its coordinate ordering does not transport; failure of a sufficient law alone would not establish changed decisions. The shared tail needs score order rather than score magnitude. We next ask which summaries suffice for a declared use, and what another fit requires.

4 What information may a reusable interface forget?

The preceding tail proof uses only comparisons and ties: scores $(1, 2, 4)$ and $(1, 8, 64)$ have the same order. Storing their order suffices for that use; we now ask whether a later operation can also use the summary (Figure 6). A *reuse contract* specifies the outputs and subsequent operations that the interface must support. At a stage with state space X_i, let $\mathcal{C}_i$ be its permitted deterministic programs from that state to a final output, including supplied inputs, queries and output conventions. Group states that no permitted use distinguishes:

$$s \sim_i t \quad \Longleftrightarrow \quad c(s) = c(t) \text{ for every } c \in \mathcal{C}_i. \tag{9}$$

This is contextual equivalence (Pitts, 2002). For partial operations, equivalence also retains definedness: both states admit the operation or both fail.

Proposition 2 (Minimal information for composable reuse)**.** *The relation $\sim_i$ is the coarsest equivalence through which every observation in $\mathcal{C}_i$ factors. If $c \circ a \in \mathcal{C}_i$ whenever $a : X_i \to X_j$ is an allowed module and $c \in \mathcal{C}_j$, then each module induces a unique map $X_i/\!\sim_i \to X_j/\!\sim_j$, and these maps respect composition. Combining observation families intersects their equivalence relations; enlarging the reuse contract can therefore require finer information.*

If a later observation distinguishes $a(s)$ from $a(t)$, the composed observation already distinguishes s from t. This proves why closure under continuations makes the operation well defined on classes. It is an information ordering, not a minimum dimension or an automatic test for arbitrary learners. Transports respecting the observation families induce transports of the classes. Their Grothendieck construction again supports fitting checks and the shared composition proof (Appendix C.1). What changes is the information each state carries. Thus we can replace raw states by sufficient summaries and reuse the same proof checks, provided the allowed operations and transports still act on those summaries.

Some interfaces can discard magnitudes. For the ranking tails, the exact summary is each candidate's *weak order*: which scores are smaller, larger or tied (Proposition 6). The necessity statement fixes a finite score table on at least four identities, varies disjoint validation, calibration and query roles, ordered candidate subsets and attained cutoff indices, and does not refit the score producer. Isolating one candidate and using singleton calibration recovers every pairwise comparison; the common tail proves sufficiency. Fixed empirical roles need only the sufficient lemma. Other readouts require different information: removing a class-common quadratic score term keeps class labels but can reverse a recalibrated OOD decision (Appendix C.6).

4.1 Current agreement can fail after one more fit

Keeping exactly what determines current labels suffices for read-only use. Figure 1 shows why another fit can require more. Fix a full-column-rank design X, positive definite penalties R, S and the no-intercept summed objective

$$\tfrac{1}{2}\|Xw - y\|^2 + \tfrac{1}{2}w^\top Rw.$$

Put $A = X^\top X$, $b = X^\top y$ and $H = A + R$, the loss's curvature matrix (Hessian). The fit is $H^{-1}b$, with label $\mathbf{1}[x^\top H^{-1}b > 0]$. Each observation supplies real targets y and query x; b remains available when summarizing H. Appending $(v, 0)$ changes H to $H + vv^\top$ and leaves b unchanged. Penalties remain fixed throughout.

Proposition 3 (Current labels and one further fit require different information)**.** *Let $d \geq 2$ and fix any nonzero appended row v. For the fixed design X, the two penalties give identical current labels for every real target vector y and every query x if and only if $A + S = c(A + R)$ for some $c > 0$. They agree for all these targets and queries both before and after appending $(v, 0)$ if and only if $R = S$.*

Among admissible Hessians, current labels identify the positive ray $[H] = \{cH : c > 0\}$, whereas agreement both before and after the append identifies H itself. The proof is short. Full column rank makes $X^\top y$ range over every b. Equal positive decision halfspaces for all b, x force $H' = cH$; agreement after the append similarly gives $H' + vv^\top = k(H + vv^\top)$. A positive definite matrix in $d \geq 2$ and a rank-one matrix are linearly independent, forcing $c = k = 1$.

For $X = I_2$, $y = (1, 1)^\top$, $R = I_2$ and $S = 2I_2$, the fits $(1/2, 1/2)$ and $(1/3, 1/3)$ give the same label on every query. Appending $(e_2, 0)$ gives $(1/2, 1/3)$ and $(1/3, 1/4)$: query $(5, -7)$ now scores $1/6$ and $-1/12$. No more exhaustive test of current labels would expose this problem. The new sample adds the same curvature to both losses, but its size relative to their previous curvature differs. The comparison also arises from the same numerical ridge penalty I after feature rescaling $g = I/\sqrt{2}$: its pulled-back penalty is $g^{-1}Ig^{-\top} = 2I$, with all rows transformed alike.

The obstruction is operational: $U_v(H) = H + vv^\top$ cannot act on rays alone. $2I_2$ and $3I_2$ share a ray, but their updated matrices $\mathrm{diag}(2, 3)$ and $\mathrm{diag}(3, 4)$ do not. Keeping H supports the update and every old readout through $H \mapsto [H]$. Two contexts already identify H; adding repeated appends

cannot coarsen this requirement. The two-context family itself need not be closed under repetition. For a feature change g, the transformed interface's fixed penalty R' becomes $S = g^{-1}R'g^{-\top}$ in the original coordinates, and the corresponding new row is gv. Applying the theorem to R, S therefore requires exactly $R' = gRg^{\top}$.

These claims quantify over all targets and queries; they do not cover a single realized target vector, $d = 1$, a fitted intercept, mean-loss rescaling or data-dependent penalties and selection. For one fixed target, even agreement before and after the append need not identify H. They are distinct from the sufficient probe construction and finite classification study. Appendix C.3 gives the full proof and excluded cases.

An exact score interface for restricted updates. For this score-based contract, let the initial state be (H, b) with $H \in \mathbb{R}^{d\times d}$ positive definite, $b \in \mathbb{R}^d$, and $w = H^{-1}b$. Supply a known full-column-rank $U \in \mathbb{R}^{d\times r}$, $1 \le r \le d$; observe every score $x^{\top}w$ and permit finite (Ua, y) appends for all $a \in \mathbb{R}^r, y \in \mathbb{R}$. Put $K = H^{-1}U$, $G = U^{\top}K$ and $D = 1 + a^{\top}Ga$. Classical RLS (Boyd, 2008) closes on (w, K) through

$$w^{+} = w + \frac{Ka}{D}(y - a^{\top}U^{\top}w), \qquad K^{+} = K - \frac{Ka\,a^{\top}G}{D}. \tag{10}$$

Two states agree under every permitted score and continuation iff (w, K) agrees: empty continuations identify w, and one-step targets 1 versus 0 identify each column of K (Proposition 7). Thus only the inverse Hessian's action on legal update directions is needed. For invertible g, transport the supplied directions too:

$$(H', b', U') = (gHg^{\top}, gb, gU) \quad \Longrightarrow \quad (w', K') = (g^{-\top}w, g^{-\top}K). \tag{11}$$

Then $U'^{\top}K' = U^{\top}K$, $U'^{\top}w' = U^{\top}w$, and $x' = gx$ preserves scores. The denominator and residual in (10) are invariant, so the update commutes with transport. The proof and numerical execution check are in Appendix C.4.

Table 1: Exactly identified structures under the displayed observations and permitted operations. The last two rows are distinct continuation contracts.

Declared uses	Exactly identified structure
All ranking tails of Proposition 6	Candidatewise weak orders and tie blocks
All two-centroid labels for fixed metrics	Positive ray of the fixed metric
All real targets and queries for a fixed quadratic design	Positive regularized-Hessian ray $[H]$
The preceding uses plus one fixed nonzero row with target zero	Full regularized Hessian H ($d \ge 2$)
All scores; finite appends in supplied $\operatorname{im} U$	$(w, H^{-1}U)$

Together, the rows show how declared observations and operations determine the state an interface must retain (Table 1). We next test complete refitting after representation change under its own contract.

5 Do the structural laws distinguish actual fitted consumers?

For an OOD rejection rule attached to a classifier, the operational question is whether the accepted set survives a representation change after the detector is fitted, selected and calibrated again. We test that question with consumers for which the structural checks predict preservation or expose a failure, and evaluate predictive utility separately. The preservation laws use exact arithmetic; finite numerical checks use the protocol's tolerances. Three CIFAR-100 ResNet-18 checkpoints and one ViT interface share fixed fitting, selection, calibration and evaluation roles. One invertible map acts on all roles, with original head compensation

$$h' = gh, \qquad W' = Wg^{-1}, \qquad b' = b, \tag{12}$$

so original logits agree. This controls the classifier while testing an additional consumer. Exact invertibility preserves feature information; it does not guarantee bitwise recovery of rounded arrays (Appendix A.5).

For independently refitted consumers, each interface rebuilds its statistics, reselects by first-maximum validation AUROC (classification: accuracy), and recalibrates an attained fifth-percentile OOD cutoff with $q > \tau$. Roles are fixed; evaluation identities are disjoint from fitting, selection and calibration and select no candidate or cutoff. An OOD cell is one model/transformation/dataset, with disagreement the fraction of changed Boolean decisions. Cells share models and images, so these are descriptive comparisons, not independent deployment draws. Appendix D fixes sample counts, grids, ties and numerical checks. Figure 3 shows the principal mild-rescaling decision comparison and its distinct utility check. The full distance and classification contrasts, including individual classification-model values, appear in Figure 8 in Appendix E.

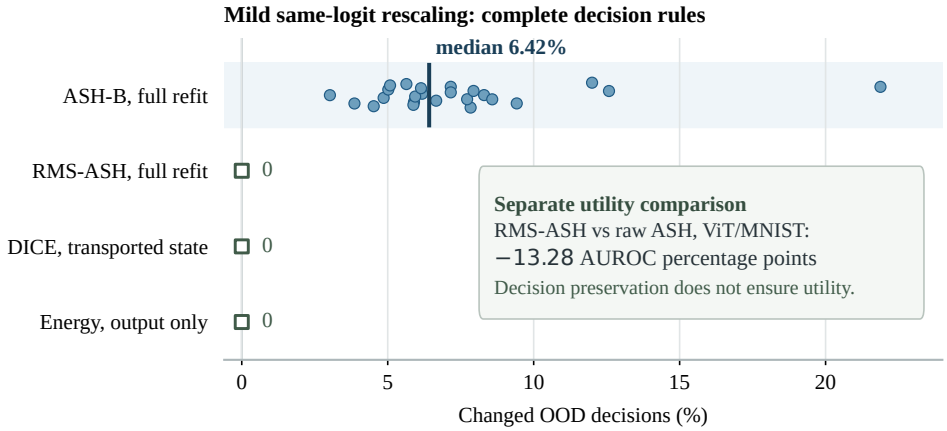


Figure 3: Mild, formula-derived positive diagonal changes preserve original classifier logits. Dots show 24 model/OOD cells: fully reselected and calibrated ASH-B decisions differ (median 6.42%), while RMS-ASH and Energy agree after endpoint-local fitting; DICE uses transported fitted state. The boxed RMS loss is a separate baseline ViT/MNIST utility comparison against raw ASH. Shared models/images make these descriptive cells, not independent draws or accuracy loss.

Complete adaptation can leave changed decisions. One SmoothQuant-formula rescaling per model has condition number $\kappa = 1.56$–1.84 (largest/smallest scale). The changed ASH-B decisions realize the coordinate-order obstruction after the entire adaptation procedure; the controls isolate which structure each consumer reads. The original DICE study transports fitted state, although its proof also covers recomputation. These are designed formula transformations without quantization or export. Stress and new-transformation CIFAR-10 results appear in Appendix E; models and images are reused in the latter. On fresh-identity EuroSAT, with a separate arrow realizing $\kappa = 6.67$, ASH AUROC moves $86.20\% \rightarrow 86.25\%$, while OOD acceptance at separately calibrated cutoffs rises $57.50\% \rightarrow 64.58\%$ (13.52% paired disagreement). ID-test acceptance also rises $94.81\% \rightarrow 96.19\%$: ranking and decisions differ, without a matched ID-test acceptance comparison (Appendix E.9). AUROC alone therefore cannot certify identity of the thresholded policy's accepted set. The changed ID operating point matters when interpreting the OOD acceptance-rate shift. A manually transformed live-graph replay on 4,972 raw images per model checks agreement with cache-based terminal outputs (Appendix E.8); it does not measure natural exporter incidence.

Fitted geometry and penalty laws predict preserving cases. Each distance consumer independently rebuilds its bank or covariance, selects and calibrates under the mild transformation and one fixed $\kappa = 2$ transformation per model. Diagonal shrinkage changes no decisions in 48 comparisons; isotropic shrinkage changes decisions in 42/48, and raw/normalized kNN in 48/48 each. Full covariance preserves all 36 valid comparisons; ViT's covariance fails the rank gate and remains undefined. Every kNN endpoint selects $k = 1$.

Both diagonal classifiers preserve decisions in all eight comparisons each. Euclidean-centroid median disagreement is 1.57% (mild) and 4.21% ($\kappa = 2$).

The new probe's local solve licenses the shared classification proof. An isotropic probe nevertheless agrees under both transformations for two ResNets: lacking a general law need not force finite failure. Accuracy changes span -0.53 to $+0.40$ percentage points. Figure 8 shows all consumers; Appendices E.6 and E.7 retain full values and protocol boundaries.

Preservation, utility and recovery have different evidence. RMS-ASH preserves decisions but fails its predeclared utility criterion: ViT/MNIST loses 0.13282 AUROC relative to raw ASH against a 0.10 bound, although mean margins pass. With the old state a and known g, complete transport instead recovers the old rule via $S_{g,a}(z) = gS_a(g^{-1}z)$ and transports the remaining selection, cutoff and tie state. Access to this additional state and map preserves old outputs and utility; the RMS failure implies no unavoidable invariance–utility tradeoff.

On new evaluation identities, the first prospective ASH/Energy study changes ASH decisions by 2.92% and 13.52% but finds no resolved AUROC ranking reversal beyond its ± 2-point band. ID accuracy is about 67.7%; this is not an equivalence result. Its ImageNetV2 successor fails the predeclared classifier-competence gate and stops before any detector evaluation, supplying no second detector outcome (Appendix E.9). State-copy diagnostics remain secondary and coordinate-dependent (Appendix F). The empirical conclusion is that full adaptation can leave structural dependence, while the proved constructions retain their own decisions on the valid tested endpoints.

6 Related work

Function preservation need not preserve learning (Kornblith et al., 2019). Structural transports and contextual equivalence are established (Maruyama, 2026; Pitts, 2002), as are RLS and projected covariance updates (Boyd, 2008; Li et al., 2014; Lai & Bernstein, 2024). We use these tools to compute declared-use information and check actual fitting, selection, calibration and decisions. ASH/DICE, kNN OOD and Mahalanobis supply consumers (Djurisic et al., 2023; Sun & Li, 2022; Sun et al., 2022; Lee et al., 2018); equalization motivates rescalings (Meller et al., 2019; Xiao et al., 2023). Same-function diagnostics can retain a fixed threshold (Karasan, 2026); ours reselects and recalibrates. Appendix G gives fuller comparisons.

7 Scope and conclusion

An invertible, same-logit change can alter a fully reselected and recalibrated OOD policy's accepted set. We make reuse checkable by declaring observations and operations, retaining what they distinguish, then proving local fitting and scoring compatible through the final decision. This gives preserving normalization recipes and exact information contrasts: weak orders for the declared ranking tails, a full Hessian for the specified all-target label-and-append contract, and $(w, H^{-1}U)$ for scores and finite updates in a known subspace. The last interface updates recursively and respects coordinate transport. The full-refit experiments test a separate use; preservation, utility and natural incidence have distinct evidential requirements.

REPRODUCIBILITY STATEMENT

The supplementary ZIP provides consumer code, finite mathematical tests, stored decision and tie records for the paired state-copy and CIFAR-10 transfer studies, and saved synthetic continuation inputs with replay commands. These checks reconstruct decision arithmetic and synthetic trajectories; they do not regenerate real-image features, detector scores, selections, or cutoffs. Included result summaries retain the failed RMS utility gate and the projected-feature continuation study's zero initial-to-final label changes. Images, weights and real-image feature caches are excluded. The ZIP's README specifies coverage, dependencies and expected outputs; Appendices D–F give the empirical protocols, results and original validation boundaries, including the distinct cached-feature and raw-image replays.

AI USE STATEMENT

OpenAI Codex was used for research design, experiments, code, analysis, literature synthesis, artifacts, mathematical checks, and writing/editing. Authors retain responsibility for the work.

REFERENCES

Amir Asiaee. Causal mechanism reduction: Mechanism replacement for neural network pruning and abstraction, 2026. URL https://arxiv.org/abs/2602.24266.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://papers.neurips.cc/paper/2021/hash/01ded4259d101feb739b06c399e9cd9c-Abstract.html.

Stephen Boyd. Least-squares applications, EE263 lecture 6. Stanford University lecture notes, 2008. URL https://ee263.stanford.edu/archive/ls-app.pdf.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral CNN. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 1321–1330, 2019. URL https://proceedings.mlr.press/v97/cohen19d.html.

Gabriele D'Acunto, Enrico Grimaldi, Valeria Avino, Mario Edoardo Pandolfo, Leonardo Di Nino, Sergio Barbarossa, and Paolo Di Lorenzo. Sheaf-based federated representation learning, 2026. URL https://arxiv.org/abs/2608.10016v1. Version 1.

Berker Demirel, Marco Fumero, and Francesco Locatello. Out-of-distribution detection with relative angles. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/863d42e2c34d5f65b9ccec5b0ddc2827-Abstract-Conference.html.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1019–1028, 2017. URL https://proceedings.mlr.press/v70/dinh17b.html.

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.09858.

Elisenda Grigsby, Kathryn Lindsey, and David Rolnick. Hidden symmetries of ReLU networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11734–11760, 2023. URL https://proceedings.mlr.press/v202/grigsby23a.html.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition, 2009. doi: 10.1007/978-0-387-84858-7.

Rahim Hossain, Md Tawheedul Islam Bhuian, Md Farhan Shadiq, and Kyoung-Don Kang. MM++: Unsupervised scale-invariant multilayer OOD detection via top-k gated feature fusion, 2026. URL https://arxiv.org/abs/2606.17352. arXiv:2606.17352v1.

Abdullah Karasan. Representation measurements under function-preserving reparameterizations, 2026. URL https://arxiv.org/abs/2608.27020.

Agnan Kessy, Alex Lewin, and Korbinian Strimmer. Optimal whitening and decorrelation. *The American Statistician*, 72(4):309–314, 2018. doi: 10.1080/00031305.2016.1277159. URL https://arxiv.org/abs/1512.00809v4.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3519–3529, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

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

Brian Lai and Dennis S. Bernstein. SIFt-RLS: Subspace of information forgetting recursive least squares, 2024. URL https://arxiv.org/abs/2404.10844v1. Version 1.

Ismail Lamaakal, Chaymae Yahyati, Khalid El Makkaoui, Ibrahim Ouahbi, and Yassine Maleh. SNAP-UQ: Self-supervised next-activation prediction for single-pass uncertainty in TinyML. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=YdK1ZRhrKi.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, 2018. URL https://arxiv.org/abs/1807.03888.

E. L. Lehmann and F.-W. Scholz. Ancillarity. In *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, Institute of Mathematical Statistics Lecture Notes–Monograph Series, pp. 32–51. Institute of Mathematical Statistics, 1992. doi: 10.1214/lnms/1215458837.

Judith Yue Li, Sivaram Ambikasaran, Eric F. Darve, and Peter K. Kitanidis. A Kalman filter powered by $\mathcal{H}$-matrices for quasi-continuous data assimilation problems. *Water Resources Research*, 50(5):3734–3749, 2014. doi: 10.1002/2013WR014607. URL https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2013WR014607.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f5496252609c43eb8a3d147ab9b9c006-Abstract.html.

Saunders Mac Lane. *Categories for the Working Mathematician*, volume 5 of *Graduate Texts in Mathematics*. Springer, 2 edition, 1998.

Zheda Mai, Ke Zhang, Fu-En Wang, Zixiao Ken Wang, Albert Y. C. Chen, Lu Xia, Min Sun, Wei-Lun Chao, and Cheng-Hao Kuo. Revisiting model stitching in the foundation model era. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41342–41351, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Mai_Revisiting_Model_Stitching_In_the_Foundation_Model_Era_CVPR_2026_paper.html.

Yoshihiro Maruyama. Foundations of equivariant deep learning: Unifying graph and sheaf neural networks, 2026. URL https://arxiv.org/abs/2607.03798.

Eldad Meller, Alexander Finkelstein, Uri Almog, and Mark Grobman. Same, same but different: Recovering neural network quantization error through weight factorization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4486–4495, 2019. URL https://proceedings.mlr.press/v97/meller19a.html.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Anthony Nguyen, Antonin Bertrand, Sylvie Le Hégarat-Mascle, Emanuel Aldea, Franck Florin, Mohammed Nabil El-Korso, and Renaud Lustrat. Fisher–rao sensitivity for out-of-distribution detection in deep neural networks. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10010515.

Andrew M. Pitts. Operational semantics and program equivalence. In *Applied Semantics*, volume 2395 of *Lecture Notes in Computer Science*, pp. 378–412. Springer, 2002. URL https://www.cl.cam.ac.uk/~amp12/papers/opespe/opespe-lncs.pdf.

Hrittik Roy, Marco Miani, Carl Henrik Ek, Philipp Hennig, Marvin Pförtner, Lukas Tatzel, and Søren Hauberg. Reparameterization invariance in approximate bayesian inference. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0f934dd2030f5740cde0aa2697a105a9-Abstract-Conference.html.

Yiyou Sun and Yixuan Li. DICE: Leveraging sparsification for out-of-distribution detection. In *European Conference on Computer Vision*, 2022. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/html/4405_ECCV_2022_paper.php.

Yiyou Sun, Chuan Guo, and Yixuan Li. ReAct: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/01894d6f048493d2cacde3c579c315a3-Abstract.html.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of the 39th International Conference on Machine Learning*, 2022. URL https://proceedings.mlr.press/v162/sun22d.html.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Kamil Szyc, Tomasz Walkowiak, and Henryk Maciejewski. Why out-of-distribution detection experiments are not reliable—subtle experimental details muddle the OOD detector rankings. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2078–2088, 2023. URL https://proceedings.mlr.press/v216/szyc23a.html.

Viet-Hoang Tran, Thieu N. Vo, Tho Tran-Huu, An T. Nguyen, and Tan M. Nguyen. Monomial matrix group equivariant neural functional networks. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/577cd5863ec73be4e6871340be0936ae-Abstract-Conference.html.

Angelo Vistoli. Notes on Grothendieck topologies, fibered categories and descent theory, 2007. URL https://arxiv.org/abs/math/0412512v4. Version 4.

Xinyue Wang and Biwei Huang. Modeling unseen environments with language-guided composable causal components in reinforcement learning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/79d86433c2acd12b6fa98553435d226e-Abstract-Conference.html.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38087–38099, 2023. URL https://proceedings.mlr.press/v202/xiao23c.html.

Kai Xu, Rongyu Chen, Gianni Franchi, and Angela Yao. Scaling for training time and post-hoc out-of-distribution detection enhancement. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/46d943bc6a15a57c923829efc0db7c7a-Abstract-Conference.html.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking generalized out-of-distribution detection. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/d201587e3a84fc4761eadc743e9b3f35-Abstract-Datasets_and_Benchmarks.html.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection, 2024. URL https://arxiv.org/abs/2306.09301v5. Version 5, accepted by Data-centric Machine Learning Research.

Shaowei Zhang, Jiahan Cao, Xunlan Zhou, Shenghua Wan, and De-Chuan Zhan. BRICKS-WM: Building reusability via interface composition kinetics for structured world models, 2026. URL https://arxiv.org/abs/2606.16489v1. Version 1.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018.

A COMMON STRUCTURE AND COMPLETE-PROGRAM GUARANTEES

Sections 2–3 derive a complete guarantee from local fits and evaluations. This appendix checks the underlying constructions, orbit criterion, domain requirement, and shared selection/calibration proof. These are classical structural principles specialized to the declared procedures.

A.1 ATTACHED STRUCTURES AND COMPATIBLE FITTED SECTIONS

Let $\mathcal{B} = P//G$ on a G-stable domain. A strict functor $\mathsf{S} : \mathcal{B} \to \mathbf{Set}$ assigns admissible structures and transports $g_* : \mathsf{S}(p) \to \mathsf{S}(gp)$ with $(kg)_* = k_* g_*$ and $1_* = 1$. Its category of elements has objects (p, s) and an arrow over g precisely when the target is $(gp, g_* s)$. A returned fit $F(p) \in \mathsf{S}(p)$ determines an object assignment. It extends to a section functor precisely when $F(gp) = g_* F(p)$.

Let $\mathcal{X}$ be the common query identity set, and let $E_p : \mathsf{S}(p) \to D^{\mathcal{X}}$ evaluate the specified output. If $E_{gp}(g_* s) = E_p(s)$ on the relevant states and the returned fit is compatible, then $T(p) = E_p(F(p))$ satisfies

$$T(gp) = E_{gp}(F(gp)) = E_{gp}(g_* F(p)) = E_p(F(p)) = T(p).$$

The fitting law and the evaluation law are distinct obligations. For means and metrics, the former follows from averaging and covariance, while the latter follows from the distance pairing. Defining a state relation solely by already equal terminal outputs would restate the desired conclusion without deriving either obligation.

For a strict $\mathsf{S} : \mathcal{B} \to \mathbf{Cat}$, the covariant Grothendieck construction has arrows

$$(p, s) \xrightarrow{(g,\alpha)} (gp, t), \qquad \alpha : g_* s \to t,$$

with composition $(k, \beta)(g, \alpha) = (kg, \beta \circ k_* \alpha)$. This convention is an opfibration; inverses in the base groupoid give the equivalent pullback convention (Vistoli, 2007). A section requires comparisons $\alpha_g : g_* F(p) \to F(gp)$ satisfying the identity and composition laws. Because the base arrows are invertible, such comparisons are necessarily isomorphisms. An observer respecting the resulting lifted arrows again gives invariant terminal outputs. A functor from the whole total category to the discrete category on $D^{\mathcal{X}}$ is a sufficient, stronger observer condition. Pseudofunctor versions insert their coherent comparison isomorphisms; the strict set-valued examples here do not require those choices.

If a chosen algorithm family is not closed under the declared structure transport, one can sometimes enlarge the state family to include transported rules. This changes what states are available. It does not make the original independent refitting procedure a compatible section of that larger family.

A.2 TOTAL AND PARTIAL DESCENT

Let BG_d^+ be the one-object groupoid whose arrows are the group elements, and let $\mathbf{Set}$ denote the category of sets. Regard the action as a functor $\mathsf{P} : BG_d^+ \to \mathbf{Set}$ with $\mathsf{P}(*) = P$. Its Grothendieck construction $\mathcal{P} = \int_{BG_d^+} \mathsf{P} \simeq P//G_d^+$ has presentations as objects and an arrow $g : p \to g \cdot p$ for each action. The map $\pi : \mathrm{Ob}(\mathcal{P}) \to \pi_0(\mathcal{P})$ is the coequalizer of source and target, or equivalently the map to presentation orbits.

Proposition 4 (0-truncated descent)**.** *For fixed x, $T_x : P \to D$ factors uniquely as $T_x = \widetilde{T}_x \circ \pi$ with $\widetilde{T}_x : \pi_0(\mathcal{P}) \to D$ if and only if $T_x(p) = T_x(g \cdot p)$ for every arrow $g : p \to g \cdot p$ in $\mathcal{P}$.*

Here "0-truncated" means that only discrete decision values are retained.

Proof of Proposition 4. If $T_x = \widetilde{T}_x \circ \pi$, then source and target of every arrow have the same π-image, hence the same decision. Conversely, endpoint equality makes T_x constant along every finite zig-zag of arrows. It is therefore constant on each connected component, and $\widetilde{T}_x([p]) = T_x(p)$ is well defined and unique. □

Let $T : \mathrm{Ob}(\mathcal{P}) \rightharpoonup D$ have domain U, and define $\overline{T} : \mathrm{Ob}(\mathcal{P}) \to D \sqcup \{\bot\}$. If $\overline{T}$ is arrow-constant and one endpoint belongs to U, its non-bottom value forces the other endpoint into U; hence U is

replete. Equality inside U follows. Conversely, a replete U leaves only two cases for any arrow: both endpoints are outside, where both values are bottom, or both are inside, where the assumed equality holds. Proposition 4 applies.

A finite check proves equality only on its listed arrows. If a finite groupoid is generated by a set S, equality on every reachable S-edge implies equality along compositions. Merely checking each generator once at one base object does not cover generator edges from newly reached objects.

A.3 Complete procedure and common downstream argument

Fix the data roles, candidate order, tie rules and deterministic operations, and write

$$(p, \mathcal{D}_{\text{fit}}) \xrightarrow{\text{fit}} a_p \xrightarrow{\text{score}} q_p \xrightarrow{\text{calibrate}} \tau_p \xrightarrow{\delta} T_p. \tag{13}$$

Here $\mathcal{D}_{\text{fit}}$ contains all non-evaluation roles; a_p includes fitted statistics and selected hyperparameters, and $T_p(x) = \delta(q_p(x), \tau_p)$ is the final Boolean decision. In our OOD protocols $\delta(q, \tau) = \mathbf{1}[q > \tau]$. A classification consumer instead returns a class ID and omits cutoff calibration. Let D denote the specified output set (Boolean decisions or class IDs). At each endpoint, the same procedure is run again; its fitted state need not be numerically identical.

A.4 Proof of composition and the common selection/calibration tail

Proof of Proposition 1. Substitute Equation (6) successively:

$$a_{m-1,gp} \cdots a_{0,gp} t_0(g) = a_{m-1,gp} \cdots a_{1,gp} t_1(g) a_{0,p} = t_m(g) a_{m-1,p} \cdots a_{0,p}.$$

Corresponding initial states therefore yield outputs related by $t_m(g)$. When this last map is identity, the terminal decisions agree. The argument applies on the declared reachable domains if they are preserved by transports and stage operations. For branches, every use of a shared state must have the same transport. For partial outputs extended by $\perp$, repleteness is additionally necessary, as in Appendix A. □

Proof of Lemma 1. For candidate j, empirical ID-positive AUROC averages $\mathbf{1}[q_j(x) > q_j(y)] + \frac{1}{2}\mathbf{1}[q_j(x) = q_j(y)]$ over validation ID/OOD pairs (x, y). A common strictly increasing map preserves both terms, so the entire ordered vector of candidate AUROCs is unchanged. The fixed first-maximum rule selects the same j_*. Sorting its calibration scores commutes with $\phi_{j_*,g,p}$, hence the order-statistic cutoff is $\tau_{gp} = \phi_{j_*,g,p}(\tau_p)$. Finally,

$$q_{j_*,gp}(x) > \tau_{gp} \iff \phi_{j_*,g,p}(q_{j_*,p}(x)) > \phi_{j_*,g,p}(\tau_p) \iff q_{j_*,p}(x) > \tau_p.$$

Only a map on the attained range, including the cutoff, is needed; a bijection on all of $\mathbb{R}$ is unnecessary. These per-arrow relations suffice for terminal invariance. To regard the recodings themselves as functorial transports, additionally require $\phi_{j,kg,p} = \phi_{j,k,gp} \circ \phi_{j,g,p}$ and identity/inverse relations on the corresponding ranges. □

The following numerical selection example makes each step explicit. Candidate A has validation ID scores $(3, 4)$ and OOD scores $(1, 2)$, giving AUROC 1; candidate B has ID $(2, 4)$ and OOD $(1, 3)$, giving $3/4$. Cubing A and applying $10s + 7$ to B preserves both AUROCs and selects A again. For A, calibration scores $(1, 2, 4)$ give second-order-statistic cutoff 2; a query score 2.5 is accepted. Cubing every score changes the cutoff to 8 and the query to 15.625, preserving the decision. The proof preserves each candidate's pairwise validation orders and ties, then selection, then its cutoff comparison. It does not cover row-dependent recodings, arbitrary selection metrics, or interpolated quantiles. For class-label consumers, candidate-label preservation instead preserves validation correct counts, first-maximum selection, and final labels. These common tails let a new consumer inherit the downstream argument once its candidate-level law and domain have been established. The following counterexample shows why attained cutoffs matter.

Interpolation is a substantive exception. Calibration scores $(1, 3)$ have interpolated median 2, so a query at 2.1 is accepted. Cubing scores produces cutoff $(1 + 27)/2 = 14$ and query $2.1^3 = 9.261$, which is rejected. The recoding preserves orders but not this calibration functional. Different row-wise recodings can also change AUROC; selection based on raw scores across candidates needs its own contract.

For classification, suppose each candidate predicts the same label on corresponding selection and query rows. Its validation correct count is then unchanged, so fixed first-maximum selection and final labels agree. A transported argmin set does not by itself specify those returned labels or coefficients: a unique solution, compatible selection, or a proved downstream equivalence is needed.

A.5 Information, recalibration, and explicit transport

Fix a measurable feature variable $H : \Omega \to \mathbb{R}^d$, the same target Y, and a fixed invertible g. Both $H' = gH$ and $H = g^{-1}H'$ are measurable functions of one another, so $\sigma(H') = \sigma(H)$. For every measurable rule d, the rule $d_g = d \circ g^{-1}$ satisfies $d_g(H') = d(H)$ pointwise, and the correspondence is bijective. Under the same nonnegative loss, the sets of attainable risks, and hence their infima, agree. This is an unrestricted information statement, not an assertion about identical fitted rules, inductive biases, or resource budgets. If the transformation G is itself fitted from data, retain it on both sides: $\sigma(H, G) = \sigma(GH, G)$. Dropping G need not preserve information. The finite endpoint check fixes the realized transformation.

These information equalities concern exact invertible maps. They do not certify bitwise recovery of rounded arrays or information preservation under quantization.

For two fitted scores q, q' on the same domain, equality of all strict pairwise orders, including ties, is equivalent to $q' = \phi \circ q$ for a strictly increasing ϕ on the attained range of q. Indeed, pairwise comparisons in both directions make $\phi(q(x)) = q'(x)$ well defined and increasing. Such a map commutes with sorting, so the order-statistic cutoff and strict decisions agree for every calibration tuple. Conversely, agreement for every tuple includes singleton calibration at each y, which recovers every comparison $q(x) > q(y)$. Our cutoff is an order statistic, not an interpolated quantile. A decision flip therefore cannot be explained by one increasing relabeling of the fitted scores on all consumed rows. This does not imply an AUROC loss or a benchmark ranking reversal.

If an increasing relabeling exists at each candidate percentile on all calibration, validation and evaluation rows, every validation AUROC and tie is preserved. The fixed first-maximum APS rule then selects the same candidate, and its calibrated decisions agree. This is sufficient, not necessary; the empirical full-refit findings are not deduced merely from failure of this condition.

Finally fix the original fitted state a and a partial shaping map S_a. With a known arrow and retained reference, define

$$S_{g,a}(z) = gS_a(g^{-1}z), \qquad \mathrm{Dom}(S_{g,a}) = g\,\mathrm{Dom}(S_a).$$

Then $(Wg^{-1})S_{g,a}(gh) + b = WS_a(h) + b$. Transport all state, including hyperparameters, statistics, cutoff and the reference-coordinate tie policy, to preserve the complete scores, decisions, and every utility functional of those outputs and labels. Successive transport by g and u is transport by ug. Retaining that arrow/reference is additional structure: different arrows to one bare endpoint can induce different consumers if a transformation that leaves the endpoint unchanged (a stabilizer element) acts nontrivially on the consumer. Thus this construction is not an automatic reference-free descent certificate. It is a legitimate exact remedy, and the failed RMS recipe does not establish an inherent invariance–utility tradeoff.

B Constructing and checking individual consumers

The complete distance example established both what refitting returns and why its evaluation agrees. This appendix verifies the common normalization principle, then gives the exact local objectives, domains and coordinate-shaping rules used by those arguments. It separates these definitions from the empirical protocols.

B.1 A geometric illustration and consumer comparison

The main-text centroid calculation isolates the geometry used after fitting means. Figure 5 shows its three distance rules; each class may have one training point, and this illustration does not estimate a covariance.

Table 2: What the framework lets a consumer reuse. Local laws include the actual fit, all consumed data roles, deterministic selection and tie rules, and the stated domains. The common tail is Lemma 1 for OOD or its class-label counterpart. Empirical entries refer to their separate studies; zero disagreements are finite checks of conditional guarantees. Distance and classification results are in Section 5. A missing sufficient certificate alone does not imply failure.

Consumer	Consumer-specific work	Reused guarantee or construction	Empirical check
Raw ASH-B	Coordinate selection can reverse calibration/query order	Exact terminal counterexample; full reselection requires its own test	Mild and stress studies: decisions change
DICE; Energy	Paired feature/head contributions cancel; raw logits agree	Candidate equality feeds the common tail	Zero changes; stress and mild DICE use transported state
RMS-ASH-B	Positive fitted RMS removes scale; completed shaping respects permutations	Residual-symmetry construction, then common tail	Zero changes; separate utility criterion fails
Full/diagonal covariance distances	Fitted means and covariance obey the distance transport law	Candidate distance equality feeds OOD or classification selection	Distance study: zero on valid domains; diagonal classifier preserved
Diagonal-penalty probe	Total-variance penalty and returned solve transform compatibly	Standardized ridge construction, then class-label tail	Classification study: zero changes after independent refitting
Euclidean / isotropic consumers	Reusing numerical geometry need not transport it	No general nonuniform-scale guarantee; test terminal outputs	Distance/classification studies: changes and agreeing cases
Complete transport	Retain the old fitted rule and known inverse map	Recover that reference rule, including selection and cutoff	OOD and classification studies: baseline decisions recovered

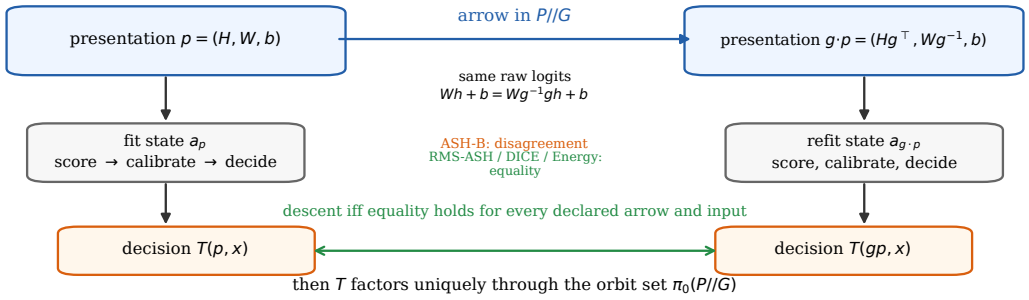


Figure 4: A controlled OOD instance of the reuse contract. The transformation preserves raw classifier logits. Repeating the complete fitting and calibration procedure may nevertheless change the rejection decision. Agreement of complete decisions for every declared transformation and input defines a decision on presentation orbits.

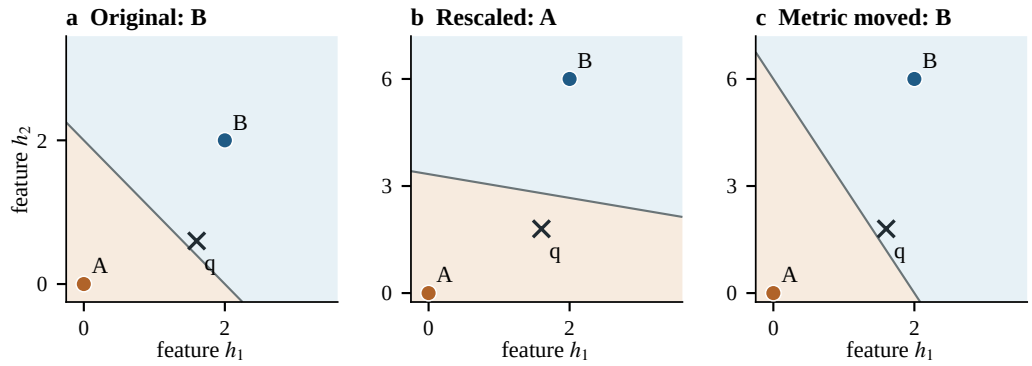


Figure 5: One query, three ways to measure distance. Left: the query is closer to center B. Middle: scale the second coordinate by three and recompute the centers; Euclidean distance now selects A. Right: give the second squared coordinate difference weight $1/9$; both original distances and the answer B return. The shaded regions show class labels. This is an exact illustration, not a dataset result.

B.2 Proof of normalization and concrete reductions

Proposition 5 (Preservation after reduction of symmetry)**.** *Let ν normalize a full interface on a G-stable domain. Suppose*

$$\nu(gp) = h(g, p) \cdot \nu(p), \qquad h(g, p) \in H. \tag{14}$$

If the complete consumer C is H-invariant on an H-stable domain containing these normalized interfaces, then $C \circ \nu$ is G-invariant. Two normalizations related by H at every presentation give the same complete decisions.

Proof of Proposition 5. The assumed domain makes both evaluations legal, and $C(\nu(gp)) = C(h(g, p)\nu(p)) = C(\nu(p))$. If $\widetilde{\nu}(p) = k_p\nu(p)$ with $k_p \in H$, the same equality proves choice independence. For the matrix construction, declare an ambient group acting on the full interface space, with G, H as subgroups and N_p in that ambient group. Then

$$h(k, gp)h(g, p) = N_{kgp}kN_{gp}^{-1}N_{gp}gN_p^{-1} = N_{kgp}kgN_p^{-1} = h(kg, p).$$

Neither normality of H nor a unique normalizer is required. Membership in H and legality of the normalized consumer must actually be checked. □

Write $g = A\Pi$ with A positive diagonal and Π a permutation matrix. For coordinate RMS, total standard deviations, or within-class standard deviations, let D_p be the corresponding positive diagonal scale matrix. Direct computation of the relevant second moments gives

$$D_{gp} = A\Pi D_p \Pi^\top, \qquad D_{gp}^{-1} g D_p = \Pi.$$

Thus the normalized coordinates differ by Π and the normalized compensated heads by its inverse. RMS uses uncentered moments; the other two choices use their specified centered covariances. This identity does not license substituting one statistic for another.

For covariance whitening, assume $\Sigma_p > 0$ and $\Sigma_{gp} = g\Sigma_p g^\top$. The residual map is

$$Q = \Sigma_{gp}^{-1/2} g \Sigma_p^{1/2}, \qquad QQ^\top = \Sigma_{gp}^{-1/2} (g\Sigma_p g^\top) \Sigma_{gp}^{-1/2} = I.$$

All square whiteners of Σ are $Q_0\Sigma^{-1/2}$ for orthogonal Q_0 (Kessy et al., 2018). Orthogonally invariant complete consumers therefore give the same behavior under these choices. This is not equality with the original unnormalized Euclidean consumer. Normalizing an interface to residual symmetry is established prior art; for example, Kornblith et al. (2019, Section 2.2) construct a general-linear invariant similarity from orthonormalized features and an orthogonally invariant statistic.

For ridge, let X contain centered row features and let Y contain centered targets. The normalized design is XD_0^{-1}. If its coefficient matrix is B_0, set $B = D_0^{-1}B_0$. Then

$$\|XD_0^{-1}B_0 - Y\|_F^2 = \|XB - Y\|_F^2, \qquad \|B_0\|_F^2 = \operatorname{tr}(B^\top D_0^2 B).$$

Taking $D_0^2 = \operatorname{diag} S > 0$ recovers the diagonal penalty. With the specified positive regularization grid the minimizer is unique, and the unpenalized intercept is recovered using the original centering convention. The normalized covariance has trace d, so its trace-scaled isotropic target is I.

For covariance shrinkage, $D_0^2 = \operatorname{diag}\Sigma > 0$ similarly gives the identity in Equation (4). The normalized covariance is positive semidefinite; every $0 < \lambda \le 1$ in the fixed grid makes its isotropic shrinkage positive definite. For monomial changes the fitted matrices obey $M' = gMg^\top$ and $R' = gRg^\top$, while means obey $\mu_c' = g\mu_c$. Distances and ridge candidate scores therefore agree. The common OOD or classification tail proves the complete guarantee for the specific task. The full definitions are given in Appendices B.3 and B.4.

B.3 Centroid and ridge consumers: objectives, returned fits and domains

The ridge calculation in Section 3 uses total centered covariance; the distance construction uses within-class covariance. Here we specify both families without identifying their statistics. Their experimental protocol is in Appendix D.8 and census in Appendix E.7.

Two fitted families. For labeled training features $(h_i, y_i)_{i=1}^n$, fit class means μ_c, overall mean $\bar{h}$, pooled within-class covariance $\Sigma = n^{-1}\sum_i (h_i - \mu_{y_i})(h_i - \mu_{y_i})^\top$, and total centered covariance $S = n^{-1}\sum_i (h_i - \bar{h})(h_i - \bar{h})^\top$. The centroid classifier minimizes $(h - \mu_c)^\top M^{-1}(h - \mu_c)$ over class IDs. We compare $M = I$ and the two shrinkage choices

$$M_{\mathrm{iso},\lambda} = (1-\lambda)\Sigma + \lambda \tfrac{\operatorname{tr}\Sigma}{d} I, \qquad M_{\mathrm{diag},\lambda} = (1-\lambda)\Sigma + \lambda \operatorname{diag}\Sigma.$$

The grid is $\lambda \in \{0.01, 0.1, 0.5, 1\}$ in that order. Scoring uses the equivalent linear discriminant $2h^\top M^{-1}\mu_c - \mu_c^\top M^{-1}\mu_c$, removing only the class-common quadratic term. No class-prior or log-determinant term is added.

For the linear probe let X_c and Y_c be centered feature and one-hot label arrays. The squared-loss fit minimizes $n^{-1}\|X_cB - Y_c\|_F^2 + \lambda \operatorname{tr}(B^\top RB)$, with an unpenalized intercept. It solves

$$(S + \lambda R)B = X_c^\top Y_c / n, \qquad \beta = \bar{y} - B^\top \bar{h}, \quad R_{\mathrm{iso}} = \tfrac{\operatorname{tr} S}{d} I, \quad R_{\mathrm{diag}} = \operatorname{diag} S.$$

The ordered ridge grid is $(10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1)$. This direct solve specifies the fitting objective without an SGD stopping-rule confound. The result is about these probes, not every algorithm called a linear probe.

Preservation predictions before evaluation. For $h' = gh$ with positive monomial g, $\mu'_c = g\mu_c$, $\Sigma' = g\Sigma g^\top$, and $S' = gSg^\top$. Because $\operatorname{diag}(gAg^\top) = g\operatorname{diag}(A)g^\top$ for such g, diagonal shrinkage satisfies $M' = gMg^\top$. Its inverse cancels the feature change in every candidate distance and discriminant. Likewise the diagonal ridge system gives $B' = g^{-\top}B$ and $\beta' = \beta$, preserving every candidate class score. With fixed labels and deterministic tie rules, candidate-wise equality preserves validation correct counts, selected λ, and final predictions. Isotropic targets satisfy the same law for $g = aP$ (uniform scale and permutation), but not for arbitrary nonuniform scaling. Euclidean distances multiply by a^2. Absence of a preservation guarantee does not require a disagreement on every dataset.

These arguments assume finite features, all classes represented in training, and a single fixed invertible map across all roles. Diagonal centroid requires every pooled within-class diagonal variance positive; diagonal ridge requires every total diagonal variance positive. Isotropic arms require the corresponding trace positive. These conditions with the stated positive grids give positive definite solve matrices; they are preserved by the declared transformations. No additive epsilon or pseudoinverse replaces an undefined candidate.

B.4 Distance scores and fitted covariance laws

Let μ_c be class means and Σ the pooled within-class covariance with weight $1/n$ per training row. The three Mahalanobis variants use

$$q_M(h) = -\min_c (h-\mu_c)^\top M^{-1}(h-\mu_c), \qquad M_{\text{full}} = \Sigma,$$

$$M_{\text{iso}} = (1-\lambda)\Sigma + \lambda\tfrac{\operatorname{tr}\Sigma}{d} I, \qquad M_{\text{diag}} = (1-\lambda)\Sigma + \lambda \operatorname{diag}\Sigma,$$

with ordered grid $\lambda \in (0.01, 0.1, 0.5, 1)$ for shrinkage. The quadratic term shared across classes at a given query is retained: dropping it preserves a class argmax but changes an OOD score. This is a final-layer Mahalanobis comparison, without the multi-layer combination or input preprocessing of the full detector of Lee et al. (2018).

For positive monomial g, $\mu'_c = g\mu_c$ and $\Sigma' = g\Sigma g^\top$. Full covariance and diagonal shrinkage satisfy $M' = gMg^\top$, so $(gh-g\mu_c)^\top (M')^{-1}(gh-g\mu_c) = (h-\mu_c)^\top M^{-1}(h-\mu_c)$. Full covariance requires positive definiteness; the positive shrinkage grid with all within-class diagonal variances positive ensures it for the diagonal variant. These exact domains are preserved by the action. Isotropic shrinkage and Euclidean distance respect uniform scaling and permutations but lack this general nonuniform-scale law. For normalized kNN, uniform scaling cancels in normalization; nonuniform scaling need not cancel. These are classical algebraic facts, not new covariance identities.

B.5 Where coordinate scale survives, cancels, or is removed

Distances and penalties are two instances. Feature-shaping detectors expose a third: they modify the feature vector before the compensated head. Following coordinate scale through these operations predicts both failures and preserving cases.

ASH: a recalibrated counterexample. ASH-B retains the k largest coordinates of h, replaces each retained value by $\sum_j h_j/k$, zeros the rest, and applies the original head (Djurisic et al., 2023). Nonuniform rescaling can alter the retained coordinates. The following minimal example shows why recalibrating the cutoff need not recover the original decision.

There is a tie-free, fully recalibrated two-coordinate decision counterexample. Take a two-class head with rows $W_1 = (1, 1)$ and $W_2 = (0, 0)$, zero bias, $g = \operatorname{diag}(1, 3)$, $k = 1$, calibration row $h_c = (6/5, 1)$, and target $h_t = (9/5, 7/10)$. Baseline ASH-B scores are $11/5$ and $5/2$, so the target is strictly above the recalibrated cutoff. After transformation the scores are $7/5$ and $13/10$, so it is strictly below. Raw calibration/target first logits remain $11/5$ and $5/2$; all top-k and target-cutoff comparisons are strict. Here and below "score" may denote the shaped first logit: the two-class log-sum-exp confidence $\log(1 + e^q)$ is strictly increasing in it. Analogous fully recalibrated decision counterexamples hold for ASH-P and ASH-S (Appendix B.6).

Here the keep count is fixed: rescaling changes the pre-head top-k support, and recalibration need not restore input ordering. Section 5 also repeats hyperparameter selection. The reversed calibra-

tion/query order rules out one common increasing recoding on these rows; the explicit unequal decisions establish failure, rather than merely the absence of a sufficient preservation certificate.

Uniform scaling and coordinate permutation preserve ASH support away from boundary ties and serve as specificity controls. Under nonuniform rescaling, ranking precedes the compensated head, so changed support need not be undone by inverse head scaling. Individual transformations and inputs can still agree.

DICE and Energy: scales cancel. DICE fits the training-feature means $m_j = \mathbb{E}_{\text{train}} h_j$ and ranks signed contributions $A_{cj} = W_{cj} m_j$ (Sun & Li, 2022). Write a monomial action as $h'_i = s_i h_{\gamma(i)}$, where $s_i > 0$ and γ permutes the coordinate indices. Then

$$m'_i = s_i m_{\gamma(i)}, \qquad W'_{ci} = W_{c,\gamma(i)}/s_i, \qquad A'_{ci} = A_{c,\gamma(i)}. \tag{15}$$

The feature-mean scale cancels its reciprocal head-weight scale. Thus the contribution mask permutes, masked logits agree, and so do the deterministic cutoff and decision when the training mean and mask are transported covariantly or recomputed from the transformed training role with the same deterministic mask rule. Energy depends only on the already invariant raw logits (Liu et al., 2020), so its decision is preserved immediately. These are stronger controls than generic numerical checks: the transformation laws give conditional preservation guarantees for DICE and Energy, while ASH admits a counterexample. For these fixed consumers the candidate set can be a singleton; if the same law is checked for each candidate of an ordered grid, Lemma 1 also handles reselection and calibration. The common tail does not require a new proof for each mean, mask, or logit construction.

RMS shaping: remove scale before ranking. RMS normalization removes coordinate scale before ASH ranking and restores it before the compensated head. Fit an ID-training RMS scale

$$r_j = \sqrt{\mathbb{E}_{\text{train}} h_j^2} > 0, \qquad z = h \oslash r, \qquad \widehat{h} = r \odot \text{ASH}(z). \tag{16}$$

Here $\oslash$ and $\odot$ denote elementwise division and multiplication. Under a positive monomial transformation, r' scales and permutes with h', so z' is only a permutation of z. ASH commutes with permutations on its unambiguous domain; reconstruction yields $\widehat{h}' = g\widehat{h}$, and the compensated head gives identical logits. This fitted RMS scale transforms covariantly with the feature coordinates as required by Equation (12). Zero-RMS coordinates are outside the strict domain; an arbitrary additive epsilon would break exact covariance. This is Proposition 5 with H the permutation group. Once permutation compatibility is established, every candidate's score equality feeds the same selection/calibration lemma as the preceding consumers.

Released top-k code can break permutation compatibility by choosing among coordinates tied at the boundary. Our *completed* rule shares mask weight equally among tied boundary coordinates; it agrees with hard ASH off ties and commutes with permutations. The RMS guarantee uses this completion, positive fitted RMS, and a nonzero retained sum for ASH-S. No additive epsilon is introduced. Appendix B.8 gives the mask, the ASH-S domain and the finite numerical certificate. Empirically, *strict-rule support* consists of rows whose required fitting and evaluation paths avoid boundary ties. *Conservative* disagreement counts unsupported rows as zero and keeps them in the denominator.

B.6 ASH FAMILY COUNTEREXAMPLES

Take $h = (2, 1)$, let the two head rows be $W_1 = (1, 1)$ and $W_2 = (0, 0)$ with zero bias, set $g = \text{diag}(1, 3)$, and retain one coordinate. Then $h' = (2, 3)$, $W'_1 = (1, 1/3)$, and both raw first logits equal 3. The shaped first logits are

$$W\,\text{ASH-P}(h) = 2, \qquad W'\,\text{ASH-P}(h') = 1, \tag{17}$$

$$W\,\text{ASH-B}(h) = 3, \qquad W'\,\text{ASH-B}(h') = 5/3, \tag{18}$$

$$W\,\text{ASH-S}(h) = 2e^{3/2}, \qquad W'\,\text{ASH-S}(h') = e^{5/3}. \tag{19}$$

All supports are strict, so no tie convention is involved. Appendix B.5 gives a fully recalibrated ASH-B decision flip. For ASH-P, calibration row $(3/2, 1)$ and target $(2, 1)$ similarly change a strict

above-calibration decision from accept to reject. For ASH-S, use calibration $(6/5, 1/2)$ and target $(11/10, 3/5)$. Before transformation their shaped first logits are $(6/5)e^{17/12}$ and $(11/10)e^{17/11}$, in increasing order; afterward they are $(1/2)e^{9/5}$ and $(3/5)e^{29/18}$, in decreasing order. Thus the strict calibrated decision flips. The second shaped logit is always zero, so the log-sum-exp confidence preserves each comparison.

B.7 DICE AND RMS PRESERVATION PROOFS

Equation (15) preserves the global multiset of signed contributions. The released mask compares every contribution strictly with a percentile of that multiset; the percentile is unchanged and the mask therefore permutes by γ, including when some contributions equal the threshold. Calling the baseline mask M and transported mask M', direct coordinate substitution gives

$$(W' \odot M')h' + b = (W \odot M)h + b.$$

Thus logits, log-sum-exp scores, deterministic ID cutoffs, and decisions agree. A stale mean next to W' is a different extended object and need not agree.

For Equation (16), write $g = DP_\gamma$ with D positive diagonal and P_γ the matrix of the coordinate permutation. Then $r' = DP_\gamma r$ and $z' = P_\gamma z$. Permutation equivariance of completed ASH yields

$$r' \odot \mathrm{ASH}(z') = DP_\gamma(r \odot \mathrm{ASH}(z)) = g\widehat{h}.$$

Multiplication by Wg^{-1} proves shaped-logit equality. Recomputing the same deterministic calibration functional then proves decision equality.

For completeness, automatic parameter selection (APS) also preserves this equality. Fix its ordered percentile set and first-maximum rule. At every candidate percentile, equality of shaped logits for every ID-calibration and OOD-validation row gives equality of the two score vectors and hence of validation AUROC. The entire finite AUROC vector is therefore equal, so deterministic first-maximum selection returns the same percentile. At that percentile, the ID-calibration score multisets agree, hence their deterministic empirical cutoffs agree; evaluation scores then agree rowwise, and the strict predicates $q > \tau$ agree. Thus the RMS preservation argument reaches the complete fit–select–calibrate–decide observable, not only its pre-calibration logits.

B.8 TIE COMPLETION

Let v be the kth largest value, r the number of coordinates strictly above v, and q the number equal to v. The barycentric mask has weight 1 above v, 0 below, and $(k-r)/q$ at each boundary coordinate. These memberships sum to k. Off ties, $q = k - r$ and every boundary membership is 1, recovering strict top-k. A permutation preserves inequalities and multiplicities and only permutes coordinates, proving $\bar{m}(Ph) = P\bar{m}(h)$. ASH-P/B follow immediately. For ASH-S, write the retained sum as $B_m(h) = \sum_j m_j h_j$. Every admissible hard mask has the same $B_m(h)$ because masks differ only among equal boundary values; on the domain $B_m(h) \neq 0$, the same barycenter argument applies to its exponential multiplier. The zero-retained-sum case is undefined and is not silently completed. In a corrected numerical implementation, the denominator and domain check use the common hard-mask sum $B = \sum_{h_j > v} h_j + (k-r)v$, computed in a canonical coordinate order with accurate summation before fractionalization. For example, $(1, -1, -1, -1)$ with $k = 2$ has $B = 0$ under every permutation, although summing its fractional barycenter in float64 can leave a nonzero residue. This auxiliary ASH-S correction leaves ASH-B and its frozen experimental implementation unchanged; no denominator epsilon is introduced.

Domains and reporting conventions used for RMS shaping. For ASH-S, its exponential multiplier additionally requires a nonzero retained sum B. We restrict this construction to presentations with every fitted $r_j > 0$ and, for ASH-S, $B \neq 0$ on every consumed row and invoked keep count. After RMS standardization the action is a permutation, so these conditions define a replete domain. The floating-point certificate further concerns only the finite tested endpoint/transformation set on which every transformed array and consumer output is finite; it is not an orbit-wide numerical guarantee for arbitrarily ill-conditioned scales.

Released top-k implementations do not specify a coordinate-free order on a boundary tie. Let $\mathcal{M}_k(h)$ be all admissible hard masks and use their average, or barycenter, $\bar{m} =$

$|\mathcal{M}_k(h)|^{-1} \sum_{m \in \mathcal{M}_k(h)} m$. Boundary coordinates receive equal fractional membership. This completion equals hard ASH off ties and commutes with permutations on ties. The RMS preservation proof uses this completion at boundary ties, on the positive-RMS domain with the additional nonzero-retained-sum condition for ASH-S. We report separately which empirical claims use the completion and which flips lie in the released rule's strict off-tie domain. We call results using the tie completion *completed*. For each comparison, *strict-rule support* consists of rows for which all required fitting and evaluation paths avoid boundary ties. *Conservative* disagreement statistics count unsupported rows as zero while retaining them in the denominator.

C MINIMAL INFORMATION, CHANGES OF OUTPUT AND CONTINUED FITTING

Section 4 asks when states that give the same current outputs can also support the next use. The proofs here establish which states may be grouped together without losing a permitted output, when operations and transports act on those groups, and the exact information required by the declared ranking and quadratic-fitting procedures. The final examples check the fixed-metric characterization and the class-label/OOD distinction.

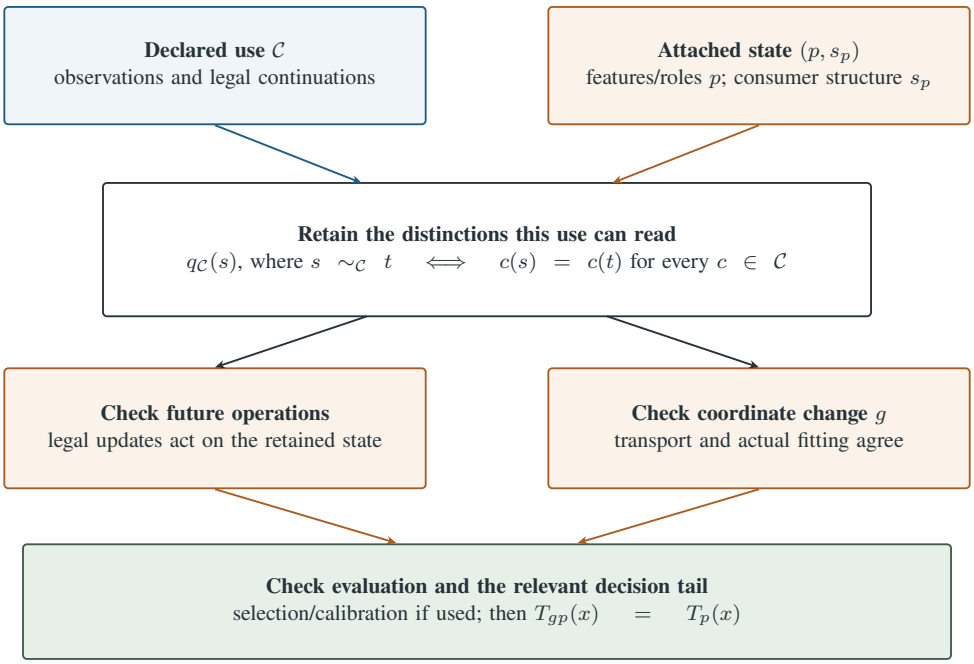


Figure 6: From a declared use to a reusable structured interface. A presentation p supplies features and data roles, while s_p supplies consumer structure such as a metric or penalty. We write $q_{\mathcal{C}}(s)$ for the state retained under the declared observations and continuations. Legal updates and coordinate transports must act on that state; actual fitting and evaluation must respect them before an applicable decision tail preserves the specified output. Table 1 gives different contract-specific summaries. Arrows indicate proof dependencies, not a new algorithm.

C.1 MINIMAL SUFFICIENT INFORMATION AND COMPATIBLE QUOTIENTS

The observations here are deterministic total maps on their declared typed domains. Their parameters, queries, input roles and any fixed random choices are included in the state or in the observation. For a partial module, require membership in its domain to be constant on each equivalence class, and equivalent outputs where defined. Alternatively, adjoin a distinct failure value and retain definedness as an observation. Reachable-state restrictions must be closed under the allowed operations.

Proof of Proposition 2. The relation $\sim_i = \bigcap_{c \in \mathcal{C}_i} \ker c$ is an equivalence relation. For another equivalence E, all observations factor through X_i/E exactly when $E \subseteq \sim_i$: equality of quotient classes must imply equality of each observation, and this condition also makes $\bar{c}([s]) = c(s)$ well-defined. Each factor is unique because the quotient projection is surjective.

If a summary $q : X_i \to Y$ admits decoders $c = d_c q$ for every c, then $\ker q \subseteq \sim_i$. The map $\rho : q(X_i) \to X_i/\sim_i$, $\rho(q(s)) = [s]$, is therefore well-defined, surjective and unique, with $\pi_i = \rho q$. Summaries with exactly this kernel are isomorphic on their attained images, compatibly with all readouts. This is the claimed information minimality.

If $s \sim_i t$ and $a : X_i \to X_j$ is allowed, prefix closure gives $c(a(s)) = (c \circ a)(s) = (c \circ a)(t) = c(a(t))$ for every $c \in \mathcal{C}_j$. Hence $\bar{a}([s]) = [a(s)]$ is well-defined. Surjectivity and evaluation on classes give uniqueness, $\overline{ba} = \bar{b}\bar{a}$ and preservation of identities. Conversely, if all modules descend and primitive terminal readouts factor, repeated substitution factors every generated terminal program. Thus, for the family generated by these modules and readouts, $\sim$ is the largest family of module congruences preserving the readouts.

For observation families $\mathcal{C}_\alpha$ on the same state space, $\sim_{\cup_\alpha \mathcal{C}_\alpha} = \bigcap_\alpha \sim_{\mathcal{C}_\alpha}$. If q_α realizes each constituent quotient, the joint quotient is the attained image of the tuple $(q_\alpha)_\alpha$, not necessarily its whole product. Enlarging the family yields a canonical map from its finer quotient to the old one. □

For a family $X_i(p)$ over interfaces, suppose a supplied transport $t_i(g) : X_i(p) \to X_i(gp)$ has the property that every target observation pulled back along it belongs to the source observation family. The same argument gives $\bar{t}_i(g)[s] = [t_i(g)s]$. Supplied identity and composition laws pass to these maps; inverses pass when the condition holds in both directions. The resulting set-valued indexed family has the usual Grothendieck construction (Vistoli, 2007).

One may therefore check fitting via $\pi_{gp} F(gp) = \bar{t}(g) \pi_p F(p)$, rather than requiring equality of the raw states. Likewise, if actual stage squares commute modulo the target equivalence, the induced quotient squares commute exactly, and Proposition 1 applies to them. The terminal readout still has to be in the declared family and obey its comparison convention. The construction neither supplies missing transports nor establishes compatibility of an arbitrary learner. Equality for one terminal program cannot establish prescribed intermediate naturality.

C.2 The exact score interface for variable ranking tails

A *weak order* records which scores are smaller, larger or tied, without retaining their magnitudes. Consider a finite candidate score table $S = (s_{ji})$ on a row-identity set I with at least four elements and a finite nonempty candidate set J. Here s_{ji} is candidate j's score on row i. Allow any nonempty ordered candidate subset and any disjoint, nonempty validation-ID, validation-OOD and calibration roles, plus a distinct query. Each tail uses first-maximum AUROC selection, an attained order-statistic cutoff and $q > \tau$; any valid cutoff index is allowed and is part of the tail. Role assignment is part of this algebraic contract; changing the tail does not refit the score producer. The ID/OOD roles may vary across these tails rather than being immutable labels of row identities. The empirical protocols instead fix their roles; their preservation uses Lemma 1, without needing the necessity direction of this stronger family-level characterization.

Proposition 6 (The complete interface for the declared ranking tails)**.** *Two score tables give the same final decision for every such tail if and only if, for every candidate, they have the same weak order, including ties, on all rows. Equivalently, their arrays of dense ranks agree, where*

$$r_{ji} = \big|\{a \in \{s_{j\ell} : \ell \in I\} : a < s_{ji}\}\big|.$$

Every summary sufficient for all these tails must determine this rank array.

Necessity follows by isolating one candidate and comparing any query with a singleton calibration row; two other rows make validation legal. Sufficiency is the common tail lemma. Thus many score producers can export the same ordinal interface and inherit one complete selection/calibration argument. Ranks as invariants of increasing recodings are classical (Lehmann & Scholz, 1992, p. 45); the characterization states exactly which declared tails can distinguish them. With immutable dataset labels, one fixed role partition, or one fixed candidate grid, this necessity need not hold:

an unused order difference may be invisible. Lemma 1 remains the sufficient result used for the empirical protocols. The proof and the fixed-role boundary follow.

Here I, J are finite, $|I| \geq 4$, and $J \neq \varnothing$. A context consists of an ordered list K of distinct candidates, disjoint nonempty row subsets V_+, V_-, L, an index $1 \leq k \leq |L|$ and a query q outside those subsets. Validation assigns the ID-positive role to V_+ and OOD role to V_-. These labels belong to the context; the theorem does not impose an immutable statistical ID/OOD label on each identity. It concerns a fixed score table and does not assert statistical validity for arbitrary role assignments.

Proof of Proposition 6. For necessity, fix candidate j and distinct rows i, ℓ. Use the singleton candidate list (j), calibration subset $\{\ell\}$ and query i. Two other rows supply singleton validation subsets, so AUROC is defined. The final decision is $\mathbf{1}[s_{ji} > s_{j\ell}]$. Equality for every such context preserves each ordered pair comparison. Testing the reverse pair also preserves equality: a tie means both strict comparisons are false.

For sufficiency, matching weak orders give a strictly increasing bijection between the attained values of each candidate in the two tables. Match the smallest distinct value to the smallest, and so on. This one map works for all rows of that candidate. Lemma 1 then preserves validation AUROCs, their first maximizer, the selected calibration order statistic and the final comparison for every context. Thus observing the selected candidate as well would leave the equivalence unchanged.

The dense-rank array records exactly the strict comparisons and tie blocks. It therefore realizes the quotient. By Proposition 2, every sufficient summary determines this array on its attained image. No increasing extension outside the finite attained values is needed. □

The following scores have the same interface before and after candidatewise recoding; columns are validation-ID, validation-OOD, calibration and query:

	original	recoded
A	$(3, 1, 2, 2)$	$(27, 1, 8, 8)$
B	$(4, 2, 1, 3)$	$(47, 27, 17, 37)$

Both candidate AUROCs are one in the displayed context. If A comes first, the query ties its cutoff and the strict decision is zero in both tables. Changing only A's original query score to 2.5 changes the weak order and makes the same context return one. Ties carry observable information.

The universal-role qualification matters. With immutable ID/OOD labels and ID-only calibration, two OOD scores above every ID score can exchange order without changing any validation ID/OOD comparison or permitted threshold decision. With one fixed candidate grid and role partition, an unselected candidate's calibration/query order can change without affecting the selected output. Interpolated quantiles or cross-candidate arithmetic require different contracts.

C.3 Quadratic fitting and a single data continuation

The design X is fixed throughout the current-readout comparison. Each observation supplies y and x, so $b = X^\top y$ remains available when replacing H by a summary. The state equivalences below concern the Hessian, not the entire stored state of a learner with its targets removed. Penalties are fixed as targets vary and during the zero-target append.

Lemma 2 (All linear decision halfspaces determine a positive ray)**.** *For positive definite H, H', equality $\mathbf{1}[x^\top H^{-1} b > 0] = \mathbf{1}[x^\top H'^{-1} b > 0]$ for every b, x holds if and only if $H' = cH$ for some $c > 0$.*

Proof. For each nonzero b, equality of strict positive halfspaces makes $H'^{-1}b$ a positive multiple of $H^{-1}b$. Thus $T = H'^{-1}H$ takes each nonzero vector to a positive multiple of itself. Writing $Te_i = c_i e_i$ and applying this fact to $e_i + e_j$ gives $c_i = c_j$. Consequently $T = c_0 I$ with $c_0 > 0$, so $H' = c_0^{-1} H$. The converse follows by positive rescaling. This is also the halfspace argument underlying the fixed-metric characterization in Appendix C.5. □

Proof of Proposition 3. Full column rank makes $A = X^\top X$ positive definite and $X^\top y$ arbitrary: $y = XA^{-1}b$ realizes any b. The unique minimizers under the displayed objective are $(A+R)^{-1}b$ and $(A+S)^{-1}b$. Lemma 2 gives the current-label criterion.

Appending $(v, 0)$ gives $\widetilde{X}^\top \widetilde{X} = A + vv^\top$ and $\widetilde{X}^\top \widetilde{y} = b$. If both contexts agree for all targets and queries, put $H = A + R$, $H' = A + S$. The lemma gives positive c, k with $H' = cH$ and $H' + vv^\top = k(H + vv^\top)$. Hence

$$(c-k)H + (1-k)vv^\top = 0.$$

As H has rank $d \geq 2$ and $vv^\top$ has rank one, they are linearly independent. Thus $c = k = 1$, and $S = R$. Equality of the penalties conversely makes both unique fits identical. The row v was any prescribed nonzero vector. □

The two contexts already separate every distinct positive definite Hessian. Adding all repetitions of the update therefore leaves the identity relation as the contextual quotient. This does not assume that a family containing only the identity and one update is itself closed under arbitrary repetition.

For the main-text example, the initial Hessians are $2I_2$ and $3I_2$ with $b = (1, 1)^\top$. Appending $(e_2, 0)$ gives $\mathrm{diag}(2, 3)$ and $\mathrm{diag}(3, 4)$. For query $(5, -7)$ the initial scores are $-1, -2/3$, and the updated scores are $1/6, -1/12$. Initial labels agree on all queries because the initial coefficients are positive multiples, not merely because this particular query agrees. This example uses the half-scaled penalty in the displayed objective; the earlier ridge example in Appendix C.6 writes an unhalved penalty.

Under a feature change g, a target penalty R' pulls back to $S = g^{-1}R'g^{-\top}$, since

$$(gx)^\top (gAg^\top + R')^{-1} gb = x^\top (A+S)^{-1} b.$$

The corresponding appended row is gv. Therefore all-target, all-query preservation before and after this append is equivalent to $R' = gRg^\top$. In the example, $g = I/\sqrt{2}$ and the numerical penalty I at both endpoints pull back to $R = I$, $S = 2I$. The same fixed-penalty refitting recipe therefore agrees on every current query under this representation change and disagrees on the exhibited query after the corresponding zero-target sample is appended. For a single design with no append, the exact condition is only $A + S = c(A + R)$. Requiring preservation for all full-rank designs also forces $S = R$: the family includes both A and $A + vv^\top$.

The quantifiers cannot be removed. For one fixed right-hand side $b = e_1$, $H = 2I_2$, $H' = 3I_2$ and $v = e_2$, labels agree on all queries in both contexts although $H \neq H'$. In dimension one, every positive Hessian and every positive addition produce the same zero-threshold label for each b, x. Penalty updates, mean-loss conventions, automatic regularization rescaling, restricted target families, fitted intercepts and hyperparameter selection need their own analysis. The theorem is an exact quadratic instance of the common information criterion; it does not add an empirical result or a new estimator.

C.4 A RECURSIVELY USABLE INTERFACE FOR RESTRICTED CONTINUATION

The preceding proposition compares binary labels across all targets for one fixed design, retaining the target-dependent b. Here the observations and admissible operations differ. Let an initial state be (H, b) with $H \succ 0$, $P = H^{-1}$, and $w = Pb$. Fix a known full-column-rank $U \in \mathbb{R}^{d\times r}$, $1 \leq r \leq d$. At every stage, the observer may read $x^\top w$ for any $x \in \mathbb{R}^d$. A legal next sample is (Ua, y) with arbitrary $a \in \mathbb{R}^r$ and $y \in \mathbb{R}$; it changes (H, b) to $(H + Uaa^\top U^\top, b + Uay)$. Any finite sequence, including the empty one, is permitted. Initial states may have different b and different histories.

Proposition 7 (Restricted continuation interface)**.** *Two states give the same scores under every such observation and continuation if and only if both w and $K = PU$ agree. The pair (w, K) is closed under all legal appends. This is information minimality for the stated observations, not a minimum number of bits among arbitrary encodings.*

Proof. Write $G = U^\top K$ and $D = 1 + a^\top Ga > 0$. The rank-one inverse update for recursive least squares (Boyd, 2008) gives

$$w^+ = w + \frac{Ka}{D}(y - a^\top U^\top w), \qquad K^+ = K - \frac{Ka\, a^\top G}{D}. \tag{20}$$

Hence equal (w, K) remain equal after each common append and give equal scores for every query. Conversely, empty continuations and all x identify w. For each column u_j of U, compare the one-step coefficients with the same row u_j and alternative permitted targets $y = 1$ and $y = 0$. Their difference, identified by all score queries, is $g_j = Pu_j/(1 + u_j^\top Pu_j)$. Since $u_j^\top g_j < 1$, it recovers $Pu_j = g_j/(1 - u_j^\top g_j)$. Thus the observations identify every column of K. □

This calculation specializes known projected-covariance filtering: a fixed observation operator can be updated through its cross-covariance without the full covariance (Li et al., 2014, Section 3.1), and the action of a positive definite matrix on a chosen subspace has a known canonical rank-r representative (Lai & Bernstein, 2024, Definition 5 and Theorem 2). The exact statement above organizes these ingredients by the allowed score and update contract; it is not a new RLS algorithm. The restriction matters. For $U = e_1$, $w = 0$ and $P_1 = I_2$, $P_2 = \mathrm{diag}(1, 2)$, both interfaces agree. If an e_2 row becomes legal, appending $(e_2, 1)$ gives e_2 scores $1/2$ and $2/3$. Both initial states can arise from distinct ridge histories with the same penalty $I_2/4$. The old interface cannot in general supply this new response.

A predeclared numerical check compared the restricted interface with full RLS and batch solves. In 30 synthetic cells (990 state times), its largest coefficient discrepancy from batch fitting was 1.6×10^{-13}. On reused 512-dimensional CNN features, 64 added feature rows were deliberately projected into a fixed rank-16 U; over 180 held-out queries, the largest restricted-versus-full score discrepancy was 7.1×10^{-14}. No test label changed from the initial to final fit (accuracy $179/180$ at both times), although scores changed. These checks support execution under the chosen operation, not an observed fixed-label continuation split, natural incidence of the restricted subspace, or end-to-end efficiency.

Fixed geometry: labels retain a metric ray. For fixed positive definite K, K' and invertible g, every two-centroid label on every prototype/query configuration is preserved exactly when

$$g^\top K' g = cK \quad \text{for some } c > 0. \tag{21}$$

The oriented halfspaces for centers $\pm v$ determine the positive ray $[K] = \{cK : c > 0\}$: the matrix is determined up to a positive scale, with the same class priority on ties (Appendix C.5). For Euclidean centroids, the universally preserving linear changes are exactly similarities. This fixes the metric pair while varying all prototypes and queries; it is not a necessity theorem for arbitrary data-dependent metric estimators.

C.5 The metric ray identified by every two-centroid label

Proposition 8 (Classical fixed-metric characterization)**.** *Fix symmetric positive definite K, K' on $\mathbb{R}^d$ and an invertible g. For every pair of labeled prototypes and every query, the nearest-prototype labels using K and using K' after transporting prototypes and query by g agree if and only if $g^\top K' g = cK$ for one $c > 0$. The same fixed label priority is used on ties.*

Proof. Let $L = g^\top K' g$. Sufficiency follows because both squared distances multiply by $c > 0$. For necessity, use prototypes $-v$ and v, $v \neq 0$. Their distance difference is $4h^\top Kv$, and under L it is $4h^\top Lv$. If Lv is not a positive multiple of Kv, some query gives opposite strict signs: write $Lv = tKv + u$ with $u \perp Kv$; if $u \neq 0$, then $h = Kv - au$ for sufficiently large $a > 0$ has positive inner product with Kv and negative inner product with Lv. Hence $Lv = t_v Kv$, with $t_v > 0$ by positive definiteness. Thus every nonzero vector is an eigenvector of $K^{-1}L$. Applying this to basis vectors and their pairwise sums forces all eigenvalues to coincide, so $L = cK$. Necessity used strict comparisons, and the resulting matrix relation also preserves all ties. □

The equivalence class of fixed metrics inducing this entire label family is $[K] = \{cK : c > 0\}$. With $K = K' = I$, the permitted invertible linear maps are exactly $g = \sqrt{c}\,Q$, Q orthogonal. This quantifies over all prototype/query configurations. For data-dependent fitted K_D, K'_{gD}, the analogous matrix relation is sufficient on each dataset, but agreement at only its actual prototypes need not imply that relation. No universal necessity for arbitrary metric-fitting algorithms follows.

C.6 EXACT FITTING AND A CHANGE OF OBSERVABLE

Changing the requested output can make previously dispensable information necessary. Take one-dimensional class centers -1 and 1. Class prediction compares $-(h+1)^2$ and $-(h-1)^2$. Removing their common term $-h^2$ leaves every class label unchanged. For OOD rejection, however, the score is the larger negative squared distance, $q(h) = -h^2+2|h|-1$. The reduced score is $\widetilde{q}(h) = 2|h|-1$. Calibrate on $h = 1$ and query $h = 3$, accepting iff the query exceeds the cutoff:

	calibration score	query score	decision
full distance	0	-4	reject
class-equivalent reduced score	1	5	accept.

Both class predictions still agree. Both cutoffs have been recalibrated. The OOD observer reads information that the class argmax could ignore; the distance study therefore keeps the query's quadratic term. The discriminant calculation below checks the same conclusion. A summary adequate for one output cannot automatically be passed to another consumer.

For the ridge example, use rows $e_1, -e_1, e_2, -e_2$ of X, targets $(1, -1, 1, -1)$, and no intercept. Minimize $\frac{1}{2}\|Xw - y\|^2 + w^\top Rw$ exactly. Initially $R = I$, $X^\top X = 2I$, $X^\top y = (2, 2)^\top$, so the unique solution is $w = (1/2, 1/2)$. For query $(1, -3/4)$ the score is $1/8$. Under $g = \mathrm{diag}(1, 3)$, the normal equation with unchanged $R = I$ has matrix $\mathrm{diag}(4, 20)$ and right-hand side $(2, 6)^\top$. Thus $w' = (1/2, 3/10)$ and the corresponding query $(1, -9/4)$ has score $-7/40$. The strict zero-threshold label flips. Using $R' = \mathrm{diag}(1, 9)$ instead gives $w' = (1/2, 1/6)$ and restores $1/8$. This illustrative fixed-penalty binary problem is distinct from the experimental multi-class trace-scaled isotropic probe and its hyperparameter selection.

For an observer change, take one-dimensional centers -1 and 1 and unit distance matrix. The class discriminants are $-2h - 1$ and $2h - 1$. Removing the shared term $-h^2$ from negative squared distances leaves every class argmax unchanged. OOD confidence is different: the true score is $q(h) = -h^2 + 2|h| - 1$, while the maximum reduced discriminant is $\widetilde{q}(h) = 2|h| - 1$. Calibrate on $h = 1$ and query $h = 3$. The true cutoff/query scores are $0, -4$, so the query is rejected; the reduced scores are $1, 5$, so it is accepted. Both classifier labels agree and all query-cutoff comparisons are strict. Therefore a quotient adequate for class labels cannot be passed to the OOD tail without proving that the latter respects it. The distance-based OOD protocol in Appendix D.10 retains the query quadratic term for this reason.

D EXPERIMENTAL CONTRACTS AND REPRODUCIBLE PROTOCOLS

Section 5 compares structural predictions with complete fitted outputs. This appendix specifies the data roles, candidate grids, frozen criteria and validation boundaries for each study. Original confirmation, reused-population development, live-graph checks and prospective ranking cohorts retain their separate identities; the shared theory does not pool their evidence populations.

D.1 ORIGINAL OOD MODELS, DATA AND COMPLETE CONSUMER

To test these predictions, the CIFAR-100 study uses three public OpenOOD v1.5 CIFAR-100 ResNet-18 best checkpoints and one public CIFAR-100 Vision Transformer (ViT)-base checkpoint. OpenOOD provides a standardized OOD evaluation framework and public checkpoint population (Yang et al., 2022). The ViT use is explicitly a classifier-interface extension of ASH, not a released ASH result. At any architecture, Equation (12) can be realized by inserting a diagonal adapter before the linear head and its inverse into the head. For suitable homogeneous ReLU interfaces, positive scaling is also a familiar network parameter symmetry (Grigsby et al., 2023; Tran et al., 2024); for the ViT we claim only exact interface equivalence.

Fitted RMS scales and DICE means use all 50,000 CIFAR-100 training rows. A disjoint 1,000/9,000 split of the official test set supplies ID calibration/evaluation. ASH's automatic parameter selection (APS) chooses the first percentile ρ maximizing validation area under the receiver-operating-characteristic curve (AUROC) in $\{65, 70, 75, 80, 85, 90, 95\}$ using 1,000 ID-calibration and 972 held-out Tiny ImageNet OOD-validation rows. Evaluation uses 5,000 rows from each of CIFAR-10 and disjoint Tiny ImageNet (near OOD), and MNIST, SVHN, Describable Textures (DTD), and Places365 (far OOD) (Krizhevsky, 2009; LeCun et al., 1998; Netzer et al., 2011; Cimpoi et al., 2014;

Zhou et al., 2018). No evaluation identity appears in fitting. Tables abbreviate Tiny ImageNet as TIN and Places365 as Places.

For every evaluated interface, the primary ASH-B consumer reruns APS, scores with log-sum-exp, sets its cutoff to the predeclared empirical 5th percentile of ID-calibration confidence, and accepts as ID iff confidence is strictly greater. The secondary holds the baseline APS percentile fixed while recalibrating the cutoff. DICE uses a global signed- contribution 90th-percentile mask. The predeclared non-vacuity criterion required the ID-evaluation rows to contain both Boolean outcomes, aggregate OOD acceptance to lie strictly between 1% and 99%, and at least three OOD sets to contain both outcomes for every model; all four passed, excluding trivial near-constant agreement.

Detection quality is measured by AUROC and the false-positive rate (FPR), the fraction of OOD inputs accepted at the fitted cutoff. The label FPR-at-ID95 refers to the 95% ID-acceptance target used for calibration; ID-test acceptance can differ.

D.2 Transformations and comparison criteria

Let κ denote the ratio of the largest to the smallest positive diagonal scale. For each of eight fixed hash-derived seeds, half the feature coordinates receive $\sqrt{\kappa}$ and half receive $1/\sqrt{\kappa}$, with $\kappa \in \{2, 4, 8, 16, 32, 64\}$. The head is divided coordinatewise by the same scales. This yields $4 \times 8 \times 6 = 192$ exact diagonal transformations and $192 \times 6 = 1{,}152$ model/seed/condition/OOD cells. A *cell* is one model, transformation, and OOD-dataset combination. The paired outcome is the fraction of 5,000 fixed OOD rows whose Boolean decision differs from the baseline interface: the finite $T_x(p)$ versus $T_x(g{\cdot}p)$ comparison, excluding score-only changes. The finite seed/condition set is the estimand; there is no claimed natural or uniform distribution on the noncompact group.

The criteria were specified before evaluation. They require disagreement across every model and both near/far OOD strata, agreement of raw logits within 2×10^{-10} and identical class argmax, and zero decision flips for RMS-ASH, DICE, and Energy at $\kappa = 64$. Identity, uniform scaling, coordinate reversal, and cyclic permutation are the specificity controls. Utility is evaluated separately by comparing RMS-ASH with raw ASH on the baseline interface. Appendix D.7 gives the full conjunctions, tolerances, and bootstrap scope.

A subsequent protocol fixed one formula-derived transformation per model before evaluating its decisions. With $a_j = \max_i |H_{ij}^{\mathrm{IDcal}}|$ and $w_j = \max_c |W_{cj}|$, the SmoothQuant-formula scale is $s_j = \max(a_j^{1/2} / \max(w_j, 10^{-5})^{1/2}, 10^{-5})$ and the tested transformation is $g_j = 1/s_j$. We fixed $\alpha = 0.5$, the ID-calibration role, no normalization, all four models and six OOD roles. The experiment applies this interface transformation without quantizing or exporting the model; its chronology is retained in Appendix D.7.

D.3 Data roles and decisions

Table 3: Disjoint frozen roles. Hash-ranked sampling is independent of pixels, labels, model outputs, and consumer outcomes.

Role	Source	Rows
RMS/DICE fit	CIFAR-100 train	50,000
ID calibration/APS	CIFAR-100 held-out test split	1,000
ID evaluation	disjoint CIFAR-100 test split	9,000
OOD validation	Tiny ImageNet validation	972
near OOD evaluation	CIFAR-10; disjoint Tiny ImageNet	5,000 each
far OOD evaluation	MNIST; SVHN; DTD; Places365	5,000 each

Network inference emits float32 features/logits; heads, transformations, downstream consumers, calibration, and metrics use float64. The cutoff is $\mathrm{sort}(q_{\mathrm{IDcal}})[\mathrm{round}(0.05n)]$ and the strict decision is $q > \tau$. APS uses first-maximum tie breaking. ASH keep count is $k = d - \mathrm{round}(d\rho/100)$ using Python/NumPy round-to-even.

Table 4: Primary strength and utility criteria. All conjunctions and outcome labels were fixed before the corresponding outcome evaluation and were unchanged by the tie amendment.

Quantity	Threshold	Result
Each model median at $\kappa = 64$	$\geq 1\%$	11.24–22.33%
Near/far pooled median at $\kappa = 64$	each $\geq 2\%$	19.53/21.87%
Fixed-ρ pooled median at $\kappa = 64$	$\geq 1\%$	22.51%
Full-refit pooled median at $\kappa = 32$	$\geq 0.5\%$	25.09%
Mean RMS-minus-ASH AUROC	≥ -0.01	-0.00933
Mean RMS-minus-ASH FPR	$\leq +0.03$	-0.00759
Maximum cell AUROC loss	≤ 0.10	0.13282 (fail)
Maximum cell FPR increase	≤ 0.10	0.03200

D.4 Frozen outcome rules

Under the prespecified conjunction, exact-logit validity and the ASH-B disagreement criteria pass, while the RMS-ASH utility criterion fails.

D.5 Predeclared rule for the formula-derived transformations

The formula-derived follow-up fixed $\alpha = 0.5$, a 10^{-5} lower clamp, the 1,000-row ID-calibration role, and no normalization. One transformation was selected for each model. Its strongest predeclared outcome required exact logits and all three controls, a strict-rule full-APS flip in every model, a 24-cell median at least 1%, and separate near/far medians at least 0.5%. No outcome could be rescued by a new alpha, split, normalization, model subset, or threshold.

D.6 CIFAR-10 confirmation protocol

A separately predeclared follow-up changes the ID dataset and checkpoints to three public CIFAR-10 ResNet-18s. It retains ASH-B, the six-role near/far taxonomy, 5,000 rows per role, condition 64, and the completed and conservative released-rule criteria. Its eight balanced transformation seeds are disjoint from the chronology-invalid development attempt; no cell from that attempt is pooled. Freshness concerns the transformation–outcome unit: models, samples, and feature caches are reused. The 144-cell study was evaluated once, retaining rows outside strict-rule support as zero.

D.7 Additional protocol details

At $\kappa = 64$, the predeclared systematic criterion required every model's median over 48 seed/OOD cells to exceed 1%, separate near/far pooled medians to exceed 2%, and the fixed-percentile median to exceed 1%. The $\kappa = 32$ full-refit median had to exceed 0.5%. Every endpoint also had to preserve checked raw logits within 2×10^{-10} absolute error and exact class argmax. At $\kappa = 64$, RMS-ASH, transported DICE, and Energy had to exhibit zero decision flips and pass tighter consumer-logit/confidence tolerances. Identity, uniform scaling, coordinate reversal, and cyclic permutation were predeclared specificity controls.

The utility criterion compares RMS-ASH with raw ASH on the baseline interface. It requires mean AUROC change at least -0.01, mean FPR-at-ID95 change at most $+0.03$, and no model/OOD cell worse by more than 0.10 AUROC or 0.10 FPR. Two thousand paired bootstrap replicates describe fixed-population mean changes but cannot override this conjunction.

After the original study was sealed, a separate protocol fixed one formula-derived transformation per model before opening those endpoints. With $a_j = \max_i |H_{ij}^{\mathrm{IDcal}}|$ and $w_j = \max_c |W_{cj}|$, the SmoothQuant-formula scale is $s_j = \max(a_j^{1/2} / \max(w_j, 10^{-5})^{1/2}, 10^{-5})$ and the tested transformation is $g_j = 1/s_j$. We fixed $\alpha = 0.5$, the ID-calibration role, no normalization, all four models and six OOD roles, and an ordered fail-closed outcome rule. This is a formula-derived interface test, not integer quantization or exporter replay.

Outcome boundary. Each outcome evaluation admitted as confirmatory evidence followed its frozen protocol; the chronology-invalid CIFAR-10 preliminary attempt is excluded. A strict-domain stop led only to the permutation-equivariant completion, and the matched-horizon protocol was run once without post-result changes. Its later coverage correction changed one sentence, not any datum, criterion result, or terminal outcome. Appendix F.6 explains the relevant exclusions and corrections.

Stored-decision validation details. The strict rule is valid for 779,991 of 960,000 row instances in 156 of 192 cells. The remaining 36 cells have a boundary tie on at least one endpoint fitting path and are retained as zero, not excluded, in every lower-bound median (six ResNet s0, none ResNet s1, twelve ResNet s2, and eighteen ViT cells). All eight joint leave-one-pair-index-out checks preserve both criteria. A separate implementation reconstructs all 13 Boolean vectors per cell, every row pattern, rate, aggregate, cancellation branch, and criterion result. It starts from serialized decisions and tie records, not an independent feature-to-score computation.

D.8 Classification study: roles, selection and numerical validation

This extension changes the terminal task from OOD rejection to CIFAR-100 class prediction. It uses the same four feature interfaces and existing 50,000/1,000/9,000 train/selection/test roles. The consumer definitions, candidate orders, transformations, analysis, and numerical checks were fixed before computing these new outputs. The models, samples, and mild arrows were previously used for OOD research; this is a development extension on reused data, not fresh-population confirmation. Nearest-centroid and regularized least-squares classification are classical (Hastie et al., 2009). Their covariance properties here provide controlled comparisons, rather than new classification methods.

Complete refitting, controls, and numerical validation. Every endpoint independently recomputes all training statistics and candidates. Choose the first candidate maximizing integer correct count on the 1,000 selection rows; class ties choose the lowest class ID. The test set never selects λ. For each model compare its baseline to the previously specified mild formula arrow and one balanced $\kappa = 2$ arrow (seed 20260920). Identity rerun, uniform scaling by 2, and a one-position cyclic permutation are specificity controls. The original head is inversely compensated and checked on all 60,000 consumed rows. Complete transport evaluates the selected baseline consumer on $g^{-1}h'$ and retains that reference and the known map.

Computation uses float64 on the retained float32 features. Raw-logit tolerance is 2×10^{-10} with exact argmax agreement. Cholesky solves require relative backward residual at most 10^{-10}. Candidate score comparisons use $10^{-8}(1 + \max|s_{\text{baseline}}|)$, and report rows with top-two margin at most twice this tolerance. The preserving comparisons additionally require identical candidate selection-row predictions, correct-count vectors, selected indices, and selected test predictions. Numerical closeness alone does not guarantee selection agreement. Synthetic direct-distance and augmented least-squares references check the scoring kernels independently of these preservation tests. An additional implementation recomputes statistics, every selection-row candidate, selection, and selected test outputs for baseline/mild/$\kappa = 2$ using direct centering and LU solves. All 540,000 selected test predictions agree; maximum selected score discrepancy is below 1.28×10^{-11}. This starts at cached features and does not repeat encoder training, control-endpoint fitting, or the bootstrap analysis.

Reported scope and uncertainty. All four models and five consumers are retained. We report class disagreement, baseline and endpoint accuracy, paired accuracy change, and both directions of correctness change. A baseline selection accuracy above 10% is the predeclared task-competence check (chance is 1%). The pointwise descriptive 95% intervals use 2,000 paired resamples of the 9,000 test-image identities, with one shared draw across all models, methods, and transformations. They condition on the fixed fits and selections and are not simultaneous confirmatory intervals. The complete finite census appears in Appendix E.7. EC/CI/CD denote Euclidean/isotropic/diagonal centroid; RI/RD denote isotropic/diagonal squared-loss probes. The machine-readable record also retains all candidate scores, margins, selected parameters, per-class accuracy, and bootstrap replicates.

D.9 DISTINCT EVIDENCE IDENTITIES OF THE LATER STUDIES

The distance-based OOD and classification extensions, the live-graph replay, and the prospective ranking cohorts have separate protocols and outcome records. They extend the original OOD study without turning reused-population development into independent confirmation. Each asks whether a specified representation change preserves a complete fitted output, but the tasks, populations and validation boundaries differ.

D.10 DISTANCE STUDY: ROLES, SELECTION AND NUMERICAL VALIDATION

The distance study uses the four existing CIFAR-100 models, 50,000 training rows, 1,000 ID-calibration rows, 9,000 ID-test rows, 972 OOD-validation rows, and six 5,000-row OOD evaluation sets. Each model has its baseline, the fixed $\alpha = 0.5$ formula arrow, and a balanced $\kappa = 2$ arrow with seed 20260921. Every endpoint independently fits its statistics and candidates. Exact kNN search uses the ordered grid $(1, 10, 50, 100, 500, 1000)$ and scores by negative squared distance to the kth neighbor in the training bank. The normalized variant divides each bank and query row by its own L2 norm; zero norm is outside its domain. The raw variant leaves features unnormalized. These are specified adaptations of nearest-neighbor OOD scoring (Sun et al., 2022) to the common protocol.

The covariance score definitions and positive-domain proof are in Appendix B.4. These are the same candidates used by the complete worked example in Section 3.1.

Select the first candidate maximizing ID-positive validation AUROC, using the fixed ID calibration and OOD validation roles. The sole full-covariance candidate needs no selection. Refit the order-statistic 5th-percentile ID cutoff and accept iff $q > \tau$. Candidate-wise score preservation therefore reaches selection, calibration, and final decisions under the fixed deterministic rules. Known-map transport separately applies g^{-1} to queries and reuses the baseline fitted bank/metric, selected parameter, and cutoff. It retains reference information unavailable to independent endpoint fitting.

All kNN endpoints select $k = 1$. Shrinkage selects $\lambda = 1$ for the CNNs and 0.01 for ViT at both transformed endpoints as well as baseline. Thus running reselection does not demonstrate a selection change here. Full covariance rejects numerical fits with $\lambda_{\min}(\Sigma) \leq 10^{-12}\lambda_{\max}(\Sigma)$; the ViT fails at all three endpoints. This finite numerical gate is not an orbit-wide rank theorem. We add neither a pseudoinverse nor an epsilon, and retain the missing comparisons explicitly.

Maximum original-head logit error is 1.78×10^{-14} with zero class changes. All valid full transports have zero decision mismatches (maximum score error 1.82×10^{-11}). Independent post-result validation checks 57 defined fits, three undefined fits, 342 OOD output cells, 144 GPU/CPU kNN reference scores, and 1,556,936 transported rows. Preserving variants' largest selected evaluation-score error is 8.73×10^{-11}; no checked score lies within the declared tolerance of its baseline cutoff. These checks start at cached features, not new trained models.

Protocol departures remain visible. The complete kNN identity/uniform/permutation synthetic check was added after real outcomes, although planned before them; it is post-result validation. Some earlier synthetic distance and Mahalanobis permutation checks preceded outcomes. Per-row identities remain with the input feature records rather than being duplicated in each output file. A post-result tie audit found 1,088 exact neighbor-boundary ties across repeated roles, candidates, and endpoints, 146 at selected $k = 1$. Neighbor identity at an equal distance cannot change the kth-distance score, and the audit reproduces saved scores. No evaluation score equals its cutoff.

D.11 MANUAL GRAPH EQUALIZATION AND RAW-IMAGE REPLAY

The live-graph check fixes ResNet-18 s0 and ViT, each with 1,000 ID-calibration, 972 OOD-validation, 1,000 ID-test, 1,000 CIFAR-10, and 1,000 Texture images already exposed in the original study. A float32 multiplication before the head and inverse weight compensation implement the formula arrows ($\kappa = 1.6474$ and 1.5561). Original replayed features match the retained cache exactly. Baseline graph/cache logits differ by at most 5.72×10^{-6}; compensated/original graph logits by at most 3.82×10^{-6}. No original classifier argmax changes. ID-test accuracies are 77.4% and 92.3%. Graph capture retains the multiplication and agrees with eager execution on its fixture; the 4,972-image comparisons concern the live forward runs.

Table 5: ASH-minus-Energy AUROC gap intervals in percentage points: 98.75% marginal paired-bootstrap percentile intervals, Bonferroni-adjusted for four gaps (nominal 95% family). They condition on the fitted procedures.

OOD source	Endpoint	Interval
Tiny ImageNet test	Baseline	[-8.63, -5.71]
Tiny ImageNet test	Transformed	[-8.28, -5.38]
EuroSAT	Baseline	[+10.13, +13.47]
EuroSAT	Transformed	[+10.13, +13.50]

ASH uses the original ordered percentile grid and strict cutoff, independently selected and calibrated at each graph endpoint. Selected percentiles are $95 \to 95$ for the CNN and $90 \to 75$ for ViT. ID decisions change on 22/1,000 and 48/1,000 images, respectively; the OOD results are in Table 16. Energy has zero changes. Selected ASH fitting/evaluation paths have no top-k boundary ties. A separate reconstruction checks ASH scores, Energy via log-sum-exp, rank-based AUROC, selection, and thresholds. Graph/cache decisions agree on all three test roles. Eight ViT OOD flips lie within 10^{-5} of a cutoff, none within 10^{-6}; corresponding graph/cache score residuals are below 8×10^{-10}. All fixed-grid candidate arrays were reconstructed and retained after the result, without changing selection.

The first ID-only smoke failed because of a TF32 runtime mismatch; the second attempt failed on a CPU/GPU tensor mismatch before OOD evaluation. Both failures precede OOD outcome access. The successful third attempt exceeds the original one-retry allowance. The arrow was computed from the pinned calibration cache, rather than newly replayed activations as planned; every replayed calibration feature is subsequently verified bit-for-bit equal. The successful scientific execution took 18.34 seconds. Peak memory was not instrumented, so the memory cap cannot be certified. These disclosed apparatus repairs and reused images preclude calling this independent population confirmation.

D.12 Prospective ranking tests and outcome access

First cohort. The frozen encoder is an ImageNet-pretrained ResNet-18. A centered ridge head is fitted on 45,000 CIFAR-10 training images and selected on the remaining 5,000. Its selected $\lambda = 1000$ is at the grid boundary. CIFAR-10.1 v6 supplies 400 class-balanced calibration and 1,600 evaluation images; their pixel hashes overlap with zero original CIFAR-10 train/test images. Tiny ImageNet test supplies 1,000 OOD-validation and 5,000 evaluation images; EuroSAT supplies 5,000 evaluation images. OOD selected paths are disjoint from prior project caches. This does not certify perceptual deduplication or disjointness from model pretraining. Before scoring, an exposed Places365 source was replaced by EuroSAT under a retained amendment. CIFAR-10 training identities are reused.

Training-internal validation accuracy is 81.82%; new ID calibration/evaluation accuracy is 67.75%/67.69%. The frozen calibration gate is 65%. The fixed formula arrow realizes $\kappa = 6.67$; maximum head-logit error is 4.44×10^{-16} with zero class changes. ASH selects percentile 90 at both endpoints. All candidate fitting paths and selected evaluation paths are free of top-k boundary ties. DICE and Energy preserve their original scores and decisions; complete known-map transport restores baseline ASH exactly at retained precision. DICE is not uniformly superior (near/far AUROC 44.38%/90.66%).

The primary ranking test compares ASH-minus-Energy AUROC within each fixed OOD source. Opposite endpoint gaps and their simultaneous intervals must both lie beyond the ± 2-percentage-point band. There are 1,000 paired image-bootstrap replicates with seed 20260922; ID resamples are shared across methods/endpoints/sources, and OOD resamples across methods/endpoints within source. Four 98.75% marginal percentile intervals give a nominal Bonferroni 95% family, conditional on the fitted procedures. Table 5 retains all four intervals. Neither source reverses. Different signs across sources already at baseline do not count as a replacement effect.

EuroSAT illustrates the distinction between the recorded consequences. ASH AUROC changes from 86.20% to 86.25%, but OOD acceptance at the fitted cutoff changes from 57.50% to 64.58%, with

13.52% paired decision disagreement. ID acceptance changes from 94.81% to 96.19%. Tiny ImageNet OOD acceptance is 96.56% to 97.56%, consistent with weak discrimination there. Average ranking, operating-point acceptance, and per-image rule identity are different measurements. This result does not prove equivalence or the absence of utility consequences.

Second cohort. A separate prospective freeze selects torchvision ResNet-50 `IMAGENET1K_V2`, ImageNetV2 matched-frequency ID (3,000 calibration, 7,000 evaluation), Oxford-IIIT Pet (1,000 OOD validation, 5,000 evaluation), and Flowers102 (5,000 OOD evaluation). A pre-score amendment skips repeated image SHA-256 digests under the original deterministic order while preserving source roles and sample counts. The 21,000 selected paths and byte hashes are all unique. The frozen evaluation top-1 gate is 70%. Calibration accuracy is 69.90%; evaluation accuracy is 69.8857%, short by 0.1143 percentage points. An independent gate audit recomputes both.

All roles were extracted, then execution stopped. No representation arrow was evaluated, ASH percentile selected, detector cutoff fitted, or ASH/Energy/DICE evaluation score, AUROC, bootstrap interval, or ranking comparison computed or opened. Launch and audit records mark detector outcome access false. The already extracted OOD features do not supply a detector outcome for this cohort. The gate is not relaxed and the cohort is not rescued by changing models, datasets, or transformation strength. The failed gate establishes only that this cohort did not satisfy its predeclared condition; the small shortfall does not itself establish substantive classifier inadequacy. A future consequence test needs a separate scientific rationale and a new pre-score protocol.

D.13 Protocol, result and validation guide

The distance protocol and its cached-feature validation are in Appendix D.10; its complete comparisons are in Appendix E.6. The classification protocol and cached-feature recomputation are in Appendix D.8, with outcomes in Appendix E.7. Appendix D.11 specifies the manually modified network and raw-image checks; the resulting decisions are in Appendix E.8. The prospective ranking cohorts' data roles and stopping gate are in Appendix D.12, with the completed first cohort in Appendix E.9. The second cohort stopped at its classifier-competence gate before any detector outcome existed. These sections are separate evidence populations and validation scopes; their counts are not pooled.

E Complete results and validation records

The main text answers the empirical questions with the principal contrasts and limits. Here we retain the complete finite censuses, response and utility tables, confirmation outcomes, and validation boundaries needed to check those answers. The tables' denominators remain study-specific.

E.1 A counterexample under mild rescaling

The four SmoothQuant-formula transformations have condition numbers $\kappa = 1.56$–1.84 and yield 6.42% median ASH-B disagreement across 24 model/OOD cells. Both hyperparameter selection and the ID cutoff are refitted. The predicted controls agree, and a separately implemented ASH replay from cached features agrees with the result (Table 6 in the appendix). Each flipped row witnesses nonconstancy of the complete fitted decision on that tested pair. Thus disagreement is present for these mild, formula-derived interface changes as well as the stress transformations below: for tested ASH-B, exact predictor-function equivalence plus refitting from scratch does not reproduce the previous rejection policy. This experiment does not measure quantized-export behavior or the frequency of such replacements in deployment.

Table 6: SmoothQuant-formula interface-transformation summary. Medians are over six OOD roles. "Full" repeats APS and cutoff calibration; "fixed-ρ" retains the baseline percentile and recalibrates the cutoff. All flips are off-tie and all three predicted invariant controls agree.

Model	κ	APS $\rho \to \rho'$	full median	fixed-ρ median
ResNet-18 s0	1.6474	$95 \to 95$	7.89%	7.89%
ResNet-18 s1	1.7897	$90 \to 90$	5.05%	5.05%
ResNet-18 s2	1.8380	$90 \to 90$	6.04%	6.04%
ViT-base	1.5561	$90 \to 75$	10.29%	3.53%

Table 7: SmoothQuant-formula full/fixed-ρ ASH-B OOD flip percentages.

	CIFAR-10	TIN	MNIST	SVHN	DTD	Places
ResNet s0	7.84/7.84	9.42/9.42	7.72/7.72	8.30/8.30	7.94/7.94	7.16/7.16
ResNet s1	4.52/4.52	5.90/5.90	4.86/4.86	6.18/6.18	5.02/5.02	5.08/5.08
ResNet s2	5.88/5.88	6.66/6.66	5.94/5.94	7.16/7.16	6.14/6.14	5.64/5.64
ViT	3.86/1.42	8.58/4.00	3.02/0.68	12.58/3.06	21.88/8.88	12.00/5.10

E.2 The failure persists under stress transformations

For the all-row census, we use the permutation-equivariant tie completion. At $\kappa = 64$, the four per-model median OOD flip fractions range from 11.24% to 22.33% (Table 8). Near/far pooled medians are 19.53%/21.87%; the fixed-percentile pooled median is 22.51%. The full-refit pooled median is already 14.17% at $\kappa = 2$ and reaches 25.09% at $\kappa = 32$ (Table 11). Across all 1,152 cells, flip fractions range from 0.68% to 60.46%, with mean 20.58% and median 21.61%. These are paired finite-population descriptions, not prevalence estimates over possible transformations.

A post-hoc directional decomposition rules out a purely one-way acceptance shift. At $\kappa = 64$, every one of the 192 OOD cells contains both baseline-reject→endpoint- accept and baseline-accept→endpoint-reject changes. Across repeated row/endpoint events, the two directions contribute 121,100 and 73,249 flips; the median smaller- direction fraction is 5.40% per cell. APS changes its selected percentile at 23 of 32 endpoints, but the fixed-percentile result remains 22.51%, so reselection alone cannot explain the obstruction. This decomposition was not part of the predeclared criteria and is reported as descriptive follow-up.

Table 8: Stress-test $\kappa = 64$ result. ASH-B medians are over 48 seed/OOD cells per model. The final column reports exact zero OOD decision disagreement for each predicted invariant consumer across all eight transformations.

Model	d	baseline APS ρ	ASH-B median flips	RMS/DICE/Energy
ResNet-18 s0	512	95	21.59%	0 / 0 / 0
ResNet-18 s1	512	90	20.59%	0 / 0 / 0
ResNet-18 s2	512	90	22.33%	0 / 0 / 0
ViT-base	768	90	11.24%	0 / 0 / 0

All 192 transformations pass the raw-logit preservation check. The maximum absolute error in these checks is 1.4211×10^{-14}, with no class-argmax mismatch. At all 32 primary endpoints, RMS-ASH-B, transported DICE, and Energy have zero recorded confidence, consumer-logit, and Boolean-decision discrepancy; all four specificity controls also pass. These controls agree with the consumer-specific predictions in Table 2. The original stress study's DICE control transports the fitted training state; it is not an independent empirical refit of that state. The DICE proof covers either this covariant transport or recomputation from the transformed training role. ASH's primary comparison independently repeats selection and calibration. Following the same transformation through each consumer identifies where coordinate scale survives, cancels, or can be removed before ranking. The tested endpoints exhibit the resulting distinction between ASH failure and the preserving consumers.

E.3 FAILURE UNDER FRESH TRANSFORMATIONS

The CIFAR-10 study tests the same comparison criteria on a second ID dataset (Table 9). Across 144 fresh-transformation cells, completed disagreement has median 35.52% and range 25.08–50.12%; the conservative median is 35.01% with six cells with zero strict-rule support retained and 689,994/720,000 strict-rule-valid row instances. Every model and near/far criterion passes. Copying the baseline APS percentile still gives 38.38%/38.22% completed/conservative medians, so reselection is unnecessary for the effect. Raw logits pass the preservation check, all specificity controls have zero flips, and a separate implementation reconstructs all 720,000 serialized decision rows and their strict-rule-support flags. This confirms bounded transfer of the policy mismatch to fresh transformations and a second ID dataset, not fresh samples or external replication.

Table 9: CIFAR-10 final confirmation. Medians are over eight fresh transformations and six OOD roles per model; conservative retains rows outside strict-rule support as zeros.

Model	completed	conservative
ResNet-18 s0	32.40%	32.40%
ResNet-18 s1	37.43%	36.41%
ResNet-18 s2	36.75%	36.75%
Near / far	34.52% / 36.06%	34.35% / 35.64%

E.4 CIFAR-10 FINAL-CONFIRMATION OUTCOMES

The V2 transformation seeds are 1000003, 1000033, 1000037, 1000039, 1000081, 1000099, 1000117, and 1000121; their intersection with the invalid V1 seed set is empty. The near roles are CIFAR-100 and disjoint Tiny ImageNet; the far roles are MNIST, SVHN, Textures, and Places365. Each of the $3 \times 8 \times 6 = 144$ cells retains denominator 5,000. Literal freeze, launch, authorization, start, and finish times are ordered, and the runner checked that order before feature access.

Table 10: CIFAR-10 V2 predeclared-criterion census. Conservative values count every row outside strict-rule support as zero and retain all six cells with zero strict-rule support.

Quantity	completed	conservative	frozen minimum
Per-model median range	32.40–37.43	32.40–36.75	1.00 each
Near median	34.52	34.35	2.00
Far median	36.06	35.64	2.00
Copied-percentile median	38.38	38.22	1.00

All 12 model/control instances have zero decision flips. The separate reconstruction recovers every packed row vector, support flag, statistic, criterion result, and terminal outcome from stored decision/tie records. It does not recompute features, scores, APS, or cutoffs, and no V1 cell enters any displayed value.

E.5 RESPONSE CURVE, UTILITY AND DIRECTIONAL COUNTS

At condition 64, full-refit OOD flip-event counts by model are 53,076, 51,608, 56,757, and 32,908 for ResNet s0/s1/s2 and ViT. Event counts repeat rows across transformation endpoints and are not distinct-image counts. All model validity flags, exact consumer flags, and specificity-control flags pass.

For the fully refitted consumer at condition 64, the exact Boolean identity $f = u + d$ and acceptance shift $\Delta = u - d$ recover 121,100 baseline-reject-to-endpoint-accept events and 73,249 events in the reverse direction. Both u and d are positive in every cell. This is a post-hoc decomposition of the recorded decisions, not an additional confirmatory endpoint.

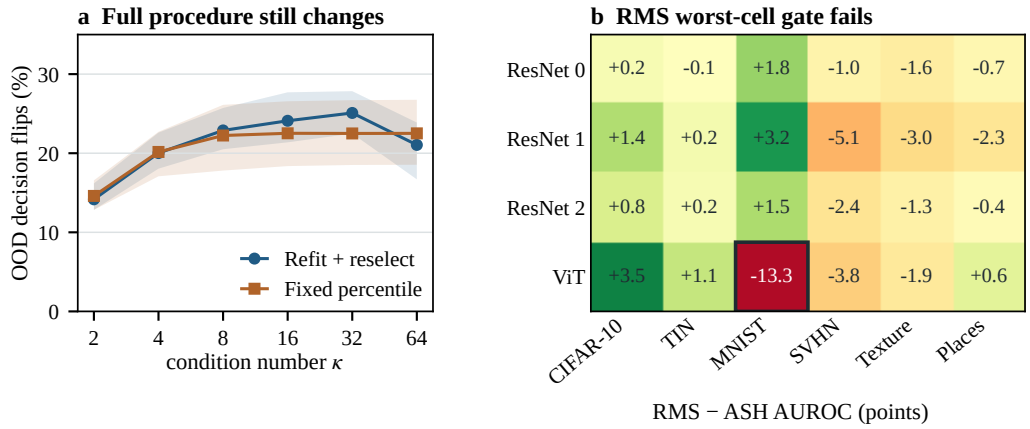


Figure 7: **a:** Completed ASH-B disagreement across the original condition-number sweep; shading is the interquartile range over 192 fixed cells per condition. **b:** RMS-ASH baseline AUROC change; ViT/MNIST violates the predeclared worst-cell utility margin.

Table 11: Completed ASH-B response curve. Each entry summarizes 192 fixed model/seed/OOD cells. Intervals are interquartile ranges (IQR), not confidence intervals. "Full" repeats APS and calibration; "fixed" retains the baseline percentile and recalibrates.

κ	full median	full IQR	fixed median	fixed IQR
2	14.17	[12.84, 16.19]	14.60	[12.83, 16.57]
4	20.02	[18.05, 22.61]	20.16	[17.08, 22.71]
8	22.88	[20.50, 25.72]	22.24	[17.81, 26.12]
16	24.10	[21.38, 27.70]	22.52	[18.37, 26.57]
32	25.09	[22.48, 27.85]	22.50	[18.51, 26.71]
64	21.05	[16.70, 23.88]	22.51	[18.54, 26.76]

Table 12: RMS-ASH minus ASH-B AUROC in percentage points. The ViT/MNIST cell is the predeclared utility-criterion failure.

	CIFAR-10	TIN	MNIST	SVHN	DTD	Places
ResNet s0	+0.16	-0.09	+1.76	-1.04	-1.61	-0.67
ResNet s1	+1.43	+0.24	+3.21	-5.15	-3.05	-2.26
ResNet s2	+0.76	+0.18	+1.48	-2.36	-1.25	-0.43
ViT	+3.53	+1.12	**-13.28**	-3.77	-1.90	+0.59

E.6 Distance-based OOD consumers: contrasting fitted metrics

On the same four CIFAR-100 interfaces, we independently rebuild each training bank or class-conditional metric, select its parameter by validation AUROC, and recalibrate its ID cutoff. Both raw and row-normalized kNN use negative squared distance to the selected kth training neighbor. Mahalanobis confidence is the negative minimum squared distance to a fitted class mean, including the query's quadratic term. The metric is full pooled within-class covariance, isotropic shrinkage, or diagonal shrinkage. These classical distance scores test a mechanism different from ASH's coordinate ranking.

Table 13: Distance-based OOD development on reused CIFAR-100 interfaces. Counts are comparisons with nonzero disagreement / valid comparisons. Median disagreement is in percent over 24 cells per arrow (18 for full covariance); ViT/full covariance is undefined. Each cell uses 5,000 OOD rows.

Consumer	Changed / valid	Mild median	$\kappa = 2$ median
Normalized kNN	48/48	2.09	6.29
Raw kNN	48/48	1.65	5.33
Isotropic shrinkage	42/48	2.64	6.49
Diagonal shrinkage	0/48	0.00	0.00
Full covariance	0/36	0.00	0.00

Under a positive diagonal map, a covariance metric satisfying $M' = gMg^\top$ preserves each candidate distance. Full covariance and diagonal shrinkage have this property on their stated domains, so Lemma 1 carries candidate equality through selection and calibration to the decisions. Equation (4) also derives the diagonal formula from standardized isotropic shrinkage. Euclidean distances and isotropic shrinkage lack this nonuniform-scale guarantee. Table 13 tests these contrasting predictions: normalized and raw kNN each change decisions in 48/48 comparisons, isotropic shrinkage in 42/48, and the preserving variants in none of their valid comparisons. These cell counts share models and images and are not independent trials or estimates of a failure probability.

The unregularized ViT covariance fails the numerical rank gate at all three endpoints; its full-covariance comparisons remain undefined. All kNN fits select $k = 1$, so this study does not exhibit changed selected k. Complete transport using the inverse map and baseline fitted state restores every valid consumer's decisions. Maximum absolute AUROC changes are 2.49, 2.43, and 4.62 percentage points for normalized kNN, raw kNN, and isotropic shrinkage, respectively; these changes do not establish ranking reversals. The result supports complete-procedure preservation conditions on this reused population, with classical covariance identities supplying the positive predictions.

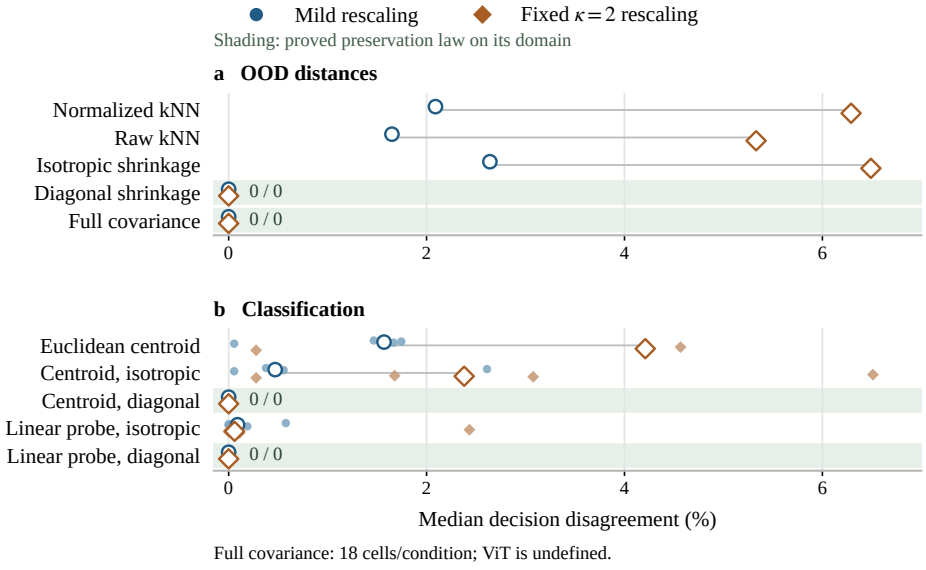


Figure 8: Supporting full-refit contrasts under mild ($\kappa = 1.56$–1.84) and fixed $\kappa = 2$ positive diagonal changes. Hollow marks are medians; faint classification marks are the four individual model disagreements per condition. OOD medians cover 24 model/dataset cells with 5,000 rows each (full covariance: 18 valid cells; ViT undefined); classification uses four models with 9,000 test rows each. Shading denotes a proved preservation law on its domain. Individual cells share models and images; the chart reports agreement, not accuracy or uncertainty intervals.

E.7 A CLASSIFICATION COMPARISON BEYOND OOD REJECTION

The same interface question applies when the consumer returns a class label. On the four existing CIFAR-100 interfaces we fit Euclidean nearest-centroid classifiers, isotropic- and diagonal-shrinkage centroid metrics, and squared-loss linear probes with isotropic or diagonal penalties. At each endpoint, 50,000 training rows refit statistics and coefficients, 1,000 labeled selection rows choose regularization, and 9,000 test rows measure disagreement and accuracy. The consumer definitions and preservation predictions preceded their evaluation; models and samples are reused. Appendix B.3 gives the objectives and covariance argument; Appendix D.8 gives the selection and numerical checks.

Table 14: Classification extension on the four existing CIFAR-100 interfaces. Baseline accuracy is a range across models; disagreement is the median over four models per transformation. Accuracy change spans all eight model/transformation comparisons per consumer. All entries are percentages or percentage points.

Consumer	Baseline acc.	Mild flips	$\kappa = 2$ flips	Acc. change
Euclidean centroid	77.02–92.09	1.57	4.21	-0.47 to $+0.10$
Centroid, isotropic	76.71–92.09	0.47	2.38	-0.53 to $+0.40$
Centroid, diagonal	76.67–92.10	0.00	0.00	$+0.00$ to $+0.00$
Linear probe, isotropic	76.56–92.31	0.09	0.06	-0.08 to $+0.02$
Linear probe, diagonal	76.56–92.32	0.00	0.00	$+0.00$ to $+0.00$

Euclidean-centroid disagreement has median 1.57% under the four mild arrows and 4.21% under one fixed $\kappa = 2$ arrow per model (Table 14). Diagonal-shrinkage centroids and diagonal-penalty probes have zero disagreement, as predicted from their covariant fitted statistics and candidate-wise score preservation. Equations (8) and (4) place both formulas under the same normalization principle, with their classification selection tail. Each diagonal variant preserves its own baseline classifier; retaining another consumer's original rule uses full transport. All specificity and complete-transport controls pass. The isotropic probe agrees on both arrows for two ResNets, despite lacking a general nonuniform-scale guarantee. Across all consumers and both arrows, accuracy changes range from -0.53 to $+0.40$ percentage points. Thus this extension supplies a competent classification instance of behavioral dependence and contrasting preservation, with small net accuracy changes. It establishes neither a large utility penalty nor compatibility of independently trained modules.

Table 15: Classification census. Each cell uses all 9,000 ID-test rows. Accuracies and disagreement are percentages; Δ accuracy and its paired 95% interval are percentage points.

Model / consumer	Arrow	Base acc.	New acc.	Flips	Δ acc. [95%]
cnn_s0 / EC	mild	77.24	77.34	1.47	$+0.10$ $[-0.08, +0.28]$
cnn_s0 / CI	mild	76.77	76.82	0.38	$+0.06$ $[-0.03, +0.14]$
cnn_s0 / CD	mild	76.67	76.67	0.00	$+0.00$ $[+0.00, +0.00]$
cnn_s0 / RI	mild	76.82	76.84	0.58	$+0.02$ $[-0.09, +0.14]$
cnn_s0 / RD	mild	76.76	76.76	0.00	$+0.00$ $[+0.00, +0.00]$
cnn_s0 / EC	kappa2	77.24	77.17	4.57	-0.08 $[-0.41, +0.24]$
cnn_s0 / CI	kappa2	76.77	77.17	6.51	$+0.40$ $[+0.03, +0.78]$
cnn_s0 / CD	kappa2	76.67	76.67	0.00	$+0.00$ $[+0.00, +0.00]$
cnn_s0 / RI	kappa2	76.82	76.77	2.43	-0.06 $[-0.29, +0.18]$
cnn_s0 / RD	kappa2	76.76	76.76	0.00	$+0.00$ $[+0.00, +0.00]$
cnn_s1 / EC	mild	77.02	77.11	1.74	$+0.09$ $[-0.10, +0.28]$
cnn_s1 / CI	mild	76.71	76.26	2.61	-0.46 $[-0.70, -0.20]$
cnn_s1 / CD	mild	77.08	77.08	0.00	$+0.00$ $[+0.00, +0.00]$
cnn_s1 / RI	mild	76.59	76.59	0.00	$+0.00$ $[+0.00, +0.00]$
cnn_s1 / RD	mild	76.59	76.59	0.00	$+0.00$ $[+0.00, +0.00]$
cnn_s1 / EC	kappa2	77.02	76.73	4.22	-0.29 $[-0.60, +0.02]$
cnn_s1 / CI	kappa2	76.71	76.56	1.68	-0.16 $[-0.36, +0.06]$
cnn_s1 / CD	kappa2	77.08	77.08	0.00	$+0.00$ $[+0.00, +0.00]$
cnn_s1 / RI	kappa2	76.59	76.59	0.00	$+0.00$ $[+0.00, +0.00]$
cnn_s1 / RD	kappa2	76.59	76.59	0.00	$+0.00$ $[+0.00, +0.00]$
cnn_s2 / EC	mild	77.28	77.14	1.67	-0.13 $[-0.32, +0.06]$
cnn_s2 / CI	mild	76.73	76.74	0.56	$+0.01$ $[-0.11, +0.13]$
cnn_s2 / CD	mild	76.74	76.74	0.00	$+0.00$ $[+0.00, +0.00]$
cnn_s2 / RI	mild	76.56	76.56	0.00	$+0.00$ $[+0.00, +0.00]$
cnn_s2 / RD	mild	76.56	76.56	0.00	$+0.00$ $[+0.00, +0.00]$
cnn_s2 / EC	kappa2	77.28	76.81	4.20	-0.47 $[-0.77, -0.16]$

Model / consumer	Arrow	Base acc.	New acc.	Flips	Δ acc. [95%]
cnn_s2 / CI	kappa2	76.73	76.20	3.08	-0.53 [-0.79, -0.28]
cnn_s2 / CD	kappa2	76.74	76.74	0.00	$+0.00$ [$+0.00$, $+0.00$]
cnn_s2 / RI	kappa2	76.56	76.56	0.00	$+0.00$ [$+0.00$, $+0.00$]
cnn_s2 / RD	kappa2	76.56	76.56	0.00	$+0.00$ [$+0.00$, $+0.00$]
vit / EC	mild	92.09	92.09	0.06	$+0.00$ [-0.04, $+0.04$]
vit / CI	mild	92.09	92.09	0.06	$+0.00$ [-0.04, $+0.04$]
vit / CD	mild	92.10	92.10	0.00	$+0.00$ [$+0.00$, $+0.00$]
vit / RI	mild	92.31	92.23	0.19	-0.08 [-0.17, $+0.00$]
vit / RD	mild	92.32	92.32	0.00	$+0.00$ [$+0.00$, $+0.00$]
vit / EC	kappa2	92.09	92.14	0.28	$+0.06$ [-0.04, $+0.16$]
vit / CI	kappa2	92.09	92.14	0.28	$+0.06$ [-0.04, $+0.16$]
vit / CD	kappa2	92.10	92.10	0.00	$+0.00$ [$+0.00$, $+0.00$]
vit / RI	kappa2	92.31	92.24	0.11	-0.07 [-0.13, -0.01]
vit / RD	kappa2	92.32	92.32	0.00	$+0.00$ [$+0.00$, $+0.00$]

E.8 Raw-image replay through live network graphs

We insert the fixed formula rescaling and compensated head into the actual ResNet-18 s0 and ViT forward graphs, then replay 4,972 raw images per model across fitting and evaluation roles. All original class argmax predictions agree; compensated logits have maximum float32 error below 3.82×10^{-6}. Both exported graphs contain the inserted multiplication and match their eager fixtures. Refitting ASH-B on graph emissions changes the OOD decisions in Table 16, while Energy stays unchanged. The ViT's selected ASH percentile changes from 90 to 75. Graph and cache-algebraic terminal decisions agree on every checked test row.

Table 16: Live-graph development replay: each row compares the same 1,000 OOD images. ASH flips and AUROC are percentages; Energy reports changed-decision counts. All original classifier argmax predictions are preserved. The transformations are manually installed.

Model	OOD	ASH flips	AUROC base	AUROC graph	Energy flips
ResNet-18	CIFAR-10	6.1	71.14	70.44	0
ResNet-18	Texture	6.9	79.97	79.21	0
ViT	CIFAR-10	9.5	64.84	62.01	0
ViT	Texture	23.9	80.72	78.07	0

This replay rules out dependence on manipulating only detached feature caches. It is a manually installed transformation on exposed models and images; graph capture does not show that a quantizer or exporter naturally produces it. It supplies an apparatus check, not independent-population confirmation or a measured incidence of reuse failures.

E.9 Prospective ranking tests delimit the consequence

The first prospective ranking cohort uses an ImageNet-pretrained ResNet-18 with a ridge head fitted on CIFAR-10 training data, 1,600 fresh CIFAR-10.1 evaluation images, and 5,000 evaluation images each from Tiny ImageNet test and EuroSAT. A calibration-derived formula arrow preserves the original logits and class predictions. ASH independently repeats selection and calibration, changing OOD decisions by 2.92% and 13.52%, respectively. Nevertheless, neither within-source ASH/Energy ranking reverses (Table 17). The four simultaneous gap intervals remain on the same side of the predeclared ± 2-point band within each source. This is not an equivalence test.

Table 17: Prospective CIFAR-10.1 ranking comparison. ASH-minus-Energy AUROC gaps are percentage points; ASH OOD disagreement is percent. A resolved reversal must cross the ± 2-point band within a fixed source, including the simultaneous intervals. Opposite rankings across different sources are not replacement-induced reversals.

OOD source	Gap base	Gap changed	Disagreement	Reversal
Tiny ImageNet test	-7.10	-6.72	2.92	No
EuroSAT	+11.70	+11.75	13.52	No

Unchanged rankings do not imply unchanged operating behavior. On EuroSAT, ASH OOD acceptance rises from 57.50% to 64.58%, while ID acceptance rises from 94.81% to 96.19%. These are endpoint-specific calibrated operating points, not a comparison at matched ID-test acceptance. They show why rule identity, acceptance rates, and AUROC rankings must be reported separately.

The first cohort passes its frozen 65% calibration-accuracy gate but has only 67.69% ID evaluation accuracy. Near-OOD ASH/Energy AUROCs are approximately 49%/56%, limiting the strength of that comparison. Energy and DICE preserve their scores and decisions; full known-map transport restores the baseline ASH rule. The arrow's realized condition number is 6.67, so the formula called "mild" in the CIFAR-100 study is not necessarily a small perturbation on this new interface.

A second prospective cohort fixes an ImageNet-pretrained ResNet-50, ImageNetV2 ID, and Pet/Flowers OOD sources. Its 7,000-image ID evaluation accuracy is 69.8857%, below the frozen 70% competence gate. Execution stops before arrow evaluation, ASH selection, detector calibration, or any detector score/AUROC comparison. No detector-level outcome exists for this attempt. It cannot be counted as either a reversal or a second no-reversal result. Together, the attempts leave fresh-identity comparative-ranking confirmation unresolved while retaining the separate evidence for changed decisions.

E.10 Preservation is distinct from OOD utility

Fitted RMS normalization eliminates interface disagreement but fails the predeclared utility criterion (Table 12 in the appendix). Mean RMS-minus-ASH AUROC and FPR changes are -0.00933 and -0.00759, satisfying their mean margins. ViT/MNIST nevertheless loses 0.13282 AUROC, exceeding the allowed 0.10; the worst FPR increase is 0.032. Descriptive paired-bootstrap 95% intervals are $[-0.01082, -0.00775]$ for mean AUROC and $[-0.00938, -0.00585]$ for mean FPR. They pair methods within each model while independently resampling models' shared images, omit cross-model image dependence, and do not determine whether the criterion passes.

RMS-ASH is invariant but does not preserve utility on the tested population; invariance alone does not certify detection quality. For an original consumer with fitted state a, shaping operator S_a, reference interface, and transformation g, define $S_{g,a}(z) = gS_a(g^{-1}z)$ and transport the remaining state, including cutoff and tie policy. Then $(Wg^{-1})S_{g,a}(gh) + b = WS_a(h) + b$. Scores, decisions, and utility agree exactly on the transported domain (Appendix A.5). This uses reference/transformation information unavailable to endpoint-local refitting; the RMS failure does not establish an unavoidable invariance–utility tradeoff. The preservation formulation thus distinguishes three obligations: agreement of endpoint-local fitted decisions, utility of an invariant consumer, and exact recovery of an existing policy using retained transformation data. Each procedure requires its own structural or reference information at the interface.

E.11 Evidence and replay boundary

The protocol and result sections above give the reader-facing index to each study's population, comparison and outcome. Replays of the original stress and formula-derived ASH studies recompute consumer outputs from cached features. The paired state-copy diagnostic and CIFAR-10 transfer reconstructions instead start from stored Boolean decisions and tie records; they reproduce row counts, aggregates and criterion results but do not recompute features, scores, selection or cutoffs. The distance and classification extensions have separate cached-feature recomputations (Appendices D.10 and D.8); the live-graph check replays raw images through manually modified networks (Appendix E.8). None of these checks retrains the original encoders or creates an independent model population.

Two corrections limit how the results are read without changing their measured outcomes. All primary ASH fitting paths and observed decision flips are off the top-k boundary, but some nonflipping evaluation rows require the stated tie completion; the evaluated domain cannot be called wholly tie-free. In the paired state-copy diagnostic, the strict-rule premise holds in 156 of 192 cells; the remaining 36 keep zero contribution in conservative summaries. The initial all-cell support statement was incorrect, while the recorded decisions and predeclared criterion results remain the same (Appendix F.2).

F SECONDARY DIAGNOSTICS AND EVIDENCE LIMITS

The main claim compares complete endpoint-local decisions. Copying numerical fitted state asks a different, secondary question. This appendix gives that diagnostic, its exact identities and coordinate dependence, then the adverse controls and excluded attempts that constrain interpretation.

F.1 A SECONDARY FITTED-STATE DIAGNOSTIC

After establishing endpoint policy preservation or failure, we predeclared a separate 32-pair diagnostic that copies the numerical ASH percentile and cutoff across transformed interfaces. Its four-corner quantities compare full-refit disagreement with sensitivity to this copied state. They are deliberately secondary: state substitutions are not presentation arrows, and the resulting quantities depend on score coordinates, so they cannot identify causal or intrinsic components of the descent failure. Appendix F.2 gives the estimands, tie-support rules, frozen criteria, and full results.

Observed fitted-state sensitivity. The paired study also finds full-refit mismatch: completed per-model median disagreement ranges from 12.49% to 23.45%. Copying fitted percentiles and cutoffs across endpoints reveals state-copy sensitivity even when opposing contrasts nearly cancel. Those auxiliary contrasts depend on numerical score coordinates; full-refit disagreement is unchanged by endpoint-specific increasing score recodings with recalibrated cutoffs. Appendix F.2 gives the estimands, conservative tie-support results, and figure; Table 18 records the criteria.

F.2 PAIRED STATE-COPY DIAGNOSTIC

The primary comparison tests endpoint policy preservation; a secondary analysis asks how sensitive its disagreement is to copied numerical fitted state. This paired study compares 32 unordered centered pairs fixed before evaluation: eight per model, with each endpoint at condition number 8 from the baseline interface and relative condition number 64. For endpoints p_0, p_1, fit ASH independently to obtain $s_i = (\rho_i, \tau_i)$, the selected APS percentile and calibrated cutoff. For this diagnostic, we copy the numerical percentile and cutoff from either endpoint into either interface. Let $e_x(p, s)$ denote the Boolean ASH-B decision on row x obtained by evaluating interface p with numerical state $s = (\rho, \tau)$, without refitting that state. The four corner decisions are

$$D_{i\leftarrow j}(x) = e_x(p_i, s_j), \qquad i, j \in \{0, 1\}. \tag{22}$$

Copying a numerical state defines a comparison convention; it is distinct from transporting the complete consumer (Appendix F.4).

For each row define

$$\begin{aligned} F_x &= \mathbf{1}[D_{0\leftarrow 0} \neq D_{1\leftarrow 1}], \\ H_x &= \tfrac{1}{2}\sum_{j=0}^{1} \mathbf{1}[D_{0\leftarrow j} \neq D_{1\leftarrow j}], & A_x &= \tfrac{1}{2}\sum_{i=0}^{1} \mathbf{1}[D_{i\leftarrow 0} \neq D_{i\leftarrow 1}], \\ U_x &= \max(F_x - H_x, 0), & V_x &= \max(H_x - F_x, 0). \end{aligned} \tag{23}$$

Here F is full-refit diagonal disagreement, H averages two held-state interface comparisons, and A measures sensitivity to which endpoint supplies fitted state. For every binary rectangle, $A_x = |F_x - H_x| = U_x + V_x$, $F_x - H_x = U_x - V_x$, and $U_x V_x = 0$ (Appendix F.4). This is an exact state-copy diagnostic, not an additive or causal decomposition. We summarize cancellation by $C = 2\min(\bar{U}, \bar{V})/(\bar{U}+\bar{V})$, where bars denote means over the fixed rows and cells in the reported stratum; C is undefined when $\bar{U} + \bar{V} = 0$.

The numerical state-copy convention is substantive. Endpoint-specific strictly increasing recodings of scores and their calibrated cutoffs preserve F, but can change H, A, U, V, C when the new numerical cutoffs are copied across endpoints. Thus these are coordinate-dependent diagnostics, not intrinsic components of the full-refit failure (Appendix F.5).

The predeclared criterion applies both to the symmetric tie-completed rule and to a conservative lower bound for the released rule: each model's median F must be at least 1%, near/far medians at least 2%, and near/far A medians at least 0.5%. A row belongs to the strict-rule support only

when both fitting paths and all four corner evaluations are off the ASH boundary; rows outside that support contribute zero while every cell keeps denominator 5,000. This gives lower bounds for F, H, A, U, V. The ratio C computed on strict-rule-valid components is a cancellation diagnostic, not a lower bound on all-row cancellation. All 32 pairs, 192 cells, and 960,000 row instances were fixed before the single permitted outcome evaluation.

The original strict-domain attempt stopped at boundary ties; the subsequent permutation-equivariant completion and the one-attempt paired study are recorded in Appendix F.6. The later coverage correction changed prose only.

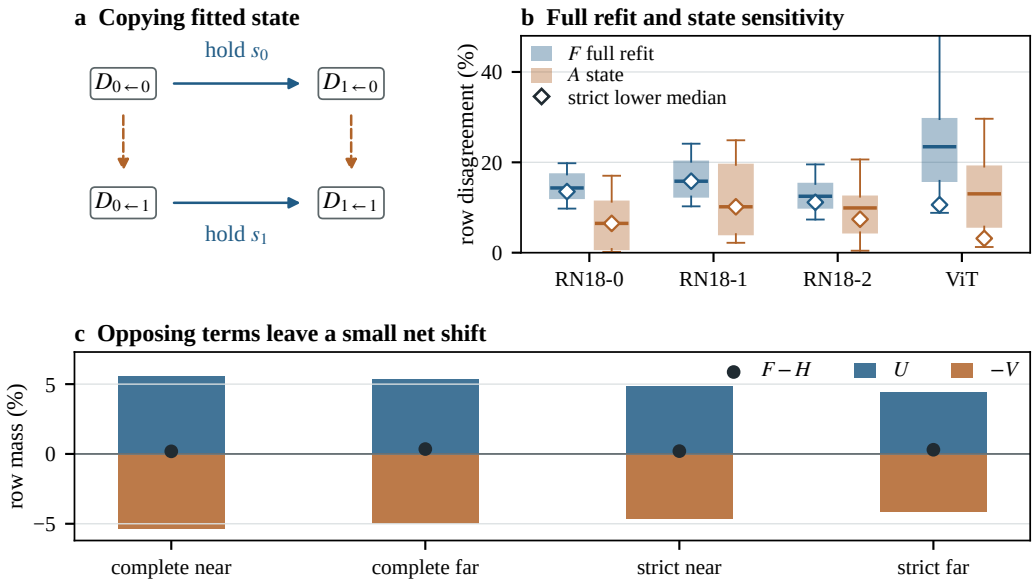


Figure 9: Paired state-copy diagnostic. **a:** The four-corner evaluation rectangle; horizontal comparisons hold fitted state fixed, while vertical state substitutions are comparators rather than groupoid arrows. **b:** Cell-level completed F and A distributions (boxes) and conservative medians (diamonds). **c:** Opposing U and V masses almost cancel, leaving a small signed $F-H$ contrast despite state-copy sensitivity. Only F is invariant to endpoint-specific increasing score recodings with recalibrated cutoffs.

Observed diagnostic behavior. The full-refit mismatch above is the preservation result; the paired study asks a secondary question about sensitivity to copied numerical fitted state. It also shows full-refit disagreement under both decision rules (Figure 9). Completed per-model median F is 14.33%, 15.82%, 12.49%, and 23.45% for ResNet s0/s1/s2 and ViT; the corresponding conservative medians, with all zeros retained, are 13.53%, 15.82%, 11.15%, and 10.60%. Completed near/far F medians are 16.44%/13.20%; lower medians are 15.97%/12.33%. Thus the result is not driven by one architecture or OOD stratum.

Under the declared numerical state-copy convention, completed near/far A medians are 12.05%/8.67%, with conservative medians 10.11%/7.73%, all far above the predeclared 0.5% threshold. Yet the mean signed contrast $\bar{F} - \bar{H} = \bar{U} - \bar{V}$ is only 0.19–0.35 percentage points. Under this convention, $\bar{U}$ and $\bar{V}$ are each approximately 0.04–0.06, producing 96.48–98.28% cancellation across the four completed/strict-rule-valid near/far strata. These are weighted disagreement contrasts; their signs concern $F - H$, rather than the direction of an ID/OOD decision flip. Small net movement coexists with large state-copy sensitivity in these score coordinates, without providing an intrinsic attribution of F to fitting.

The strict rule is valid for 779,991 of 960,000 row instances in 156 of 192 cells. The remaining 36 cells contain fitting-path boundary ties and remain zero in every lower-bound median. All eight

leave-one-pair-index-out checks preserve both criteria. A separate implementation checks stored decisions and tie records; it does not replay the feature-to-score computation (Appendix D.7).

F.3 Predeclared outcomes of the paired study

The successor fixed 32 centered unordered pairs, six OOD roles, the full four-corner decision vectors, both completed and conservative estimands, and conjunctive thresholds before its single outcome evaluation. Cancellation was a predeclared diagnostic and could not override a missed F or A criterion.

Table 18: Paired-study criterion census. All entries are percentages; cells with zero strict-rule support remain zero in conservative medians. The last two entries in that column are strict-rule-valid cancellation diagnostics, not lower bounds on all-row cancellation.

Quantity	completed	conservative	frozen minimum
Per-model F median range	12.49–23.45	10.60–15.82	1.00 each
Near F median	16.44	15.97	2.00
Far F median	13.20	12.33	2.00
Near A median	12.05	10.11	0.50
Far A median	8.67	7.73	0.50
Near cancellation C	98.28	97.85	diagnostic
Far cancellation C	96.60	96.48	diagnostic

The strict rule is valid for 779,991/960,000 row instances and 156/192 cells; all eight joint leave-one-pair-index-out analyses preserve both criteria. No row is dropped and no denominator changes.

F.4 Binary evaluation-rectangle identity

The numerical state-copy convention. The decisions $D_{i\leftarrow j}$ evaluate interface i with the numerical state fitted at interface j; Appendix F.2 defines the four-corner diagnostics. Let S contain the numerical percentile–cutoff pairs and extend the action by $g \cdot (p, s) = (g \cdot p, s)$. The state-bearing groupoid $\mathcal{E} = (P \times S)//G_d^+ \simeq \mathcal{P} \times S_{\text{disc}}$ projects to $\mathcal{P}$ by forgetting state; S_{disc} has those states as objects and only identity arrows. Each fixed-state comparison lifts a presentation arrow to $\mathcal{E}$. State substitutions are comparisons, not arrows of $\mathcal{E}$; the fitted diagonal is not a lift when $s_0 \neq s_1$.

Let $a = D_{0\leftarrow 0}$, $b = D_{1\leftarrow 0}$, $c = D_{0\leftarrow 1}$, and $d = D_{1\leftarrow 1}$ be bits. If $a = d$, complementing all four bits if needed gives $a = d = 0$, and Equation (23) yields $F = 0$ and $A = H = (b + c)/2$. If $a \neq d$, complementing if needed gives $a = 0, d = 1$; then $F = 1$, $H = (b + 1 - c)/2$, and $A = (c + 1 - b)/2 = 1 - H$. Hence in both cases $A = |F - H|$. The identities $A = U + V$, $F - H = U - V$, and $UV = 0$ follow from the positive and negative parts of $F - H$. This exhaustive bit identity justifies the diagnostic rectangle, but does not make its edges interventions or its terms causal components.

F.5 Score-coordinate dependence of the state-copy diagnostic

Write $q_{i,\rho}$ for the score at presentation i with candidate percentile ρ. Replace all scores at endpoint i by $\phi_i \circ q_{i,\rho}$ for a strictly increasing bijection $\phi_i : \mathbb{R} \to \mathbb{R}$. Candidate validation AUROCs and the first-maximum APS choice are unchanged; the order-statistic cutoff becomes $\phi_i(\tau_i)$. Hence the two endpoint-local decisions, and therefore F, are unchanged. A numerical cross-copy comparison instead becomes

$$D'_{i\leftarrow j}(x) = \mathbf{1}[\phi_i(q_{i,\rho_j}(x)) > \phi_j(\tau_j)],$$

which need not equal $D_{i\leftarrow j}(x)$ when $i \neq j$.

For an exact minimal example, fix one candidate and let both endpoints have singleton calibration score 0 and target score 1. All four decisions are 1, so $F = H = A = 0$. Translate endpoint 1's entire score function and cutoff by 10. Its calibration score is now 10 and target score 11. The ordered bits $(D'_{0\leftarrow 0}, D'_{1\leftarrow 0}, D'_{0\leftarrow 1}, D'_{1\leftarrow 1}) = (1, 1, 0, 1)$ give $F' = 0$ but $H' = A' = 1/2$. No

endpoint-local decision changed and there is no cutoff equality. A common recoding at both endpoints preserves every corner. More generally, translating a foreign cutoff into the recipient's score coordinates as $\phi_i(\phi_j^{-1}(\phi_j(\tau_j)))$ also preserves the original rectangle.

Thus the binary identity is exact for each declared rectangle, but its state-copy quantities H, A, U, V and aggregate C are not intrinsic to the two fitted decision rules. The example qualifies their interpretation; it is not a new classifier transformation, an empirical reanalysis, or a counterexample to the measured nonzero F. The experiment retains its original numerical percentile–cutoff convention throughout; its numbers and criterion outcomes are unchanged.

F.6 Adverse controls and excluded attempts

The ReAct comparison is an adverse result rather than support for a broader consumer-failure claim. An initial attempt assigned validation and test source roles incorrectly and is excluded. The corrected fixed test passes source, arrow-validity and specificity controls, but its largest near-OOD median flip is 8.01% against the required 10% (far OOD: 10.26%), and its Energy-referenced utility conjunction fails. The exploratory flips therefore do not satisfy the prespecified released-consumer-failure or useful-repair conditions.

The activation-source check initially compared best weights with receipts for final weights; source roles were corrected before OOD outcomes were opened. The initial strict ASH top-k test stopped at a boundary tie without yielding a gauge result. The later symmetric, permutation-equivariant completion kept the original utility failure. Under that completion every row remains in the estimand: observed primary flips are off-tie, while some nonflipping evaluation rows use the tie rule (Appendix D.7).

The paired state-copy analysis initially described all 192 cells as having strict-rule support. The correct count is 156; the other 36 remain zero in conservative medians. This correction changes the support description, not the decisions or criterion results. A separate CIFAR-10 transfer attempt had freeze labels entered after its outcomes and is excluded from confirmation. Only the later, separately fixed 144-cell study contributes the reported transfer result; none of the earlier cells is pooled. Its transformation outcomes were unopened at its own freeze, although the earlier attempt's outcomes were already known. This corrects an overbroad prior-exposure description without changing the follow-up's samples or statistics.

F.7 Detailed scope and responsible-use boundary

The study can reveal a representation-dependent downstream decision, but it does not establish harm frequency, adversarial accessibility in a deployment, or the safety of a particular system. Positive diagonal adapters are exact stress transformations; the workflow successor selects four of them with a deployment formula but does not observe an actual export trace. An incidence study would need separately predeclared exporter populations, logged transformations, and operational costs; tuning it on the present outcomes would not be confirmatory.

The result should not be used to claim that ASH broadly fails OOD detection. ASH may retain strong OOD utility on its intended interface, and our own canonical repair loses unacceptable utility in one cell. The correct engineering implication is to state and test a preservation contract for any internal-feature consumer under declared representation transformations alongside accuracy, calibration, coverage, and non-vacuity.

G Further related-work context

Representation reuse and interface compatibility. Model stitching tests compatibility by fitting an adapter between frozen model parts (Bansal et al., 2021); Mai et al. (2026) extend it to heterogeneous vision foundation models and show that the adapter's training objective affects downstream performance. WM3C identifies language-controlled causal components up to within-block invertible transformations (Wang & Huang, 2025). Such identifiability specifies information recovery; whether a particular downstream fitting procedure respects the remaining coordinate freedom is a separate question. BRICKS-WM explicitly conditions reuse of a frozen dynamics module on matching conditional interface distributions (Zhang et al., 2026). These works already address adaptation

and compatibility. Our framework specifies the structures required by a complete fitted consumer. Its controlled empirical comparison uses invertible feature changes with a fixed original classifier, and includes fitting, selection, and calibration in the tested behavior. It does not test those methods or infer that their reuse mechanisms fail.

Same-function preservation. Implementation invariance is an attribution axiom (Sundararajan et al., 2017); sharpness (Dinh et al., 2017), Bayesian inference, and pruning (Roy et al., 2024; Asiaee, 2026) can depend on reparameterization. Karasan (2026) studies fixed-threshold validity under orthogonal reparameterization; we test released ASH under positive diagonal rescaling with full APS reselection and recalibration. Kornblith et al. (2019, Sections 2.1–2.2) already distinguish function-preserving feature changes from learning invariance and construct general-linear invariant similarities by orthonormalization. Prediction-only analysis is therefore a conditional foil, not a description of all existing representation theory. Positive-rescaling symmetry and monomial equivariance are prior art (Grigsby et al., 2023; Tran et al., 2024); the categorical ingredients follow (Mac Lane, 1998; Maruyama, 2026). Here they define a preservation test for fully refitted consumers under same-function transformations. Deployment equalization motivates exact scaling (Meller et al., 2019; Xiao et al., 2023), without implying exporter incidence.

Structural frameworks and residual symmetry. Gauge-equivariant CNNs organize feature types and transport through common compatibility laws (Cohen et al., 2019). Category-equivariant networks explicitly compose natural layers; their Grothendieck example combines varying configurations, supplied alignment, and removal of residual rotation (Maruyama, 2026, Appendix A). Sheaf-based federated representation learning jointly fits heterogeneous representations and alignment maps (D'Acunto et al., 2026). Our framework specializes these structural principles to a complete downstream procedure, including its returned fit, candidate selection, calibration, and terminal observable. It distinguishes covariance of transported structure, invariance of an independently refitted recipe, and recovery with a known reference. If a lawful normalization supplies identical complete inputs to a deterministic learner, preservation already follows; the work here is to establish the required structure and remaining compatibility for consumers that need not receive such identical inputs. We do not claim to introduce naturality, symmetry reduction, or categorical equivariance.

Contextual equivalence and minimal information. Program semantics defines equivalence by indistinguishability under declared contexts (Pitts, 2002). Ranks are classical maximal invariants for common increasing transformations (Lehmann & Scholz, 1992, p. 45). We apply these principles to typed consumer states and compute the information required by explicit selection/calibration and quadratic-refitting contracts. The common quotient explains both why distinct numerical states can be reusable and why adding a fitting operation can require a finer state description. The rank and matrix arguments are elementary; their role is to make the framework's information requirements and compositional obligations concrete.

Feature shaping and OOD evaluation. ASH, ReAct, and DICE modify late activations or weights; Energy is output-only (Djurisic et al., 2023; Sun et al., 2021; Sun & Li, 2022; Liu et al., 2020), while SCALE attributes ASH's gain to scaling rather than pruning (Xu et al., 2024). Scale-aware methods provide invariances (Demirel et al., 2025; Lamaakal et al., 2026; Nguyen et al., 2026; Hossain et al., 2026) absent from released ASH under our nonuniform same-function rescalings. OpenOOD standardizes modular evaluation (Yang et al., 2022; Zhang et al., 2024); Szyc et al. (2023) show that training details can change OOD performance and rankings at similar accuracy. We instead hold classifier function and data fixed and test full-refit Boolean agreement; we do not explain those runs or claim ranking reversals.

Distance-based consumers. Deep nearest-neighbor OOD scoring (Sun et al., 2022) and class-conditional Mahalanobis confidence (Lee et al., 2018) provide established alternatives to activation shaping. Our comparison uses the specified final-layer scores, exact neighbor search, and a common endpoint-local selection and calibration protocol; it is not a reproduction of all components of either published detector. Covariance transformation laws and regularized least squares are classical (Hastie et al., 2009). We use them to derive and test preservation through the complete procedure, without claiming new distance methods or superior detection accuracy.