CONSISTENCY LOSS IS SECRETLY PICKING A SIDE: RISK-DIRECTED CONSENSUS FOR HIGH-STAKES DECISIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Consistency regularization makes predictions agree across semantically equivalent inputs. Under asymmetric decision risk, however, agreement is underspecified: it contracts within-orbit disagreement without determining where the shared prediction settles relative to a risk-critical boundary. In emergency triage, stronger symmetric consistency cut permutation-induced prediction flips by more than two-thirds but also increased under-triage. We formalize this missing degree of freedom as *shared-solution selection*, distinct from agreement contraction, and introduce *disagreement-conditioned risk-directed consensus* (DGN), which biases the direction in which inconsistent views approach consensus while letting that directional force vanish as agreement forms. On three triage corpora, DGN lowers under-triage relative to matched symmetric consistency at less than one point of over-triage per point removed. Across four additional 7B–14B LLM backbones, matched CE→BWOT→DGN interventions reproduce the same two-stage contraction–selection geometry, and DGN lowers both under-triage and flips relative to CE without a corresponding loss of ordinal predictive quality. A frozen-readout intervention roughly halves MIMIC under-triage without relearning the representation, reaching operating points that no train-selectable global threshold attains. A CIFAR-10-LT study with ResNet-32 and unordered labels reproduces both the contraction–selection separation and directional control outside clinical language, LLMs, and ordinal prediction. These results identify consensus location as a distinct and controllable design variable for consistency learning under asymmetric risk.

1 INTRODUCTION

Consistency learning is built on a simple principle: semantically equivalent inputs should produce the same prediction. This principle underlies consistency regularization, perturbation-based training, and architectures designed to be invariant to task-irrelevant transformations (Tarvainen & Valpola, 2017; Xie et al., 2020; Sohn et al., 2020; Wu et al., 2021; Zaheer et al., 2017), and it is especially pressing for language models, whose decisions can change under presentation or ordering variations that preserve much of the underlying task content (Hager et al., 2024; Gourabathina et al., 2025; Liu et al., 2024). Yet agreement alone does not determine which common prediction is selected. This matters when errors on opposite sides of a decision boundary have asymmetric consequences, as in ordinal emergency triage: increasing symmetric transport regularization (BWOT) from $\lambda = 0$ to $\lambda = 6$ reduces permutation-induced prediction flips from 24.65% to 7.55%, yet increases under-triage at the critical boundary from 46.91% to 51.63%, a reversal that holds for every seed (Section 3.2). Stronger agreement can thus improve stability while worsening the risk-relevant decision. This raises a question that standard consistency objectives do not answer: *once equivalent views agree, which shared solution do they converge to?*

We study this missing degree of freedom as *shared-solution selection*. Existing consistency regularizers and invariant architectures constrain variation across equivalent views (Tarvainen & Valpola, 2017; Sohn et al., 2020; Zaheer et al., 2017; Lee et al., 2019), but do not by themselves determine the location of the common prediction, so comparably invariant predictors can make materially different downstream decisions, a structured instance of underspecification (D'Amour et al., 2022). We

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

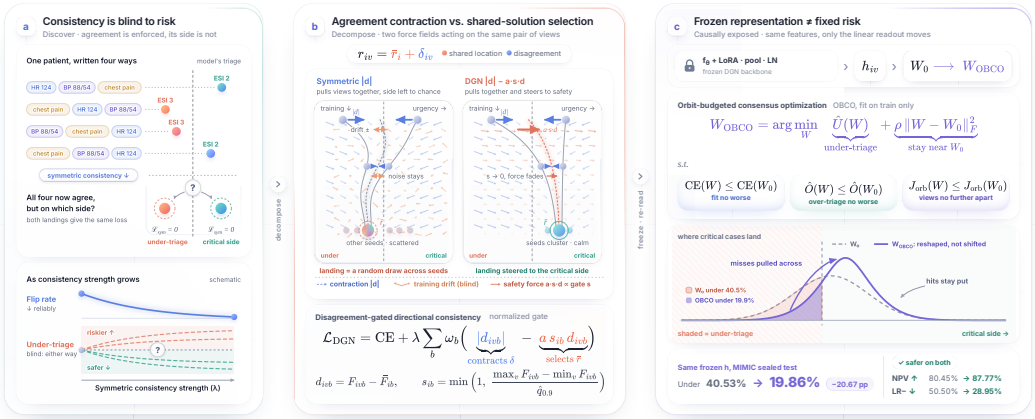


Figure 1: **Consistency fixes agreement but not where the shared solution lands. (a)** Symmetric consistency contracts predictions across equivalent views but vanishes once they agree, regardless of which side of the risk-critical boundary their common prediction occupies. **(b)** Writing the score of view v as $r_{iv} = \bar{r}_i + \delta_{iv}$ separates shared location from within-orbit deviation. Symmetric contraction acts on δ_{iv}, whereas DGN adds a disagreement-gated signed component that tilts consensus toward the critical side while views disagree and disappears at agreement. **(c)** OBCO refits only the linear readout of a frozen representation under predictive, opposite-risk, and orbit-dispersion budgets anchored at W_0. On sealed MIMIC, under-triage falls from 40.53% to 19.86%, NPV rises from 80.45% to 87.77%, and $100 \times \mathrm{LR}^-$ falls from 50.50 to 28.95, at 0.56 points of over-triage per point removed.

therefore distinguish two components of invariance learning: *agreement contraction* reduces differences among predictions within an orbit of equivalent inputs, whereas *shared-solution selection* determines where their common prediction lies relative to a task-relevant boundary. In the two-view ordinal transport studied here, this distinction is exact: symmetric transport acts only on the disagreement mode and exerts exactly zero pseudo-gradient on the common mode (Appendix G.1). Supervision, optimization, and inductive bias may still select a solution implicitly, but the symmetric consistency term itself supplies no signed preference for which invariant solution should survive.

This distinction becomes consequential under asymmetric risk. In emergency triage, under-triaging a high-acuity patient and over-triaging a lower-acuity patient are not interchangeable errors: under-triage remains common in routine practice, and triage guidelines prioritize avoiding it over the resource cost of over-triage (Sax et al., 2023; Newgard et al., 2022; Lupton et al., 2023). Analogous asymmetries arise whenever false negatives and false positives carry different consequences. Cost-sensitive and constrained methods already encode such preferences at the level of labels, prediction losses, or aggregate constraints (Elkan, 2001; Cotter et al., 2019a;b). Our question is different: *can risk preference act directly on how equivalent predictions resolve disagreement and form consensus?*

We introduce *disagreement-conditioned risk-directed consensus* (DGN) to control this process. DGN augments symmetric contraction with a directional component whose strength is set by scale-normalized residual disagreement, so direction acts while equivalent views still disagree and vanishes at agreement, avoiding the persistent shift induced by an always-on directional preference. The novelty is not asymmetric cost itself, but where the asymmetry enters learning: the geometry through which equivalent predictions approach a common solution. We further introduce *OBCO* as a complementary frozen-readout test of the same degree of freedom. Building on the decoupling of representation learning and classifier fitting (Kang et al., 2019), OBCO freezes the learned representation and changes only the final linear readout under predictive, opposite-risk, and orbit-dispersion budgets, asking whether the risk-relevant location of a shared solution remains selectable after representation learning has ended.

Figure 1 summarizes the evidence: symmetric transport (BWOT) suppresses flips yet raises under-triage over cross-entropy on all three triage corpora (panel a); DGN's directional coefficient relocates under-triage by more than thirty points while Flip stays in a narrow band, without the over-steering

of an always-on direction (panel b); and a readout-only change on a frozen representation moves 34.9% of unanimously under-triaged critical orbits entirely to the critical side, relocating agreed decisions rather than only resolving disagreement (panel c; Appendix E.5).

Contributions. (1) We identify *shared-solution selection* as a degree of freedom distinct from agreement contraction and formalize it for two-view ordinal transport (Appendix G): symmetric transport exerts exactly zero force on the common mode of sibling predictions; DGN's directional force vanishes at agreement and preserves the cross-entropy fixed point, whereas an always-on direction shifts it; and critical-boundary crossing, set by orbit width, is separated from under-triage, set by the shared landing. (2) We introduce DGN, a training-time objective that lowers under-triage relative to matched symmetric consistency on all three triage corpora at a below-parity over-triage exchange; among trained objectives, it alone lowers both under-triage and Flip relative to cross-entropy on all three corpora without ordinal collapse. (3) We introduce OBCO, a readout-only convex program with a unique solution that roughly halves MIMIC under-triage at operating points unreachable by global threshold translation, and lowers under-triage on all eight frozen MIMIC representations tested (Appendix E). (4) The contraction–selection geometry replicates across four additional 7B–14B LLM backbones, where DGN lowers both under-triage and Flip relative to cross-entropy without a corresponding loss of ordinal quality (Appendix D), and on CIFAR-10-LT with a ResNet-32, image augmentations, and unordered labels (Appendix F), suggesting that shared-solution selection is not specific to clinical language, LLMs, or ordinal prediction, but is a general design variable for consistency learning under asymmetric risk.

2 Related Work and Problem Positioning

Consistency and invariance. Consistency learning encourages equivalent or augmented views to produce similar predictions, through weight-averaged targets (Tarvainen & Valpola, 2017), perturbation consistency (Xie et al., 2020; Sohn et al., 2020), or agreement between stochastic forward passes (Wu et al., 2021). A complementary line builds invariance into the architecture or inference procedure (Zaheer et al., 2017; Lee et al., 2019); for language models, PINE removes position bias at inference time (Wang et al., 2025), and Set-LLM makes permutation invariance architectural (Egressy & Stühmer, 2025). These approaches determine how strongly, and by what mechanism, equivalent views agree. In medicine, MG-LoRA shows that pure paraphrase consistency admits degenerate agreement and therefore combines it with task supervision (Sadanandan & Behzadan, 2026b); related audits quantify paraphrase sensitivity (Sadanandan & Behzadan, 2026c) and show that consistency can mask unsafe predictions (Sadanandan & Behzadan, 2026a). Our question begins after task relevance is imposed: even among task-relevant predictors with comparable invariance, the common prediction need not occupy the same decision location. This is a structured instance of underspecification, in which predictors with similar aggregate performance differ materially under downstream criteria (D'Amour et al., 2022). We study the degree of freedom that remains once both agreement and task relevance are enforced, namely which invariant solution is selected.

Cost-sensitive and constrained learning. Cost-sensitive learning alters training costs or decision rules according to asymmetric consequences (Elkan, 2001); ordinal methods exploit ordered labels (Cao et al., 2020; Hou et al., 2016; Pedregosa et al., 2017), and OCE models direction- and magnitude-dependent ordinal costs in medical applications (Dvora et al., 2026). Constrained learning expresses requirements such as recall or churn limits directly in the optimization problem (Cotter et al., 2019a;b), and semi-supervised methods combine consistency with cost-sensitive or class-balanced objectives (Rangwani et al., 2022; Lee et al., 2021). Emergency-triage work likewise treats recognition of high-acuity patients and consistent ESI prediction as joint clinical objectives (Shen et al., 2025). These methods act on the relation between a prediction and its target, on aggregate constraints, or on pseudo-label selection; none of them specifies how two equivalent predictions should reach a common value. DGN introduces a different site of intervention: it conditions the asymmetric component of ordinal transport on residual within-orbit disagreement, so direction is active precisely while views disagree and vanishes as agreement is reached. Asymmetric preferences themselves are not new; what DGN changes is where the asymmetry enters learning, namely the geometry of agreement among equivalent predictions.

Frozen readouts. Decoupling representation learning from classifier fitting can substantially change class-level behavior without relearning features (Kang et al., 2019). We do not claim novelty for last-layer retraining and use it only as a diagnostic. OBCO freezes the representation and selects a new readout under predictive, asymmetric-risk, and orbit-level budgets. This tests whether the risk location of an already learned shared solution can still be changed when representation learning is held fixed. As controls, we compare against bias shifts, balanced cost-sensitive refitting, and OBCO without the orbit budget.

3 Agreement Does Not Determine the Shared Decision

3.1 Orbit agreement and risk location

For each clinical case i, let $\mathcal{O}_i = \{x_{iv}\}_{v=1}^{V}$ denote an *orbit* of semantically equivalent inputs: all views contain the same clinical facts and use the same narrative voice, but differ in fact ordering, a presentation change to which LLM clinical decisions are known to be sensitive (Hager et al., 2024; Gourabathina et al., 2025; Liu et al., 2024). They share the same ordinal label $y_i \in \{1, \dots, K\}$, where smaller ESI values indicate greater urgency. Our training implementation uses $V = 2$ sibling permutations per case. Let $p_{iv} \in \Delta^{K-1}$ be the predicted class distribution, and let $F_{ivk} = \sum_{j=1}^{k} p_{ivj}$ denote its cumulative probability at ordinal boundary k.

The ESI algorithm separates patients requiring immediate or high-risk care from those entering the lower-acuity resource-based branch (Eitel et al., 2003; Espejo et al., 2025; Wandl et al., 2025); we therefore use the ESI-$2/3$ cut as the risk-critical boundary. Missing a high-acuity patient is the safety-critical error in triage. It remains common in routine ESI practice (Sax et al., 2023), and triage guidelines prioritize avoiding it over avoiding over-triage (Newgard et al., 2022; Lupton et al., 2023). We therefore evaluate this boundary primarily through under-triage (U), negative predictive value (NPV), and the negative likelihood ratio (LR^{-}), and report over-triage (O) as the resource cost of any relocation. Exact metric definitions, diagnostic conventions, and aggregation rules are given in Appendix A.

The key distinction is between *within-orbit deviation* and the *location of the shared prediction*. We decompose each cumulative prediction as

$$F_{iv} = \bar{F}_i + \delta_{iv}, \qquad \bar{F}_i = \frac{1}{V}\sum_{v=1}^{V} F_{iv}, \qquad \sum_{v=1}^{V} \delta_{iv} = 0.$$

Here δ_{iv} describes disagreement among equivalent views, whereas $\bar{F}_i$ locates their common ordinal prediction. At the critical boundary, the same decomposition becomes $F_{iv,2} = \bar{F}_{i,2} + \delta_{iv,2}$: the deviations describe how much sibling views disagree about criticality, while the shared component determines where the case lies relative to the operating boundary.

This decomposition shows that agreement alone is under-identified: any consistency criterion minimized whenever $F_{i1} = \cdots = F_{iV}$ admits the whole family of zero-disagreement solutions $(F, \dots, F)$ and does not by itself select a member. For the two-view ordinal transport used here, Appendix G.1 makes this exact: symmetric transport has zero pseudo-gradient in the common CDF mode, and the directional term is the only transport component acting on it. Supervision and optimization may select a member implicitly, but a symmetric agreement term supplies no *signed preference* for which side of a risk boundary to favor; predictors with comparable contraction can therefore land at substantially different under–over operating points.

3.2 Stronger consistency can move risk in the wrong direction

We first vary agreement strength without adding any directional preference. BWOT is the symmetric member of the ordinal transport family introduced in Section 4; it contracts sibling cumulative predictions toward a shared target and assigns equal cost to positive and negative drift. We vary only its consistency coefficient λ on MIMIC-4B.

Increasing λ from 0 to 6 reduces prediction flips from 24.65% to 7.55% and critical-boundary crossings from 11.46% to 3.38%, so the consistency objective does what it is designed to do. Over the

same sweep, however, under-triage rises from 46.91% to 51.63%: as the model becomes more consistent, more high-acuity patients are assigned to the lower-acuity side.

This reversal is not a tuning artifact. Flip is non-increasing in λ, under-triage is higher at $\lambda = 6$ than at $\lambda = 0$ for every seed, and no $\lambda > 0$ on the grid lowers mean under-triage below the unregularized endpoint (Appendix G.3). The decision geometry also accounts for it exactly, which makes it more specific than the general observation that stability and accuracy can differ. For a two-view orbit with critical-boundary margins r_{i1}, r_{i2} (positive on the high-acuity side), let $m_i = (r_{i1} + r_{i2})/2$ be its landing and $\Delta_i = |r_{i1} - r_{i2}|/2$ its half-width. With X_i the indicator that the two views cross the boundary and $\mathcal{H}$ the set of high-acuity orbits, Theorem G.6 gives

$$X_i = \mathbf{1}[\,|m_i| < \Delta_i\,], \qquad \Big| U - \Pr_{i\in\mathcal{H}}(m_i < 0) \Big| \le \tfrac{1}{2} \Pr_{i\in\mathcal{H}}(X_i = 1). \tag{1}$$

Symmetric contraction shrinks Δ_i and hence crossing. As crossing vanishes, however, under-triage converges to the fraction of high-acuity orbits whose landing lies on the lower-acuity side, and a symmetric agreement term supplies no signed preference on this quantity. Stronger agreement therefore makes risk depend increasingly on where the shared prediction lands, while leaving that location unselected.

3.3 Direction and contraction are distinguishable controls

We next fix the consistency strength at $\lambda = 3$ and vary only DGN's directional coefficient a. As a increases from 0 to 0.9, under-triage falls from 49.51% to 16.37% while prediction flips stay within 11.40–15.01%: direction moves the critical-boundary operating point by more than thirty points without a comparable change in orbit disagreement, and for every $a \le 0.75$ each point of under-triage removed costs less than one point of over-triage. The main configuration ($a = 0.5$) was frozen before any test set was loaded and not re-tuned after this sweep, so the main-table results are conservative: they do not use the best point on this curve. Indexed by the directional scale λa, the λ-sweep points at $a = 0.5$ with $\lambda a \in \{0.5, 1\}$ are bracketed by a-sweep points, consistent with the common-mode force $-\lambda a\, \omega_k g_{ik}$ in Proposition 4.1, and varying the gate quantile $q \in [0.50, 0.95]$ moves the operating point smoothly along the same curve (Appendix G.3). Agreement contraction and solution-location selection are thus empirically distinguishable control directions, and the same two-stage pattern reappears on four additional LLM backbones (Section 5.2).

4 Risk-Directed Solution Selection

We make shared-solution location an explicit object of optimization at two complementary stages: **DGN** acts during training and controls the direction in which residual disagreement is resolved, while **OBCO** acts after representation learning and relocates the shared solution using only the final linear readout.

4.1 Disagreement-conditioned directional transport

We first define a common ordinal-transport family containing symmetric contraction, fixed-direction transport, and DGN. For view v of training example i, let p_{ivj} be the predicted probability of ordinal class j and $F_{ivk} = \sum_{j\le k} p_{ivj}$ its cumulative probability at boundary k. We estimate the reliability of each sibling using its expected ordinal error, $e_{iv} = \sum_{j=1}^{K} p_{ivj}|j - y_i|$, and assign the stop-gradient weight $\alpha_{iv} = \mathrm{softmax}_v(-\mathrm{sg}(e_{iv})/\tau)$. The resulting sibling teacher is $T_{ik} = \mathrm{sg}(\sum_v \alpha_{iv} F_{ivk})$, and the signed deviation from that teacher is $d_{ivk} = F_{ivk} - T_{ik}$. Thus, the teacher is constructed entirely from sibling predictions observed during supervised training and introduces no additional inference-time model.

Because our primary clinical boundary is ESI-2/3, ordinal boundaries are weighted by $\omega_k \propto \exp(-(k-2)^2/(2\sigma^2))$, normalized so that $\sum_k \omega_k = 1$. Symmetric transport penalizes $|d_{ivk}|$ and therefore contracts sibling predictions without preferring either direction. A fixed directional term can instead favor one sign of d, but applying the same asymmetry after disagreement has largely disappeared can over-steer the operating point.

DGN therefore conditions the asymmetry on the disagreement that remains. At each ordinal boundary, the sibling spread is $s_{ik} = \max_v F_{ivk} - \min_v F_{ivk}$. We normalize this quantity using an

exponential moving scale $m_t = \beta m_{t-1} + (1-\beta) Q_q(\{s_{ik}\}_{k=1}^{K-1})$, maintained separately for each narrative voice, and define the detached gate $g_{ik} = \min(1, s_{ik}/m_t)$. The training objective is

$$\mathcal{L}_{\mathrm{DGN}} = \mathcal{L}_{\mathrm{CE}} + \lambda \frac{1}{V} \sum_{v=1}^{V} \sum_{k=1}^{K-1} \omega_k \left(|d_{ivk}| - a\, g_{ik} d_{ivk} \right). \tag{2}$$

Equation 2 contains three useful regimes. Setting $a = 0$ recovers symmetric BWOT, while setting $g_{ik} \equiv 1$ gives fixed-direction OQT. DGN instead lets residual disagreement determine how much directional preference is active. Because smaller ESI indices denote greater urgency, increasing F_{ivk} moves probability mass toward the urgent side. Hence, for $d > 0$ the effective slope is $1 - ag$, whereas for $d < 0$ it is $1 + ag$: less-urgent drift is penalized more strongly than equally sized more-urgent drift. For $0 \le a \le 1$ and $0 \le g \le 1$, both slopes remain nonnegative, so the asymmetric term changes the direction of contraction without reversing contraction itself. Moreover, in our two-view setting the sibling teacher is a convex combination of the two predictions, so $|d_{ivk}| \le s_{ik}$. Since $0 \le g_{ik} \le 1$, the directional component satisfies $|a\, g_{ik} d_{ivk}| \le a\, s_{ik} \to 0$ as siblings agree; the directional force therefore vanishes at agreement even though the normalized gate itself need not converge to zero as its reference scale adapts. This construction admits an exact mode decomposition. Write the two sibling CDFs as $F_{i1k} = \bar{F}_{ik} + \delta_{ik}$ and $F_{i2k} = \bar{F}_{ik} - \delta_{ik}$, and let $\mathcal{R}_i$ denote the transport term of Eq. 2, differentiated with the stop-gradient teacher and gate held fixed ($\widetilde{\partial}$).

Proposition 4.1 (Contraction and selection act on different modes; informal)**.** *For $V = 2$ and any teacher weights $\alpha_{iv} \in (0,1)$,*

$$\frac{\widetilde{\partial}\, \mathcal{R}_i}{\partial \delta_{ik}} = \lambda\, \omega_k \operatorname{sign}(\delta_{ik}), \qquad \frac{\widetilde{\partial}\, \mathcal{R}_i}{\partial \bar{F}_{ik}} = -\lambda\, a\, \omega_k\, g_{ik}. \tag{3}$$

Thus symmetric BWOT ($a = 0$) exerts exactly zero common-mode force, and the directional term is the only transport force that can relocate the shared prediction. For DGN this force vanishes at agreement ($g_{ik} = 0$), so the cross-entropy fixed point is preserved; fixed-direction OQT ($g_{ik} \equiv 1$) retains the force $-\lambda a \omega_k$ after agreement and shifts that fixed point toward the urgent classes.

Appendix G gives the exact statements and proofs (Theorem G.2 and Corollary G.5) together with the corresponding single-boundary resolution dynamics (Proposition G.4). Unless otherwise stated, we use $\tau = 0.05$, $\sigma = 1$, $q = 0.9$, $\beta = 0.99$, $\lambda = 3$, and $a = 0.5$.

4.2 Orbit-budgeted readout selection

DGN controls where agreement forms during training. We next ask whether the risk location remains selectable once representation learning has ended. Let $h_{iv} = [\mathrm{LN}(\mathrm{meanpool}(f_\theta(x_{iv}))); 1]$ denote the frozen feature with an appended bias coordinate. If W_0 is the learned linear classifier, then $z_{iv}(W) = W h_{iv}$. OBCO freezes the backbone, adapters, pooling operation, and LayerNorm, and modifies only W.

Let $\mathcal{I}_H = \{i : y_i \le 2\}$ contain high-acuity examples and $\mathcal{I}_L = \{i : y_i > 2\}$ the remaining examples. For competing labels S, define the boundary-specific surrogate $\ell_S(z, y) = (\log 2)^{-1} \log(1 + \sum_{k \in S} e^{z_k - z_y})$. The under-triage objective is the average of $\ell_{\{3,\dots,K\}}$ over $i \in \mathcal{I}_H$ and all views, yielding $U_{\mathrm{sur}}(W)$; analogously, averaging $\ell_{\{1,2\}}$ over $i \in \mathcal{I}_L$ gives the over-triage surrogate $O_{\mathrm{sur}}(W)$. Unlike a binary probability difference, these surrogates compare the gold logit directly against every competing class on the opposite side of the critical boundary.

We additionally constrain changes to orbit behavior. Let $P = I - \mathbf{1}\mathbf{1}^\top / K$ remove the softmax-invariant common-logit component, and let $\bar{z}_i^P = V^{-1} \sum_u P z_{iu}(W)$. We measure within-orbit dispersion by $J(W) = (NVK)^{-1} \sum_{i,v} \|P z_{iv}(W) - \bar{z}_i^P\|_2^2$. Let $C(W)$ denote mean multiclass cross-entropy, and abbreviate $C_0 = C(W_0)$, $O_0 = O_{\mathrm{sur}}(W_0)$, and $J_0 = J(W_0)$. OBCO then solves

$$\begin{aligned} \min_{W} \quad & U_{\mathrm{sur}}(W) + \frac{\rho}{2} \|W - W_0\|_F^2 \\ \text{s.t.} \quad & C(W) \le C_0, \qquad O_{\mathrm{sur}}(W) \le O_0, \qquad J(W) \le J_0. \end{aligned} \tag{4}$$

The three budgets allow no increase over the original head in overall predictive loss, opposite-side risk, or within-orbit dispersion. The ridge term, with $\rho = 10^{-3}$ by default, keeps the readout close to W_0 within this feasible set. OBCO therefore searches for another shared-solution location only inside these budgets, instead of refitting the head toward a more aggressive operating point.

We solve Eq. 4 with an augmented-Lagrangian method on the scale-normalized constraints $g_r(W) = G_r(W)/G_r(W_0) - 1 \leq 0$, $G_r \in \{C, O_{\mathrm{sur}}, J\}$, using L-BFGS-B for the inner problem (Hestenes, 1969; Powell, 1969; Byrd et al., 1995), and retain a candidate only once all normalized violations fall below a fixed numerical tolerance. All objectives and budgets are computed from the training feature cache; sealed-test labels are used only for evaluation.

Eq. 4 has a nonempty convex feasible set and, for $\rho > 0$, a unique global minimizer (Appendix G.6).

5 Experiments

We test whether explicit solution selection moves the shared prediction under a fixed, sealed evaluation protocol, and at what over-triage exchange rate. Section 5.2 summarizes the cross-backbone and cross-domain replications, with details in Appendices D and F.

5.1 Main Comparison on Sealed Test Sets

Protocol. We evaluate on three triage corpora with different label sources: MIMIC (ESI recorded during emergency-department stays), EMS-M (four-level weak labels mapped from EMS protocols), and ER-Reason (acuity annotations mapped onto the ESI scale). Each case is written in two narrative voices, each voice defining an orbit of fact permutations; the test sets contain 503, 883, and 500 source cases ($5{,}276$, $8{,}830$, and $5{,}000$ permutation rows). All methods share one backbone (Llama-3.1-Nemotron-Nano-4B-v1.1), rank-16 QLoRA adapters (Hu et al., 2021; Dettmers et al., 2023), two training epochs, and two sibling views per training case. Every configuration was frozen and hashed before any test set was loaded, and we report five-seed means. At the ESI-$2/3$ boundary, under-triage (U), NPV, and LR^- are the pre-specified primary safety endpoints, and the resource cost of each relocation is summarized by the exchange rate $\eta = \Delta O/|\Delta U|$, the over-triage incurred per point of under-triage removed. Because symmetric summaries weight both errors equally, contrary to the asymmetric-risk premise of this paper, we keep the two errors separate; Table 1 reports the primary endpoints together with QWK and Flip, and exact definitions are given in Appendix A.

Compared methods. Plain cross-entropy (CE) is the common anchor. Label-side cost control is represented by inverse-prior weighted cross-entropy (WCE, $\rho = 1$) and the ordinal cost-sensitive loss OCE (Dvora et al., 2026). Agreement- and invariance-side controls are MG-LoRA (Sadanandan & Behzadan, 2026b), adapted to text-only permutation views; PINE (Wang et al., 2025), with strict fact canonicalization; Set-LLM (Egressy & Stühmer, 2025), which enforces permutation invariance architecturally; and SDTA, the domain-adaptation-then-tuning pipeline of Shen et al. (2025). PINE and Set-LLM have zero Flip by construction; we mark these entries and do not rank them against learned consistency. Within the transport family of Eq. 2, BWOT ($a = 0$) and DGN share $\lambda = 3$, isolating disagreement-conditioned direction at matched consistency strength; the ungated ablation OQT ($g \equiv 1$) shares both λ and a with DGN and isolates the gate. OBCO is applied to the frozen DGN representation of the same run, so DGN + OBCO differs from DGN only in its linear readout.

Exact or stronger agreement does not determine the operating point. Agreement- and invariance-side controls consistently reduce permutation sensitivity relative to CE, but their effect on risk location is inconsistent. On MIMIC, MG-LoRA and SDTA reduce Flip from 24.98% to 19.28% and 17.59%, and PINE and Set-LLM reach Flip $= 0$ by construction, yet their under-triage ranges from 42.16% (PINE) to 48.42% (MG-LoRA), on both sides of CE's 44.73%. On ER-Reason, both exact-invariance methods have zero Flip, but PINE under-triages 46.36% of high-acuity rows and Set-LLM 51.78%. Exact invariance thus guarantees that equivalent views agree, but it does not determine where their common prediction lies. Strengthening agreement shows the same pattern. BWOT reduces MIMIC Flip to 10.54% but raises U on all three datasets (MIMIC $44.73\% \rightarrow 49.96\%$, EMS-M $16.64\% \rightarrow 17.17\%$, ER-Reason $46.67\% \rightarrow 49.89\%$), which reproduces the λ-sweep of Section 3.2 on the sealed test sets. Label-side cost acts on a different axis. WCE lowers MIMIC U to 30.67%, but its Flip stays at the CE level on MIMIC and EMS-M and

Table 1: Sealed-test comparison on three triage datasets using Llama-3.1-Nemotron-Nano-4B (five-seed mean). Entries are shown as $100\times$ the underlying metric; definitions are given in Appendix A. For U, NPV, LR^-, and QWK, best is shown in **bold** and second best is underlined. Flip is reported descriptively because exact-invariance methods can attain structural zeros. The gray row is the ungated ablation of DGN ($g \equiv 1$ in Eq. 2) and is not ranked. †Ours. ‡Flip is zero by construction. BWOT is the symmetric member ($a = 0$) of the same transport family.

	MIMIC					EMS-M					ER-Reason				
Method	$U\downarrow$	NPV↑	$\mathrm{LR}^-\downarrow$	QWK↑	Flip↓	$U\downarrow$	NPV↑	$\mathrm{LR}^-\downarrow$	QWK↑	Flip↓	$U\downarrow$	NPV↑	$\mathrm{LR}^-\downarrow$	QWK↑	Flip↓
CE	44.73	79.83	52.52	56.17	24.98	16.64	90.77	17.20	80.36	5.45	46.67	69.78	61.85	27.95	7.10
WCE	30.67	84.01	39.59	**56.40**	24.95	16.42	90.81	17.13	78.57	5.49	41.39	69.28	63.43	21.23	16.40
OCE	45.37	79.68	53.04	55.91	23.79	16.43	90.88	16.99	80.72	5.42	46.86	69.68	62.13	27.54	7.63
MG-LoRA	48.42	78.61	56.56	54.07	19.28	16.47	90.85	17.05	80.11	5.36	47.54	69.58	62.45	27.76	7.30
SDTA	45.53	79.45	53.80	55.90	17.59	16.65	90.79	17.16	80.60	4.27	46.67	69.43	62.83	27.12	8.70
PINE	42.16	80.23	51.23	55.08	0.00‡	15.96	91.10	16.52	80.71	0.00‡	46.36	70.00	61.26	**28.93**	0.00‡
Set-LLM	46.44	79.28	54.33	54.59	0.00‡	16.11	91.01	16.73	80.30	0.00‡	51.78	67.68	68.17	23.32	0.00‡
BWOT	49.96	78.24	57.81	54.54	10.54	17.17	90.50	17.77	79.66	4.89	49.89	68.59	65.42	25.36	5.37
DGN†	40.53	80.45	50.50	51.37	14.21	14.28	91.88	14.95	**81.19**	3.81	36.23	72.36	54.54	27.46	6.60
DGN + OBCO†	**19.86**	**87.77**	**28.95**	54.30	28.53	**11.19**	**93.34**	**12.07**	80.19	6.38	**20.16**	**75.09**	**47.39**	20.56	14.40
DGN w/o gate (OQT)	18.07	85.46	35.45	28.20	14.45	12.94	92.51	13.71	80.24	3.53	18.07	78.73	38.64	24.64	6.93

more than doubles on ER-Reason (16.40%), and OCE stays within one point of CE in U on every dataset. None of these mechanisms by itself determines the risk location of the shared prediction.

Direction moves the consensus; conditioning controls how far. At matched consistency strength, DGN lowers U relative to BWOT by 9.43, 2.90, and 13.66 points on MIMIC, EMS-M, and ER-Reason, improves NPV and LR^- on all three, keeps Flip below CE on every dataset, and attains the best EMS-M QWK (81.19%), mirroring the mostly vertical selection step in Figure 2c. Its over-triage cost stays below parity ($\eta = 0.65$, 0.39, 0.72), an exchange in the direction triage guidelines endorse (Newgard et al., 2022; Lupton et al., 2023). The matched ungated ablation OQT ($g \equiv 1$) shows the opposite failure: on MIMIC its over-triage rises from 13.59% (BWOT) to 49.23% and QWK falls from 54.54% to 28.20% ($\eta = 1.12$; 0.94 on ER-Reason); only on EMS-M, where low baseline U leaves little room to overshoot, is it within 1.4 points of DGN on every metric. A raw gate $g_{\mathrm{raw}} = \min(1, s)$ is a negative control: it stays within 1.17 points of BWOT in U on all three datasets, while normalization lowers U by a further 9.41, 2.47, and 12.49 points at much smaller Flip changes, so normalization is what unlocks relocation (Appendix G.4). Among trained objectives, only DGN lowers both U and Flip relative to CE on all three datasets without ordinal collapse, at no inference cost; PINE needs roughly $12\times$ the evaluation time (Table 2) for U reductions of at most 2.6 points.

Readout selection relocates risk on a frozen representation. With the DGN representation frozen, OBCO lowers MIMIC under-triage from 40.53% to 19.86% at $\eta = 0.56$, attains the best NPV (87.77%) and $100 \times \mathrm{LR}^-$ (28.95), and raises QWK from 51.37% to 54.30%; it also ranks first in U, NPV, and LR^- on EMS-M and ER-Reason, where U falls from 36.23% to 20.16%. The effect is not specific to DGN: OBCO lowers U and LR^- on all eight frozen MIMIC representations tested (Appendix E.1), including the exactly invariant PINE representation, which has no cross-view disagreement to resolve yet whose U drops from 42.16% to 20.89%, with 55 of 152 unanimously under-triaged critical cases becoming correct in every view. Nor is it a threshold shift: critical-boundary AUROC rises from 81.24% to 82.83%, which a global critical-bias shift leaves unchanged, and on MIMIC no train-selectable global threshold is at least as good in both U and O, even when OBCO is restricted to a rank-one change $W_0 + c\,u^\top$ (Appendix E.4); balanced cost-sensitive refitting does not reproduce it, and removing the ridge term changes endpoints by only tenths of a point (Appendix E.2). The higher Flip of DGN + OBCO reflects these repairs: on the development split, 34.9% of unanimously under-triaged critical orbits move entirely to the critical side and another 24.2% are partially corrected, which Flip counts as disagreement (Appendix E.5); without the orbit budget J, Flip nearly doubles to 53.64% and over-triage rises (Appendix E.3). Shared-solution location therefore remains selectable after representation learning, and the effect

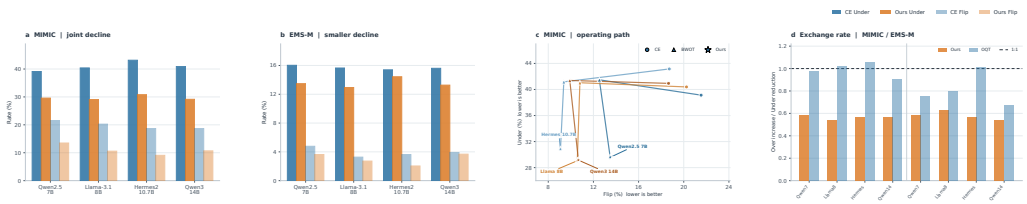


Figure 2: **The contraction–selection geometry replicates across LLM backbones. (a,b)** Under-triage and permutation Flip for CE and DGN (Ours) on MIMIC and EMS-M across four backbones. **(c)** Matched CE→BWOT→DGN trajectories on MIMIC in the Flip–Under plane: symmetric contraction (CE→BWOT) moves predominantly left, and disagreement-gated selection (BWOT→DGN, same $\lambda = 3$) moves predominantly down. **(d)** Additional over-triage per point of under-triage removed relative to CE; DGN stays well below the dashed 1:1 line in every backbone–dataset cell, whereas fixed-direction OQT approaches or exceeds it.

does not reduce to disagreement resolution, cost-sensitive refitting, or a one-dimensional threshold shift.

5.2 Replication Across Backbones and Modalities

The two control directions reappear on every backbone. We repeat the matched CE→BWOT→DGN intervention with frozen hyperparameters on Qwen2.5-7B, Llama-3.1-8B, Hermes2-10.7B, and Qwen3-14B (Figure 2; details in Appendix D). On MIMIC, the contraction step lowers Flip by 8.74–9.45 points and changes under-triage by $+0.36$ points on average. The selection step lowers under-triage by 10.36–12.23 points and changes Flip by $+0.30$ points on average. Relative to CE, both quantities improve: four-backbone mean under-triage falls from 40.90% to 29.65% and Flip from 19.79% to 10.95%, while QWK stays within -0.56 to $+0.23$ points of CE. EMS-M shows the same pattern at a smaller scale. The gate again determines the exchange rate. Relative to CE, DGN incurs 0.54–0.63 points of over-triage per point of under-triage removed in all eight backbone–dataset cells, compared with 0.68–1.06 for fixed-direction OQT.

The separation also holds across modality, architecture, and label geometry. On CIFAR-10-LT with a ResNet-32, image augmentations as equivalent views, and the unordered tail class C9 as the pre-specified target (Appendix F), the same separation appears over five seeds. Strengthening symmetric consistency (CE→SYM-4) lowers Flip from 27.01% to 19.84% but also C9 recall from 57.67% to 50.64%. With the contraction mechanism held fixed, direction raises C9 recall from 55.47% at $a = 0$ to 58.29% at $a = 0.5$ while Flip stays within 23.64–25.11%, controllably though not monotonically in a; replacing the normalized gate with the raw spread drops C9 recall to 52.64% with Flip nearly unchanged.

6 Conclusion

Consistency learning is usually evaluated by how much equivalent views agree. Under asymmetric risk, it must also be evaluated by where they agree. We showed that agreement contraction and shared-solution selection are distinct. A symmetric agreement term exerts no common-mode force, so the landing of the shared prediction is left to supervision and optimization, and stronger agreement can move it in the wrong direction. This is a concrete case of underspecification (D'Amour et al., 2022). Across backbones, under-triage and permutation sensitivity change substantially while ordinal quality stays essentially the same, so both vary along directions that standard supervision leaves unconstrained. DGN makes the landing an explicit training objective whose directional strength is calibrated by relative, scale-normalized disagreement, vanishes at agreement, and adds no inference-time cost. OBCO shows that the landing can still be selected on a frozen representation. The same separation on CIFAR-10-LT suggests that the principle extends to any setting where invariance is imposed under asymmetric costs: *invariance specifies agreement, but the shared solution still has to be selected.*

Ethics Statement

This work uses de-identified clinical records from MIMIC-IV-ED and ER-Reason, accessed through PhysioNet after completing the required training and signing the corresponding data use agreements; no restricted records or derived patient-level files are redistributed. EMS-M is derived from the synthetic EMSDialog corpus (CC BY-NC 4.0), and CIFAR-10 is publicly available. The models studied here are research prototypes for analyzing consistency learning under asymmetric risk and are not intended for clinical deployment; the reported reductions in under-triage come with increased over-triage and would require prospective clinical validation before any real-world use.

AI Use Statement

We used generative AI tools only to edit the manuscript for grammar, clarity, and readability. The research ideas, methodology, theoretical results, data processing, and interpretation of results were produced entirely by the authors without generative AI. All AI-assisted edits were reviewed by the authors, who take full responsibility for the content of this paper.

References

Patrick M Bossuyt, Johannes B Reitsma, David E Bruns, Constantine A Gatsonis, Paul P Glasziou, Les Irwig, Jeroen G Lijmer, David Moher, Drummond Rennie, Henrica CW De Vet, et al. Stard 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Radiology*, 277 (3):826–832, 2015.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition*, pp. 3121–3124. IEEE, 2010.

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020.

Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213, 1968.

Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In *International Conference on Machine Learning*, pp. 1397–1405. PMLR, 2019a.

Andrew Cotter, Heinrich Jiang, Maya Gupta, Serena Wang, Taman Narayan, Seungil You, and Karthik Sridharan. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, 20(172):1–59, 2019b.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9260–9269. IEEE, 2019.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36:10088–10115, 2023.

Tal Dvora, Rotem Haba, and Gonen Singer. Deep neural networks with ordinal loss for medical applications. *arXiv preprint arXiv:2606.25769*, 2026.

Orestis Efthimiou, Michael Seo, Konstantina Chalkou, Thomas Debray, Matthias Egger, and Georgia Salanti. Developing clinical prediction models: a step-by-step guide. *BMJ*, 386, 2024.

Beni Egressy and Jan Stühmer. Set-llm: A permutation-invariant llm. *Advances in Neural Information Processing Systems*, 38:62798–62834, 2025.

David R Eitel, Debbie A Travers, Alexander M Rosenau, Nicki Gilboy, and Richard C Wuerz. The emergency severity index triage algorithm version 2 is reliable and valid. *Academic Emergency Medicine*, 10(10):1070–1080, 2003.

Charles Elkan. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence*, volume 17, pp. 973–978. Lawrence Erlbaum Associates Ltd, 2001.

MD Tanguy Espejo, Florian F Grossmann, MD Henk B Riedel, MD Roland Bingisser, and MD Christian H Nickel. The emergency severity index (esi) version 5: simulation of predictive validity and triage level distribution. *The Journal of Emergency Medicine*, 78:57–70, 2025.

Xueren Ge, Sahil Murtaza, Anthony Cortez, and Homa Alemzadeh. Emsdialog: Synthetic multi-person emergency medical service dialogue generation from electronic patient care reports via multi-llm agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 35081–35110, 2026.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Abinitha Gourabathina, Walter Gerych, Eileen Pan, and Marzyeh Ghassemi. The medium is the message: how non-clinical information shapes clinical decisions in llms. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1805–1828, 2025.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30(9):2613–2622, 2024.

Magnus R Hestenes. Multiplier and gradient methods. *Journal of Optimization Theory and Applications*, 4(5):303–320, 1969.

Le Hou, Chen-Ping Yu, and Dimitris Samaras. Squared earth mover's distance-based loss for training deep neural networks. *arXiv preprint arXiv:1611.05916*, 2016.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

Hyuck Lee, Seungjae Shin, and Heeyoung Kim. Abc: Auxiliary balanced classifier for class-imbalanced semi-supervised learning. *Advances in Neural Information Processing Systems*, 34: 7082–7094, 2021.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pp. 3744–3753. PMLR, 2019.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Joshua R Lupton, Cynthia Davis-O'Reilly, Rebecca M Jungbauer, Craig D Newgard, Mary E Fallat, Joshua B Brown, N Clay Mann, Gregory J Jurkovich, Eileen Bulger, Mark L Gestring, et al. Under-triage and over-triage using the field triage guidelines for injured patients: a systematic review. *Prehospital emergency care*, 27(1):38–45, 2023.

Nikita Mehandru, Niloufar Golchini, David Bamman, Travis Zack, Melanie F Molina, and Ahmed Alaa. Er-reason: A benchmark dataset for llm-based clinical reasoning in the emergency room. *arXiv e-prints*, pp. arXiv–2505, 2025.

Craig D Newgard, Peter E Fischer, Mark Gestring, Holly N Michaels, Gregory J Jurkovich, E Brooke Lerner, Mary E Fallat, Theodore R Delbridge, Joshua B Brown, Eileen M Bulger, et al. National guideline for the field triage of injured patients: recommendations of the national expert panel on field triage, 2021. *The journal of trauma and acute care surgery*, 93(2):e49, 2022.

Fabian Pedregosa, Francis Bach, and Alexandre Gramfort. On the consistency of ordinal regression methods. *Journal of Machine Learning Research*, 18(55):1–35, 2017.

Michael JD Powell. A method for nonlinear constraints in minimization problems. *Optimization*, pp. 283–298, 1969.

Harsh Rangwani, Sho Takemori, Kato Takashi, Yuhei Umeda, et al. Cost-sensitive self-training for optimizing non-decomposable metrics. *Advances in Neural Information Processing Systems*, 35: 26994–27007, 2022.

Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020.

Binesh Sadanandan and Vahid Behzadan. Consistent but dangerous: Per-sample safety classification reveals false reliability in medical vision-language models. *arXiv preprint arXiv:2603.20985*, 2026a.

Binesh Sadanandan and Vahid Behzadan. Mechanistically guided lora improves paraphrase consistency in medical vision-language models. *arXiv preprint arXiv:2603.00148*, 2026b.

Binesh Sadanandan and Vahid Behzadan. Psf-med: Measuring and explaining paraphrase sensitivity in medical vision language models. *arXiv preprint arXiv:2602.21428*, 2026c.

Dana R Sax, E Margaret Warton, Dustin G Mark, David R Vinson, Mamata V Kene, Dustin W Ballard, Tina J Vitale, Katherine R McGaughey, Aaron Beardsley, Jesse M Pines, et al. Evaluation of version 4 of the emergency severity index in us emergency departments for the rate of mistriage. *JAMA Network Open*, 6(3):e233404, 2023.

Qingyang Shen, Xiaozhi Zhang, Haomin Ren, Quan Guo, and Zhang Yi. Knowledge-embedded large language models for emergency triage. *Knowledge-Based Systems*, 318:113431, 2025.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33: 596–608, 2020.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30, 2017.

Bettina Wandl, Jan D Kellerer, Verena Fuhrmann, Karina Tapinova, Dominik Roth, and Gerhard Müller. Diagnostic test accuracy of the emergency severity index: a systematic review and meta-analysis. *European Journal of Emergency Medicine*, 32(5):325, 2025.

Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham Kakade, Hao Peng, and Heng Ji. Eliminating position bias of language models: A mechanistic approach. In *International Conference on Learning Representations*, volume 2025, pp. 37090–37117, 2025.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905, 2021.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268, 2020.

William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander Smola. Deep sets. *Advances in Neural Information Processing Systems*, 30, 2017.

A Metric Definitions and Evaluation Protocol

Clinical boundary and notation. We treat ESI levels 1–2 as the high-acuity class and ESI levels 3+ as the lower-acuity class. This cut follows the structure of the ESI algorithm, which first identifies patients requiring immediate or high-risk care before entering the lower-acuity resource-based branch (Eitel et al., 2003; Espejo et al., 2025). Diagnostic evaluations of ESI likewise treat discrimination of high-acuity patients as a central operating characteristic (Wandl et al., 2025). Let high acuity be the positive class and define

$$TP = \#\{y \le 2, \hat{y} \le 2\}, \qquad FN = \#\{y \le 2, \hat{y} > 2\},$$
$$TN = \#\{y > 2, \hat{y} > 2\}, \qquad FP = \#\{y > 2, \hat{y} \le 2\}.$$

The diagnostic quantities used below are established measures rather than task-specific composite scores. Sensitivity, specificity, predictive values, and likelihood ratios are standard quantities in diagnostic-accuracy reporting (Bossuyt et al., 2015; Efthimiou et al., 2024); balanced accuracy is a standard class-balanced classification measure (Brodersen et al., 2010); and weighted kappa is a classical measure of agreement that permits graded disagreement (Cohen, 1968). The ESI literature motivates the clinical boundary, whereas the diagnostic-accuracy literature supplies the evaluation measures; among these, we treat U, NPV, and LR^- as primary safety endpoints.

Under- and over-triage. Under-triage is the false-negative rate at the ESI-$2/3$ boundary,

$$U = \frac{FN}{TP + FN} = \Pr(\hat{y} > 2 \mid y \le 2),$$

whereas over-triage is the corresponding false-positive rate,

$$O = \frac{FP}{TN + FP} = \Pr(\hat{y} \le 2 \mid y > 2).$$

Equivalently,

$$\text{Sensitivity} = 1 - U, \qquad \text{Specificity} = 1 - O.$$

Both U and O are class-conditional error rates. Their sum is therefore not an overall mistriage rate. If $\pi = \Pr(y \le 2)$ denotes the prevalence of high-acuity cases, the overall binary error rate at this boundary is

$$\mathrm{Err}_{2/3} = \pi U + (1 - \pi) O.$$

We report U and O separately because collapsing them with a task-specific weight would require an assumed clinical weighting that is not part of our evaluation protocol; the observed trade-off is instead summarized by the exchange rate η defined below.

Exchange rate and balanced accuracy. When a method relocates the operating point relative to a reference, we summarize the cost of the relocation by the exchange rate

$$\eta = \frac{\Delta O}{|\Delta U|},$$

the additional over-triage incurred per point of under-triage removed. Balanced accuracy at the critical boundary,

$$\mathrm{BA}_2 = \frac{\text{Sensitivity} + \text{Specificity}}{2} = 1 - \frac{U + O}{2},$$

is the standard binary balanced-accuracy statistic (Brodersen et al., 2010) and an affine transformation of Youden's index (Youden, 1950). For a relocation that lowers U, $\Delta\mathrm{BA}_2 = (|\Delta U| - \Delta O)/2$, so $\eta < 1$ holds exactly when BA_2 improves; η therefore carries the same ordering information while keeping the two error types separate. Because BA_2 weights both errors equally, contrary to the asymmetric-risk premise of this paper, we use it only as a secondary check against degenerate collapse (e.g., the U-only readout in Table 4, which attains $U = 0$ with $\mathrm{BA}_2 = 50\%$).

Negative predictive value and negative likelihood ratio. When a lower-acuity prediction ($\hat{y} > 2$) is interpreted as a negative triage decision, its negative predictive value is

$$\mathrm{NPV} = \frac{TN}{TN + FN}.$$

NPV answers the clinically interpretable question: among cases predicted to be lower acuity, what fraction are truly below the high-acuity boundary? Because predictive values depend on class prevalence, we additionally report the negative likelihood ratio

$$\mathrm{LR}^- = \frac{1 - \text{Sensitivity}}{\text{Specificity}} = \frac{U}{1 - O}.$$

Smaller LR^- indicates stronger evidence against high acuity after a lower-acuity prediction. NPV and likelihood ratios are standard diagnostic-accuracy quantities (Bossuyt et al., 2015; Efthimiou et al., 2024).

Full-scale classification and ordinal metrics. Boundary metrics do not capture errors occurring entirely within the ESI-1–2 or ESI-3+ regions. We therefore retain conventional full-label metrics as complementary measures. For M evaluated prediction rows and K ordinal classes,

$$\mathrm{Accuracy} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}[\hat{y}_m = y_m],$$

and macro recall is

$$\mathrm{MacroR} = \frac{1}{K}\sum_{k=1}^{K} \frac{TP_k}{TP_k + FN_k}.$$

Mean absolute error preserves the ordinal distance between predicted and gold ESI levels,

$$\mathrm{MAE} = \frac{1}{M}\sum_{m=1}^{M} |\hat{y}_m - y_m|.$$

We also report quadratic weighted kappa (QWK) to measure full-scale ordinal agreement. Let A_{ab} denote the normalized observed confusion matrix between gold class a and prediction b, and let E_{ab} denote the expected matrix obtained from the outer product of their marginals. Using quadratic disagreement weights

$$w_{ab} = \frac{(a-b)^2}{(K-1)^2},$$

QWK is

$$\mathrm{QWK} = 1 - \frac{\sum_{a=1}^{K}\sum_{b=1}^{K} w_{ab}A_{ab}}{\sum_{a=1}^{K}\sum_{b=1}^{K} w_{ab}E_{ab}}.$$

Weighted kappa originates from the classical framework of graded agreement (Cohen, 1968); the quadratic weighting used here penalizes larger ordinal disagreements more strongly.

Permutation-invariance metrics. For each case–voice orbit i, let

$$\mathcal{O}_i = \{x_{iv}\}_{v=1}^{V_i}$$

denote the set of semantically equivalent fact permutations and let $\hat{y}_{iv}$ be the predicted ESI label for view v. We define the any-flip rate as

$$\text{Flip} = \frac{1}{N_{\text{orb}}} \sum_{i=1}^{N_{\text{orb}}} \mathbf{1}\left[|\{\hat{y}_{iv} : 1 \le v \le V_i\}| > 1\right].$$

Flip therefore counts an orbit whenever at least two equivalent permutations receive different predicted labels.

For analyses specifically concerned with the critical ESI-$2/3$ boundary, define

$$b_{iv} = \mathbf{1}[\hat{y}_{iv} \le 2].$$

The critical-boundary crossing rate is

$$\text{Crossing} = \frac{1}{N_{\text{orb}}} \sum_{i=1}^{N_{\text{orb}}} \mathbf{1}\left[\exists\, v, v' \text{ such that } b_{iv} \neq b_{iv'}\right].$$

Thus Flip detects any permutation-induced label change, whereas Crossing only detects changes that move the prediction across the high-acuity decision boundary.

Row-weighted and source-equal evaluation. The main paper uses row-weighted evaluation: every permutation row receives equal weight when computing the predictive metrics above. This preserves comparability with the training and historical evaluation pipeline.

As a complementary audit, we also use source-equal evaluation. All voice/permutation predictions belonging to the same source record are first collapsed to a single modal prediction, after which every source record receives equal weight. This prevents records with more evaluated views from receiving greater influence solely because of orbit size.

Metrics are computed independently for each random seed and then summarized by their arithmetic mean. No seed or configuration is selected using sealed-test performance. In tables where heterogeneous metrics are displayed on a common numerical scale, entries are shown as $100\times$ the underlying metric; this scaling is only for presentation and does not change any comparison.

B Datasets, Construction, and Access

Overview. We evaluate the clinical experiments on three emergency-care datasets with different data-generation and target-construction processes: MIMIC-IV-ED, EMS-M, and ER-Reason. We additionally use CIFAR-10-LT as a non-clinical cross-domain test of the contraction–selection mechanism. Importantly, only MIMIC-IV-ED contains native Emergency Severity Index (ESI) labels recorded as part of the source emergency-department data. EMS-M and ER-Reason are mapped into the same ordered acuity space for the purposes of our experiments, but their targets are constructed differently: EMS-M uses protocol-derived weak triage labels, whereas ER-Reason uses an explicit semantic mapping from its five-level acuity terminology to ESI 1–5. We keep these provenance distinctions explicit throughout the paper rather than treating all three datasets as natively ESI-annotated benchmarks.

MIMIC-IV-ED: native ESI triage labels. MIMIC is derived from MIMIC-IV-ED, a de-identified emergency-department database released through PhysioNet (Johnson et al., 2023; Goldberger et al., 2000). Among the three clinical datasets used in this work, MIMIC-IV-ED is the only one for which the target corresponds directly to an ESI value recorded in the source emergency-department data. We use emergency-department triage records with these recorded ESI 1–5 labels and construct the task-specific textual inputs used in our experiments. The clinically critical boundary used throughout the paper is therefore the native ESI 1–2 versus ESI 3–5 distinction in this dataset.

MIMIC-IV-ED is a credentialed-access resource. Access requires the corresponding PhysioNet training and data-use agreement. We therefore do not redistribute the underlying records or derived patient-level files.

EMS-M: protocol-derived weak triage labels. EMS-M is constructed from EMSDialog, a corpus of synthetic multi-speaker emergency-medical-service conversations generated from electronic patient care reports (Ge et al., 2026). EMSDialog does not provide native ESI annotations for the prediction problem studied here. We therefore construct a four-level ordinal triage target from the emergency protocol associated with each case and refer to the resulting task as *EMS-M*.

The mapping is based on the urgency and intervention semantics of the protocol rather than on an ESI field in the source data. For example, protocols corresponding to conditions requiring immediate life-saving intervention, such as *Cardiac Arrest* or *Failed Airway*, are assigned to the highest-acuity level corresponding to ESI 1. High-risk protocols that require rapid evaluation or intervention but not necessarily an immediate life-saving procedure (for example STEMI, severe symptomatic bradycardia, or severe shock/anaphylaxis) are assigned to the next acuity level, corresponding to ESI 2. The remaining protocols are analogously grouped into lower-acuity levels according to their intervention urgency and resource requirements.

Thus, references to ESI-like levels for EMS-M denote our *protocol-derived weak ordinal labels*, not clinician-recorded ESI assessments. The ESI 1–2 versus lower-acuity boundary used in the analysis should therefore be interpreted as a harmonized high-acuity decision boundary for this dataset rather than as an evaluation of native ESI documentation. The upstream EMSDialog dataset is released under CC BY-NC 4.0.

ER-Reason: semantic acuity-to-ESI mapping. ER-Reason is a benchmark of de-identified longitudinal emergency-room records designed for clinical reasoning evaluation (Mehandru et al., 2025). The original resource contains longitudinal clinical narratives for 3,984 patients, including information spanning emergency-department presentation, history and physical examination, ECG and imaging findings, consultations, treatment, disposition or ICU admission, and final diagnoses. Because many of these elements occur after the initial triage decision, our prediction inputs retain only the information admitted by the triage-oriented construction used in our experimental pipeline; downstream information is not intentionally introduced into the triage input.

ER-Reason does not provide a native numeric ESI field. Instead, its target acuity is expressed using a five-level emergency-severity vocabulary whose clinical semantics align naturally with the five ESI levels. For a common ordinal evaluation space, we use the following fixed semantic mapping:

$$\begin{aligned} \text{Immediate} &\longrightarrow \text{ESI } 1, \\ \text{Emergent} &\longrightarrow \text{ESI } 2, \\ \text{Urgent} &\longrightarrow \text{ESI } 3, \\ \text{Less Urgent} &\longrightarrow \text{ESI } 4, \\ \text{Non-Urgent} &\longrightarrow \text{ESI } 5. \end{aligned} \tag{5}$$

The mapping preserves the ordered meaning of the original acuity categories: *Immediate* denotes the need for immediate resuscitative care; *Emergent* denotes a high-risk emergency that should not wait; *Urgent* denotes a clinically urgent presentation; and the final two levels represent progressively lower-acuity care.

Accordingly, the "ESI" values reported for ER-Reason in our experiments are a deterministic semantic re-indexing of the benchmark's original five-level acuity annotation, not ESI values recorded by the treating emergency department. The ESI 1–2 versus 3–5 boundary therefore corresponds to the original *Immediate/Emergent* versus *Urgent/Less Urgent/Non-Urgent* distinction.

ER-Reason is distributed through PhysioNet under restricted access; credentialing, required training, a data-use agreement, and contributor review are required. As with MIMIC-IV-ED, no restricted clinical records are redistributed.

Why use a harmonized critical boundary? The three clinical datasets therefore instantiate the same experimental question through different sources of supervision. MIMIC-IV-ED supplies native five-level ESI labels; EMS-M supplies protocol-derived ordinal urgency labels; and ER-Reason supplies five-level acuity labels that are deterministically mapped into the ESI ordering. Our primary binary analysis does not assume that these annotation mechanisms are interchangeable. Instead, it uses the common high-acuity distinction represented by levels 1–2 versus the remaining levels as a

harmonized decision boundary, while retaining each dataset's original provenance when interpreting results.

CIFAR-10-LT. To test whether the proposed distinction between agreement contraction and shared-solution selection extends beyond clinical language and ordinal prediction, we additionally evaluate on a long-tailed CIFAR-10 setting following the standard artificially imbalanced CIFAR construction used in long-tail recognition (Cui et al., 2019). CIFAR-10 does not possess a natural clinical-style severity ordering. We therefore pre-specify one target class before evaluation and treat its one-vs-rest decision as the analogue of the risk-relevant boundary. Image augmentations replace textual fact permutations as equivalent views, while the symmetric and directional objectives respectively control cross-view contraction and the location of the resulting target-class decision. Full experimental details are reported in Appendix F.

Orbit construction and evaluation units. For each clinical source case, we construct two narrative voices and multiple fact-order permutations within each voice. A case–voice pair defines an orbit: all members contain the same underlying input facts admitted by the dataset-specific construction and share the same target, while differing only in fact order. The purpose of these permutations is to alter presentation rather than clinical content.

Predictive metrics in the main tables are computed over permutation rows. Because datasets differ in the number of rows generated from each underlying source record, we additionally compute a source-equal audit in which predictions are first aggregated within each source case and source cases are then weighted equally. Exact definitions of under-triage, over-triage, BA_2, Flip, critical-boundary crossing, and the two aggregation schemes are given in Appendix A.

C Implementation and Computational Environment

Software environment. All experiments were implemented in Python 3.11.14 with PyTorch 2.14.0, built against CUDA 13.0. The main model-training stack used Transformers 4.57.6, PEFT 0.18.1, Accelerate 1.13.0, bitsandbytes 0.49.2, tokenizers 0.22.2, and safetensors 0.7.0. Unless otherwise specified, the common LLM training setup uses 4-bit NF4 quantization with bfloat16 computation and LoRA adaptation. The remaining optimization and model-specific settings are reported with their corresponding experiments.

Training configuration and hyperparameters. Unless otherwise specified, all LLM experiments use QLoRA with rank $r = 16$, scaling parameter $\alpha = 16$, zero LoRA dropout, and no bias parameters. LoRA adapters are applied to the query, key, value, and output projections, as well as the gate, up, and down projections. Optimization uses AdamW with weight decay 0.01. The learning rate is 1×10^{-4} for LoRA parameters and 2×10^{-4} for the LayerNorm and classification head.

For the disagreement-conditioned model, the frozen default configuration is $\lambda = 3$, $a = 0.5$, $q = 0.9$, $m = 0.99$, and $\sigma = 1$. Here, q specifies the disagreement quantile used for gate normalization and m is the EMA momentum. Separate gating EMA states are maintained for the clinical and patient voices.

Classification models are trained for two epochs. Each optimization batch contains 10 source cases, with two permutation views drawn from the same voice for each source case. The maximum input length is 768 tokens. Unless a method requires a specialized inference path, evaluation uses a batch size of 8.

Hardware and standardized cost measurement. The full experimental campaign was executed across several compute clusters equipped with NVIDIA H200, RTX 5090, and A100 GPUs. Because wall-clock time and memory consumption are hardware dependent, we measure computational cost separately under a standardized environment: a single NVIDIA RTX A6000 GPU. Table 2 reports end-to-end training time, evaluation time, and peak GPU memory for representative methods on the MIMIC 4B setting. All entries in this table were measured under the same hardware environment, so the reported costs are directly comparable across methods; they are not intended to represent throughput on the heterogeneous clusters used for the full experimental sweep.

Table 2: Standardized computational cost on MIMIC 4B using a single NVIDIA RTX A6000. Times are wall-clock minutes and memory denotes peak GPU memory usage.

Method	Train (min)	Eval. (min)	GPU memory (GB)
CE	32.7	7.7	4.65
BWOT	51.3	12.0	4.80
OQT	51.1	11.9	4.80
WCE	51.0	12.2	4.80
MG-LoRA	51.1	11.7	4.65
OCE	49.3	11.8	4.65
SDTA	49.2	7.1	6.23
PINE	56.1	142.8	12.27

D Cross-LLM Replication of the Contraction–Selection Geometry

Purpose and protocol. The main experiments establish the contraction–selection distinction using Llama-3.1-Nemotron-Nano-4B. We next test whether this geometry depends on that particular representation or model scale. We keep the training objective, data construction, sealed-test evaluation, and frozen hyperparameters fixed while replacing the backbone with Qwen2.5-7B, Llama-3.1-8B, Hermes2-10.7B, and Qwen3-14B. The key comparison remains matched within the transport family: CE provides the unregularized anchor, BWOT uses symmetric transport with $a = 0$, fixed-direction OQT uses $g \equiv 1$, and DGN uses the same $\lambda = 3$ and $a = 0.5$ with disagreement-conditioned gating.

The same two control directions reappear across backbones. Panel 2c provides the most direct cross-architecture test of the paper's central decomposition. On MIMIC, CE→BWOT reduces Flip by 8.74–9.45 points across the four backbones, whereas under-triage changes by at most 2.35 points in magnitude. Averaged across backbones, this first intervention moves Flip by -9.14 points but Under by only $+0.36$ points. The matched BWOT→DGN intervention produces the complementary trajectory: under-triage falls by 10.36–12.23 points, while Flip changes by less than one point for every backbone, corresponding to mean changes of -11.62 and $+0.30$ points, respectively. Thus, replacing the LLM does not erase the empirical division of labor identified by the single-backbone intervention sweeps: symmetric transport predominantly contracts cross-view disagreement, whereas disagreement-conditioned direction predominantly relocates the shared risk operating point.

The same structure is visible on EMS-M at a smaller absolute scale. There, CE→BWOT changes mean Flip by -0.82 points and mean Under by only $+0.07$ points, while BWOT→DGN changes mean Under by -2.21 points and mean Flip by only -0.06 points. The magnitude therefore depends on the dataset, but the intervention geometry remains aligned with the contraction–selection decomposition.

Combining contraction and selection improves both stability and risk location. The two stages together produce a favorable net displacement relative to CE. Across the four MIMIC backbones, DGN reduces under-triage by 9.55–12.34 points and Flip by 7.98–9.68 points. The corresponding four-backbone means move from 40.90% to 29.65% in Under and from 19.79% to 10.95% in Flip. EMS-M shows the same direction of change, with mean Under decreasing from 15.67% to 13.53% and mean Flip from 3.90% to 3.02%.

These gains are not obtained through a comparable loss of ordinary predictive quality. Relative to CE, MIMIC BA_2 improves by approximately 2.0–2.7 points across the four backbones, while QWK remains within -0.56 to $+0.23$ points and accuracy within -1.07 to $+0.89$ points. On EMS-M, BA_2, QWK, and accuracy all increase across the four backbones. The cross-backbone result therefore strengthens the interpretation of DGN as a shared-solution selector rather than a mechanism that simply exchanges stability for ordinary predictive quality.

Disagreement conditioning improves the directional exchange. Panel 2d compares DGN with the fixed-direction member of the same transport family. Across the displayed MIMIC and EMS-M settings, DGN incurs only 0.54–0.63 additional points of over-triage for each point of under-triage

removed relative to CE. Fixed-direction OQT spans 0.68–1.06, reaching approximately a one-for-one exchange in several backbone–dataset combinations. This reproduces across architectures the role identified in the main experiments: direction can move the shared solution, while conditioning that direction on residual disagreement limits continued movement after the views have largely converged.

Connection to the decomposition. The cross-backbone trajectories provide an empirical counterpart to the function-space result in Appendix G.1. The symmetric component has no signed common-mode preference in the transport objective, whereas the directional component supplies the common-mode force. Figure 2c shows the same separation at the level of sealed-test behavior across distinct LLM representations: the intervention that changes symmetric contraction predominantly moves cross-view disagreement, while the intervention that activates directional selection predominantly moves the risk-relevant landing location.

E Mechanism-Focused Frozen-Readout Controls

Purpose. The main experiments use OBCO as a frozen-readout intervention: after the representation has been learned, the backbone, adapters, pooling operation, and normalization are held fixed, and only the final linear readout is changed. This section asks whether the resulting relocation of the risk operating point can be explained by simpler alternatives. In particular, we test whether the effect reduces to (i) unconstrained suppression of under-triage, (ii) preservation of ordinary predictive loss alone, (iii) generic cost-sensitive head refitting, (iv) a global critical/non-critical threshold shift, (v) the same constrained optimization without the orbit budget, or (vi) resolution of residual cross-view disagreement rather than relocation of decisions on which the views already agree.

Let W_0 denote the original frozen readout and let h_{iv} be the frozen feature of view v from case i. The full intervention solves the readout problem

$$\begin{aligned} \min_{W} \quad & U_{\mathrm{sur}}(W) + \frac{\rho}{2}\|W - W_0\|_F^2 \\ \text{s.t.} \quad & C(W) \le C(W_0), \\ & O_{\mathrm{sur}}(W) \le O_{\mathrm{sur}}(W_0), \\ & J(W) \le J(W_0), \end{aligned} \tag{6}$$

where C is multiclass cross-entropy, O_{sur} is the opposite-risk surrogate, and J measures centered within-orbit logit dispersion. All budgets are anchored at the original readout on frozen training features. The readout is fit without using test labels; sealed-test examples are used only for evaluation.

Appendices E.2 and E.3 report sealed-test metrics. The threshold-path and shared-error audits in Appendices E.4 and E.5 are conducted on the development split, where all readout variants of a run share the same frozen representation, training feature cache, and W_0; no development result is used to select a threshold, budget, or constraint.

E.1 Clinical risk remains adjustable across frozen representations

The main experiments apply OBCO to the representation learned by DGN. We next test whether this readout-level freedom is specific to DGN or remains present after substantially different training procedures. For each model, we freeze the backbone, adapters, pooling operation, and normalization, initialize from its own learned head W_0, and apply the same OBCO solver without method-specific retuning. Thus, within each row of Table 3, the learned representation is unchanged and only the final linear readout is modified.

The same qualitative relocation appears across all eight frozen representations. Under-triage decreases in every case, with corresponding increases in high-acuity sensitivity. More importantly, the negative likelihood ratio also decreases consistently, showing that the improvement is not confined to a re-expression of sensitivity: after the readout intervention, a lower-acuity prediction provides stronger evidence against true high acuity across all eight representation families. The effect therefore does not depend on DGN having learned a uniquely "repairable" feature space.

PINE provides the cleanest counterfactual. Its structured-document representation is already exactly permutation invariant, so all equivalent views share the same frozen feature before the readout.

Table 3: **Clinical high-acuity detection before and after OBCO on frozen MIMIC representations.** Each row uses exactly the same learned representation before and after the intervention; only the linear readout changes. Entries are percentages except $100 \times \mathrm{LR}^{-}$. Sensitivity equals $1 - U$ and is shown for clinical interpretability rather than as an independent metric.

Frozen representation	$U \downarrow$	Sensitivity$\uparrow$	$100 \times \mathrm{LR}^{-} \downarrow$
CE	44.73 → **21.14**	55.27 → **78.86**	52.52 → **31.29**
BWOT	49.96 → **23.92**	50.04 → **76.08**	57.81 → **34.72**
OQT	18.07 → **9.86**	81.93 → **90.14**	35.45 → **20.62**
DGN	40.53 → **19.86**	59.47 → **80.14**	50.50 → **28.95**
PINE	42.16 → **20.89**	57.84 → **79.11**	51.23 → **31.97**
OCE	45.37 → **22.46**	54.63 → **77.54**	53.04 → **32.78**
WCE	30.67 → **16.61**	69.33 → **83.39**	39.59 → **26.78**
SDTA	45.53 → **24.89**	54.47 → **75.11**	53.80 → **35.21**

Table 4: **Mechanism-focused frozen-readout controls on the sealed MIMIC test set.** All entries are $100\times$ the underlying metric. U and O denote under- and over-triage. The rows share the same frozen representation and differ only in the readout intervention. The table is intended as a mechanism comparison rather than a ranking: different columns expose different failure modes of a readout repair.

Readout intervention	$U \downarrow$	$O \downarrow$	$\mathrm{BA}_2 \uparrow$	Acc.$\uparrow$	QWK$\uparrow$	Flip$\downarrow$	Cross.$\downarrow$
U only	0.00	100.00	50.00	21.09	-0.36	5.83	0.00
U + CE	9.47	55.70	67.42	37.52	38.96	53.38	24.59
Balanced-WCE head	34.97	22.19	71.42	44.87	52.51	58.68	31.81
Critical-bias shift	24.93	30.32	72.37	43.18	48.38	15.24	6.96
U + CE + O (no J)	20.05	36.21	71.87	40.39	48.68	53.64	30.19
Full OBCO	**19.86**	31.41	**74.36**	43.95	54.30	28.53	14.35
Full OBCO, $\rho = 0$	19.99	31.70	74.16	43.88	54.45	28.23	14.45

Nevertheless, changing only that shared linear readout reduces under-triage from 42.16% to 20.89%, increases sensitivity from 57.84% to 79.11%, and reduces $100 \times \mathrm{LR}^{-}$ from 51.23 to 31.97. In the corresponding shared-stable-error audit, 152 critical source cases were under-triaged for every evaluated view under the original readout; after OBCO, 55 are correct for every view and another 74 are partially recovered. Because these cases were already in consensus before the intervention, their recovery cannot be attributed to resolving residual cross-view disagreement.

These results strengthen the interpretation of OBCO as a frozen-representation counterfactual: clinically consequential risk location remains adjustable after representation learning has ended and across representations produced by markedly different training objectives. The following controls ask whether this relocation can instead be explained by simpler forms of readout refitting.

E.2 The constraint ladder excludes trivial risk relocation

Table 4 reports the controls most directly associated with the alternative explanations above. The first row removes every behavioral constraint and optimizes only the under-triage surrogate. The second retains only the multiclass-CE budget. Balanced-WCE instead performs a conventional cost-sensitive refit on the same frozen representation. Critical-bias changes only a single global critical-side intercept, chosen on training data from the path described in Appendix E.4. No-J retains the CE and opposite-risk budgets but removes the orbit constraint. Finally, the $\rho = 0$ control retains all three behavioral constraints while removing the proximal penalty.

Predictive preservation alone is insufficient. The unconstrained objective reaches $U = 0$, but only by moving essentially the entire non-critical population to the critical side, giving $O = 100\%$ and $\mathrm{BA}_2 = 50\%$. Notably, this collapsed solution also has very low Flip and zero critical crossing. Thus a readout can be highly stable while occupying a useless shared operating point, which is precisely why agreement alone cannot define the desired solution.

Adding only the CE budget prevents complete collapse but does not determine a useful risk exchange: U falls to 9.47% while O rises to 55.70%, and both Flip and boundary crossing become large. Ordinary predictive preservation therefore does not by itself prevent a low-under solution

Table 5: **Effect of the orbit budget.** J_{train} and J_{dev} measure continuous centered logit dispersion on the frozen feature caches; the remaining columns are sealed-test metrics. Removing J substantially enlarges orbit dispersion and produces a corresponding increase in discrete prediction instability.

Variant	$J_{\mathrm{train}} \downarrow$	$J_{\mathrm{dev}} \downarrow$	$U \downarrow$	$O \downarrow$	Flip↓	Cross.↓	$\mathrm{BA}_2 \uparrow$
Full OBCO	**.0036**	**.0241**	**19.86**	**31.41**	**28.53**	**14.35**	**74.36**
No J	.2421	.4253	20.05	36.21	53.64	30.19	71.87

from being obtained through excessive escalation. The opposite-risk budget is needed to constrain this direction of movement.

The effect is not reproduced by generic cost-sensitive refitting. Balanced-WCE uses the same frozen representation and directly changes the label-side cost of the linear head, but reaches a different operating point: $U = 34.97\%$ and $\mathrm{BA}_2 = 71.42\%$. It also produces substantially larger permutation sensitivity (58.68% Flip). Thus the frozen representation certainly admits ordinary cost-sensitive movement, but the relocation produced by OBCO is not simply the result of reweighting critical versus non-critical examples.

The near-identical Full and $\rho = 0$ rows provide a complementary control. Removing the proximal term changes U, O, BA_2, Flip, and Crossing by less than a few tenths of a percentage point. The observed operating-point relocation is therefore governed primarily by the constrained risk geometry rather than by the small $\|W - W_0\|_F^2$ regularizer.

E.3 The orbit budget changes the attainable repair

The most direct internal control removes only J while retaining the CE and opposite-risk constraints. Table 5 separates the continuous orbit quantities used by the optimizer from the discrete behavior observed on sealed test examples.

Removing the orbit budget increases training orbit dispersion by roughly two orders of magnitude and development-set dispersion from 0.024 to 0.425. The same intervention nearly doubles sealed-test Flip ($28.53\% \rightarrow 53.64\%$) and more than doubles critical-boundary crossing ($14.35\% \rightarrow 30.19\%$). Importantly, this is not merely a stability-only difference: the no-J candidate also has higher over-triage and lower BA_2. The orbit budget therefore changes the attainable frozen-readout repair rather than acting as a cosmetic regularizer on logits.

E.4 Shared-solution relocation is not a global threshold shift on MIMIC

The most natural alternative explanation for a frozen-readout relocation is that it merely raises the critical-side logits, i.e., that it is a global decision-threshold change. We test this explanation against the entire one-dimensional family of such changes rather than against a hand-chosen range. Holding every parameter of W_0 fixed, we modify only the critical-class intercepts,

$$z'_k = z_k + \beta\, \mathbf{1}[k \le 2]. \tag{7}$$

For a training row with logits z, the predicted side of the ESI-$2/3$ boundary changes exactly at

$$\beta(z) = \max_{k \ge 3} z_k - \max_{k \le 2} z_k.$$

We sort these breakpoints over all training rows and take one value strictly below the smallest, one midpoint between every pair of consecutive distinct breakpoints, and one value strictly above the largest. Any global critical-side threshold chosen from training data alone is therefore equivalent on the training set to one point of this path. On MIMIC, the path contains $6{,}957$ points per frozen representation.

A single train-selected threshold. The critical-bias row of Table 4 is the train-selected point of this path, frozen before sealed-test evaluation. It reaches $U = 24.93\%$, $O = 30.32\%$, and $\mathrm{BA}_2 = 72.37\%$. Full OBCO reaches a substantially lower $U = 19.86\%$ while increasing O by only 1.09 points, yielding a higher $\mathrm{BA}_2 = 74.36\%$. A single point, however, cannot exclude that some other threshold would match OBCO, so we next evaluate the complete path.

Table 6: **Full OBCO against the complete train-selectable critical-bias path on the MIMIC development split.** Entries are $100\times$ the underlying metric. "Dominating" counts path points whose under- and over-triage are both no worse than those of Full OBCO. ΔO_{path} is the additional over-triage the path requires to reach Full OBCO's under-triage, and ΔU_{path} is the additional under-triage it requires to reach Full OBCO's over-triage. The maximum path BA_2 is selected with development labels and is therefore an oracle upper bound for threshold translation; Full OBCO uses no development labels.

	Full OBCO			Critical-bias path (6,957 points)			
Run	$U\downarrow$	$O\downarrow$	$\mathrm{BA}_2\uparrow$	Dominating	ΔO_{path}	ΔU_{path}	Max BA_2 (oracle)
1	19.80	27.67	76.26	0	$+6.11$	$+5.29$	74.39
2	20.14	28.93	75.47	0	$+3.44$	$+4.72$	74.16
3	20.25	28.93	75.41	0	$+0.93$	$+2.47$	75.33
Mean	20.06	28.51	75.71	0	$+3.49$	$+4.16$	74.63

The complete path does not reach the OBCO operating point. Table 6 shows that, in every run, no point on the path is at least as good as Full OBCO in both under- and over-triage. Matching Full OBCO's under-triage by threshold translation costs 0.93–6.11 additional points of over-triage (mean 3.49), and matching its over-triage costs 2.47–5.29 additional points of under-triage (mean 4.16). Full OBCO therefore lies strictly outside the under–over frontier attainable by global threshold translation.

The balanced-accuracy comparison is deliberately tilted toward the baseline. Even when the threshold is chosen with development labels, the best path point reaches 74.63% BA_2 on average, whereas Full OBCO, selected without development labels, reaches 75.71%. Because the margin is small in the third run (75.33% versus 75.41%), we read this as Full OBCO matching or exceeding the oracle threshold rather than as a uniform improvement; the frontier result above does not depend on this margin. The separation is also consistent with the sealed-test AUROC reported in Section 5.1: a monotone threshold translation leaves critical-boundary AUROC unchanged, whereas OBCO raises it from 81.24% to 82.83%.

A capacity-restricted readout, $W = W_0 + c\,u^\top$, in which c is a fixed centered critical-versus-non-critical class contrast and only the feature direction u is fitted under the same three budgets, is likewise undominated by the path in every run. Case-dependent readout changes, even when confined to a single class direction, therefore reach shared-solution locations that a global offset cannot.

Scope. We state this separation as mechanistic evidence on MIMIC, whose development split contains 258 source records. On ER-Reason, whose development split contains only 20 source records, the same audit does not separate OBCO from threshold translation (OBCO is undominated in one of three runs); we therefore restrict this claim to MIMIC.

E.5 Relocation repairs consensus errors, not only disagreements

Appendix E.4 asks whether the relocation is a global shift. A different alternative explanation is that the frozen-readout improvement arises mainly from resolving residual cross-view disagreement, i.e., from turning orbits whose views straddle the critical boundary into orbits that agree on the correct side. We test this explanation on orbits that contain no such disagreement.

A *shared under-triage error* is a critical case–voice orbit ($y_i \le 2$) for which the original readout W_0 predicts every evaluated permutation on the non-critical side ($\hat{y}_{iv} > 2$ for all v). In the terminology of Appendix G.5, such an orbit has no critical-boundary crossing under W_0: its views already agree, and they agree on the wrong side. After the readout intervention, we classify each such orbit as *all-correct* if every permutation moves to the critical side, *partial* if some but not all permutations do, and *still all-wrong* otherwise. Only an all-correct transition keeps the orbit free of crossing while reversing the side on which its views agree.

Agreed errors are relocated. Across runs, Full OBCO moves $34.9\% \pm 7.5\%$ of these orbits to all-correct: every permutation of an orbit that was unanimously under-triaged is now placed on the

Table 7: **Transitions of shared under-triage errors on the MIMIC development split.** Each counted orbit is a critical case–voice orbit whose permutations are all predicted non-critical by the original readout W_0; the development split contains 164 critical orbits. Percentages are within-run fractions of the shared errors, and the mean row averages these fractions.

Run	Shared errors under W_0	$\rightarrow$ All-correct	$\rightarrow$ Partial	$\rightarrow$ Still all-wrong
1	55	19 (34.5%)	13 (23.6%)	23 (41.8%)
2	58	16 (27.6%)	16 (27.6%)	26 (44.8%)
3	61	26 (42.6%)	13 (21.3%)	22 (36.1%)
Mean	58.0	34.9%	24.2%	40.9%

critical side. Because these orbits had no disagreement to resolve, their recovery cannot be attributed to disagreement resolution; it changes where an already-agreeing orbit lands. The exactly invariant PINE representation in Appendix E.1 shows the same pattern in its cleanest form, since there all equivalent views share one frozen feature and cross-view disagreement is absent by construction.

Partial repairs account for part of the Flip increase. A further 24.2% of the shared errors become partially correct. These transitions turn a consistent error into a disagreement, which is the orbit-level counterpart of the higher test-time Flip of OBCO reported in Table 1. The remaining 40.9% stay all-wrong, so the relocation is substantial but incomplete.

What this audit does and does not show. Any sufficiently large movement toward the critical side, including points of the threshold path in Appendix E.4 and the degenerate U-only readout in Table 4, can also repair shared errors. This audit therefore does not by itself distinguish OBCO from threshold translation; that question is answered by Appendix E.4. Its role is complementary: at the level of individual orbits, it shows that the frozen readout changes the location of an agreed decision rather than only the amount of disagreement, which is exactly the distinction between shared-solution selection and agreement contraction. On ER-Reason, the corresponding denominators are at most seven shared errors per run, which is too small to report a rate.

Summary. These controls isolate the role of the OBCO components. Minimizing the critical-side surrogate alone gives a degenerate but highly stable solution; CE preservation alone still permits excessive escalation; generic balanced reweighting reaches a different and substantially less stable operating point; removing the orbit budget enlarges both continuous orbit dispersion and discrete test-time instability; and, on MIMIC, no train-selectable critical-side threshold matches the full readout in both under- and over-triage (Appendix E.4). Together with the shared-error audit, which shows that the relocation reverses orbits already agreeing on the wrong side (Appendix E.5), these results support the interpretation used in the main text: even after representation learning is frozen, the risk-relevant location of the shared prediction remains a selectable degree of freedom, and OBCO exposes that freedom without reducing it to either agreement contraction or a global threshold shift.

F Cross-Domain Mechanism Validation on CIFAR-10-LT

Purpose. The clinical experiments expose shared-solution selection in a setting where equivalent views are textual permutations, the predictor is an LLM, and the decision space is ordinal. We use CIFAR-10-LT to ask whether the same distinction survives after removing all three of these ingredients simultaneously. Here equivalent views are image transformations, the predictor is a CIFAR-style ResNet-32 trained from scratch, and the labels form an unordered ten-class problem. The purpose of this experiment is therefore not to establish a new long-tailed recognition state of the art, but to test whether *agreement contraction* and *shared-solution selection* remain separately observable and controllable outside clinical language.

Protocol. We use CIFAR-10-LT with imbalance ratio 10. The training set contains $4{,}084$ images, with class counts decreasing from $1{,}000$ for C0 to 100 for C9. Before running the experiments, we fix the least frequent training class, C9 (*truck*), as the target class $c^\star$; it is not selected from test performance. For each seed, we reserve 100 images per class from the original CIFAR-10 test set

as development data and use the remaining 900 images per class as a sealed test set, yielding 9,000 balanced test images. No test result is used for hyperparameter or method selection.

The backbone is a CIFAR-style ResNet-32 with 466,906 parameters, trained for 200 epochs using SGD with batch size 128, learning rate 0.1, momentum 0.9, weight decay 5×10^{-4}, and cosine decay. Each training image produces two independently augmented views using `RandomCrop(32,padding=4)` and `RandomHorizontalFlip`. At evaluation time, each image defines a four-view orbit containing the identity image, its horizontal flip, and translations by $+2$ and -2 pixels.

Because CIFAR-10 has no natural severity ordering, we use the pre-specified C9-vs-rest decision as the analogue of the critical boundary. Target recall measures how often C9 examples remain on the target side, target false-positive rate (FPR) measures movement of non-C9 examples onto that side, and target balanced accuracy is

$$\mathrm{BA}_{\mathrm{C9}} = \frac{1}{2}\left(\mathrm{Recall}_{\mathrm{C9}} + 1 - \mathrm{FPR}_{\mathrm{C9}}\right).$$

For invariance, Any-flip records whether any member of the four-view orbit changes the predicted class, Mean JS measures distributional disagreement, and Target Crossing records whether the orbit crosses the C9-vs-rest decision boundary.

F.1 Non-ordinal realization of contraction and selection

Let $p_i^{(1)}$ and $p_i^{(2)}$ denote the class distributions of two augmented views. The symmetric component is a stop-gradient bidirectional KL,

$$\mathcal{L}_{\mathrm{sym}} = \frac{1}{2B}\sum_{i=1}^{B}\left[\mathrm{KL}\Big(\mathrm{sg}(p_i^{(2)}) \,\|\, p_i^{(1)}\Big) + \mathrm{KL}\Big(\mathrm{sg}(p_i^{(1)}) \,\|\, p_i^{(2)}\Big)\right].$$

It acts on the full ten-class distributions and vanishes whenever the two views agree, irrespective of the common distribution on which they agree.

To instantiate solution selection without an ordinal CDF, define the target-class probabilities

$$q_i^{(v)} = p_i^{(v)}[c^\star].$$

A stop-gradient sibling teacher T_i is formed as a convex combination of the two $q_i^{(v)}$, with weights obtained from a softmin of the per-view gold-label cross-entropies. We write

$$d_i^{(v)} = q_i^{(v)} - T_i, \qquad s_i = \mathrm{sg}\Big(\max_v q_i^{(v)} - \min_v q_i^{(v)}\Big).$$

DGN uses the normalized disagreement gate

$$g_i = \mathrm{clip}\left(\frac{s_i}{\max(m_t, 10^{-6})}, 0, 1\right),$$

where m_t is an exponential moving average of the minibatch 0.9 quantile of s_i. The directional risk term is

$$\mathcal{L}_{\mathrm{risk}} = \frac{1}{2B}\sum_{i=1}^{B}\sum_{v=1}^{2}\left(|d_i^{(v)}| - a\, g_i d_i^{(v)}\right),$$

and the complete objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{sup}} + \lambda_{\mathrm{sym}}\mathcal{L}_{\mathrm{sym}} + \beta_{\mathrm{risk}}\mathcal{L}_{\mathrm{risk}}.$$

The matched directional family below fixes $\lambda_{\mathrm{sym}} = 1$, $\beta_{\mathrm{risk}} = 1$, the teacher, optimization, and all data exposure, and changes only a. Thus $a = 0$ provides an exact no-direction anchor for testing whether the shared target-class location can be moved while the contraction mechanism is held fixed.

Table 8: Symmetric-consistency dose response on sealed CIFAR-10-LT test data. Values are five-seed mean ± sample standard deviation.

Method	Acc. ↑	Flip ↓	Mean JS ↓	Crossing ↓	C9 Recall ↑	C9 BA ↑
CE	69.36±1.66	27.01±2.38	.0993±.0114	3.08±.35	57.67±3.66	78.63±1.83
SYM-0.25	69.33±1.78	25.18±.86	.0879±.0060	2.73±.16	56.71±3.22	78.15±1.59
SYM-1	69.06±1.03	23.85±.80	.0742±.0020	2.63±.14	55.33±4.02	77.43±2.02
SYM-4	66.38±3.37	**19.84±.92**	**.0473±.0023**	**2.05±.15**	50.64±4.87	75.10±2.46

Table 9: Different interventions move different parts of the CIFAR-10-LT evaluation profile. Values are five-seed mean ± sample standard deviation.

Method	Acc. ↑	Flip ↓	Mean JS ↓	C9 Recall ↑	C9 BA ↑
CE	69.36±1.66	27.01±2.38	.0993±.0114	57.67±3.66	78.63±1.83
SYM-1	69.06±1.03	23.85±.80	.0742±.0020	55.33±4.02	77.43±2.02
Balanced Softmax	71.99±2.05	26.22±1.66	.0964±.0060	66.53±1.28	82.79±.71
Balanced Softmax + SYM	71.58±1.60	23.04±1.93	.0720±.0078	66.04±4.25	82.63±2.10

F.2 Symmetric agreement contracts deviation but does not select a favorable shared location

Table 8 varies only the strength of symmetric consistency. All entries are five-seed means with sample standard deviations. Increasing λ_{sym} progressively reduces disagreement across equivalent image views, but the target-side operating point does not improve with contraction.

From CE to SYM-4, Any-flip falls by 7.17 points, Mean JS falls from 0.0993 to 0.0473, and target-boundary crossing falls from 3.08% to 2.05%. The model therefore becomes substantially more invariant. Over the same intervention, however, accuracy falls by 2.98 points, C9 recall falls by 7.03 points, and C9 balanced accuracy falls by 3.53 points. Thus stronger agreement can successfully contract equivalent views while moving their shared target-side decision to a worse operating point.

The milder SYM-1 intervention shows the separation before this strong regularization regime is reached. Relative to CE, accuracy changes by only -0.30 points, whereas Flip falls by 3.16 points and Mean JS by 0.0251. C9 recall nevertheless moves by -2.34 points. Agreement can therefore change substantially without fixing either ordinary predictive quality or the target-side location of the shared prediction.

F.3 Contraction and shared location form empirically distinct control directions

The separation is also visible when the same symmetric term is applied to a strong label-side baseline. Table 9 compares ordinary CE and symmetric consistency with Balanced Softmax, which modifies the class-prior correction but contains no cross-view agreement term.

Balanced Softmax (Ren et al., 2020) moves mainly along the label-side axis: relative to CE, accuracy increases by 2.63 points, C9 recall by 8.86 points, and C9 balanced accuracy by 4.16 points, whereas its changes in Flip and Mean JS are comparatively small. Adding SYM to this stronger classifier produces the complementary intervention: relative to Balanced Softmax alone, Flip falls by 3.18 points and Mean JS by 0.0244, while accuracy changes by only -0.41 points and C9 recall by -0.49 points.

The same symmetric objective can therefore contract cross-view disagreement without determining the quality or target-side location of the classifier on which the views agree. Conversely, substantially improving the target-side classifier does not by itself produce the same contraction. These interventions provide an empirical counterpart to the contraction–selection distinction without relying on ordinal labels.

Table 10: Matched directional sweep on sealed CIFAR-10-LT test data. Only the directional coefficient a changes across rows. Values are five-seed mean ± sample standard deviation.

a	Acc. ↑	Flip ↓	C9 Recall ↑	C9 BA ↑	C9 FPR ↓
0	68.14±3.05	25.11±1.42	55.47±5.09	77.52±2.56	.42±.10
.25	69.42±1.55	23.76±2.15	56.13±2.29	77.85±1.11	.43±.09
.50	69.92±1.69	24.04±.58	58.29±3.58	78.93±1.76	.43±.10
.75	68.98±2.43	23.64±2.14	57.73±4.02	78.63±1.98	.48±.08

Table 11: Raw disagreement gating versus normalized DGN at $a = .5$ on sealed CIFAR-10-LT test data. Values are five-seed mean ± sample standard deviation.

Method	Acc. ↑	Flip ↓	Mean JS ↓	C9 Recall ↑	C9 BA ↑	C9 FPR ↓
Raw-DG	68.37±1.79	25.08±2.89	.0791±.0124	52.64±4.67	76.13±2.33	.39±.07
DGN	69.92±1.69	24.04±.58	.0758±.0034	58.29±3.58	78.93±1.76	.43±.10

F.4 A MATCHED DIRECTIONAL INTERVENTION MOVES THE SHARED SOLUTION

We next hold the contraction mechanism fixed and vary only the directional coefficient a. The four configurations in Table 10 share the same supervised loss, full-distribution symmetric term, $\lambda_{\mathrm{sym}} = 1$, risk-transport skeleton, teacher, optimizer, and data exposure. The $a = 0$ row is therefore the exact no-direction anchor for this family.

Introducing direction moves the mean target-side operating point while the agreement statistics remain in a comparatively narrow range. C9 recall rises from 55.47% at $a = 0$ to 58.29% at $a = .5$ and remains 57.73% at $a = .75$, while Any-flip stays between 23.64% and 25.11% and C9 FPR between 0.42% and 0.48%. Relative to the $a = 0$ anchor, the $a = .75$ endpoint increases C9 recall by 2.26 points and C9 balanced accuracy by 1.11 points while changing Flip by -1.47 points.

The response is not strictly monotone in a: the largest mean C9 recall is observed at $a = .5$, not at the largest directional coefficient. Our claim is therefore controllability rather than monotonic performance improvement. Once the contraction mechanism is held fixed, changing the directional term can separately relocate the shared target-side operating point. This provides a non-ordinal analogue of the clinical contraction–selection intervention without requiring a severity ordering over the ten CIFAR classes.

F.5 WHY DISAGREEMENT NORMALIZATION MATTERS

CIFAR also provides a direct control for the scale normalization used by DGN. With the raw gate

$$g_i = \mathrm{clip}(s_i, 0, 1),$$

the target-probability spread is numerically small, so the directional component receives little effective weight. We compare this raw gate with the normalized gate s_i/m_t while holding $a = .5$, the symmetric term, the risk skeleton, and all other training choices fixed.

Replacing the raw gate with the normalized gate increases mean C9 recall by 5.65 points and C9 balanced accuracy by 2.80 points. Accuracy increases by 1.55 points, while Flip changes by -1.04 points, Mean JS by -0.0033, and C9 FPR by only $+0.04$ points. Thus normalization primarily restores an effective directional intervention rather than producing a comparably large change in cross-view contraction.

This control should not be read as a claim that larger direction or disagreement gating uniformly improves every CIFAR metric. In particular, the matched five-seed fixed-direction and DGN configurations are mixed across metrics rather than uniformly ordered. The CIFAR evidence is used here for the narrower mechanism claim: agreement contraction and target-side solution selection can be manipulated separately, and scale normalization is necessary for the directional component to remain numerically active in this setting.

F.6 SCOPE OF THE CROSS-DOMAIN CLAIM

These experiments validate the mechanism; they do not claim state-of-the-art long-tailed image classification. Balanced Softmax changes the label prior and contains no cross-view agreement term. Table 9 shows that it composes with symmetric consistency: Balanced Softmax + SYM keeps both the Flip reduction of SYM and the target-recall gain of Balanced Softmax, so label-side and agreement-side interventions complement each other. Our conclusion is therefore structural. After the input modality, model family, and label geometry are all changed, symmetric agreement still mainly controls cross-view deviation, and an explicit directional term can separately move the shared target-side operating point. This reproduces the contraction–selection distinction outside the clinical LLM setting and supports shared-solution selection as a general design variable for invariance learning, not one specific to ordinal triage.

G THEORY OF SHARED-SOLUTION SELECTION

This appendix formalizes the distinction between agreement contraction and shared-solution selection used throughout the paper. The first result is exact for the two-view transport objective used in our experiments: in cumulative-probability coordinates, symmetric transport acts only on within-orbit disagreement, whereas the directional component provides the common-mode force that can relocate the shared prediction. We then study an explicitly idealized single-boundary gradient flow to expose how symmetric and directional transport resolve a disagreement. Because these objective-level statements do not automatically determine the behavior of a shared-parameter network, we next report a post-freeze sensitivity analysis asking whether the same transport coordinates are visible in held-out decisions. We then separate critical-boundary crossing from the location of the shared decision algebraically, and finally establish the convexity and uniqueness properties of OBCO. The exact theoretical results should not be interpreted as guarantees on held-out risk or as claims of parameter-space orthogonality; the sensitivity analysis is complementary empirical evidence rather than an extension of those guarantees.

G.1 EXACT MODE DECOMPOSITION OF THE TWO-VIEW TRANSPORT

We suppress the example index and consider the two sibling views used during training. Let $z_v \in \mathbb{R}^K$ be the logits of view $v \in \{1, 2\}$, $p_v = \mathrm{softmax}(z_v)$, and

$$F_{vk} = \sum_{j=1}^{k} p_{vj}, \qquad k = 1, \dots, K-1, \tag{8}$$

so that, under the ESI ordering used in the paper, larger F_{vk} places more probability mass on the more urgent side of boundary k. For the clinical boundary of interest, $k^\star = 2$.

The per-example transport term shared by BWOT, OQT, and DGN can be written as

$$\mathcal{R}_a = \frac{1}{2} \sum_{v=1}^{2} \sum_{k=1}^{K-1} \omega_k \left(|d_{vk}| - a g_k d_{vk} \right), \qquad d_{vk} = F_{vk} - T_k, \tag{9}$$

where $0 \le a < 1$ and

$$T_k = \mathrm{sg}\left(\alpha_1 F_{1k} + \alpha_2 F_{2k} \right), \qquad \alpha_v = \frac{\exp[-\mathrm{sg}(e_v)/\tau]}{\sum_{u=1}^{2} \exp[-\mathrm{sg}(e_u)/\tau]}. \tag{10}$$

Here sg denotes stop-gradient. The gate $g_k \in [0, 1]$ is also treated as stop-gradient. BWOT is obtained with $a = 0$, fixed-direction OQT uses $g_k \equiv 1$, and DGN uses a detached disagreement gate of the form

$$g_k = \min\left(1, \frac{s_k}{m}\right), \qquad s_k = |F_{1k} - F_{2k}|, \qquad m > 0, \tag{11}$$

where m denotes the detached running normalization scale. Thus DGN satisfies $g_k = 0$ whenever the two views agree at boundary k.

Following the implemented automatic differentiation, let $\widetilde{\partial}$ denote differentiation with all stop-gradient quantities held fixed and with the convention $\partial|x|/\partial x = 0$ at $x = 0$. Define

$$G_{vk} := \frac{\widetilde{\partial}\mathcal{R}_a}{\partial F_{vk}}, \qquad \bar{F}_k = \frac{F_{1k} + F_{2k}}{2}, \qquad \delta_k = \frac{F_{1k} - F_{2k}}{2}. \tag{12}$$

The coordinates $\bar{F}_k$ and δ_k respectively represent the common CDF location and the half-disagreement of the two views.

Lemma G.1 (Two-view teacher cancellation)**.** *Let*

$$\chi_k = \operatorname{sign}(F_{1k} - F_{2k}).$$

For finite e_1, e_2 and $\tau > 0$, the softmax weights in Eq. 10 satisfy $\alpha_1, \alpha_2 \in (0, 1)$, and

$$G_{1k} = \frac{\omega_k}{2}\left(\chi_k - a g_k\right), \qquad G_{2k} = \frac{\omega_k}{2}\left(-\chi_k - a g_k\right). \tag{13}$$

Consequently, the pseudo-gradient of the transport term is independent of the teacher weights α_v, the temperature τ, and the particular per-view reliability scores e_v.

Proof. If $F_{1k} = F_{2k}$, then $T_k = F_{1k} = F_{2k}$ and hence $d_{1k} = d_{2k} = 0$. Under the stated convention for the derivative of the absolute value,

$$G_{1k} = G_{2k} = -\frac{\omega_k}{2} a g_k,$$

which is Eq. 13 with $\chi_k = 0$.

Now suppose $F_{1k} \neq F_{2k}$. Since $\alpha_1 + \alpha_2 = 1$,

$$d_{1k} = F_{1k} - \alpha_1 F_{1k} - \alpha_2 F_{2k} = \alpha_2 (F_{1k} - F_{2k}), \tag{14}$$
$$d_{2k} = F_{2k} - \alpha_1 F_{1k} - \alpha_2 F_{2k} = -\alpha_1 (F_{1k} - F_{2k}). \tag{15}$$

Because both weights are strictly positive,

$$\operatorname{sign}(d_{1k}) = \chi_k, \qquad \operatorname{sign}(d_{2k}) = -\chi_k.$$

With T_k and g_k held fixed by stop-gradient,

$$\frac{\widetilde{\partial}}{\partial F_{vk}}\left(|d_{vk}| - a g_k d_{vk}\right) = \operatorname{sign}(d_{vk}) - a g_k.$$

Multiplying by $\omega_k/2$ gives Eq. 13. The values of α_v, τ, and e_v enter only through the positive multipliers above and therefore disappear from the pseudo-gradient. □

Theorem G.2 (Common-mode/disagreement-mode decomposition)**.** *Under the conditions of Lemma G.1,*

$$\boxed{\frac{\widetilde{\partial}\mathcal{R}_a}{\partial \bar{F}_k} = -a\omega_k g_k, \qquad \frac{\widetilde{\partial}\mathcal{R}_a}{\partial \delta_k} = \omega_k \chi_k.} \tag{16}$$

Therefore:

- *(i) for symmetric transport ($a = 0$), the transport regularizer has exactly zero common-mode pseudo-gradient for every boundary and every amount of disagreement;*
- *(ii) within this transport family, the directional component is the only regularizer term that acts directly on the common CDF mode, with training-loss magnitude $\lambda a \omega_k g_k$;*
- *(iii) the disagreement-mode pseudo-gradient is independent of a and g_k in these coordinates;*
- *(iv) for DGN, $g_k = 0$ at agreement and the directional common-mode force therefore vanishes exactly, whereas fixed-direction OQT retains the common-mode force $-\lambda a \omega_k$.*

Proof. Since

$$F_{1k} = \bar{F}_k + \delta_k, \qquad F_{2k} = \bar{F}_k - \delta_k,$$

the chain rule gives

$$\frac{\widetilde{\partial}\mathcal{R}_a}{\partial \bar{F}_k} = G_{1k} + G_{2k}, \qquad \frac{\widetilde{\partial}\mathcal{R}_a}{\partial \delta_k} = G_{1k} - G_{2k}.$$

Substituting Eq. 13,

$$G_{1k} + G_{2k} = -a\omega_k g_k, \tag{17}$$

$$G_{1k} - G_{2k} = \omega_k \chi_k, \tag{18}$$

which proves Eq. 16. Multiplication by the outer transport coefficient λ gives the corresponding force in the complete training loss. For DGN, Eq. 11 gives $g_k = 0$ whenever $F_{1k} = F_{2k}$; for OQT, $g_k \equiv 1$. □

Interpretation. Theorem G.2 is an objective-level statement in cumulative probability coordinates. It formalizes the distinction used in the main text: symmetric transport directly contracts the disagreement coordinate δ_k but supplies no signed preference to the common coordinate $\bar{F}_k$, while the directional term supplies precisely such a preference. This does *not* imply that $\bar{F}_k$ remains numerically fixed during neural-network optimization. The map from shared parameters to (F_{1k}, F_{2k}) is nonlinear, the cross-entropy term also acts on the predictions, and a parameter update shared across cases can couple their common and differential modes. Accordingly, the theorem supports function-space separation rather than strict parameter-space orthogonality.

G.2 Idealized resolution dynamics

The zero common-mode gradient of symmetric transport does not imply that the eventual meeting point of two predictions must equal their arithmetic mean in probability space. Optimization acts through logits, and the probability map introduces a nonlinear geometry. The following deliberately idealized dynamics isolate the rule induced by the transport regularizer itself.

Assumption G.3 (Single-boundary free-logit flow)**.** *Only the critical boundary $k^\star$ is considered. Each view is represented by its urgent-side probability*

$$u_v = F_{vk^\star} = \sigma(\zeta_v) = \frac{1}{1 + e^{-\zeta_v}}, \qquad \zeta_v \in \mathbb{R}.$$

The two logits ζ_1, ζ_2 are free parameters, cross-entropy and parameter sharing are removed, and gradient flow follows only the transport term:

$$\dot{\zeta}_v = -\lambda \frac{\widetilde{\partial}\mathcal{R}_a}{\partial \zeta_v}.$$

For DGN the normalization scale is held fixed while a pair is resolved, so the gate can be written

$$g = \gamma(|u_1 - u_2|),$$

where $\gamma : [0,1] \to [0,1]$ is Lipschitz and $\gamma(0) = 0$.

Define the confidence potential

$$\Phi(u) = \frac{1}{1-u} - \frac{1}{u} + 2\log\frac{u}{1-u}, \qquad u \in (0,1). \tag{19}$$

Its derivative is

$$\Phi'(u) = \frac{1}{u^2(1-u)^2} > 0, \tag{20}$$

so Φ is strictly increasing from $(0,1)$ onto $\mathbb{R}$. It is also antisymmetric around the decision cut:

$$\Phi(1-u) = -\Phi(u), \qquad \Phi(1/2) = 0.$$

Proposition G.4 (How a disagreement is resolved)**.** *Under Assumption G.3, suppose $u_1(0) > u_2(0)$ and write*

$$\Phi_v^0 = \Phi(u_v(0)).$$

After rescaling time by $\lambda\omega_{k^\star}/2$, the following statements hold.

(a) ***BWOT.*** *The two views meet at*

$$t^\star = \frac{\Phi_1^0 - \Phi_2^0}{2} \tag{21}$$

at a consensus $u^\star$ satisfying

$$\Phi(u^\star) = \frac{\Phi_1^0 + \Phi_2^0}{2}. \tag{22}$$

Under the automatic-differentiation convention above, the consensus then remains stationary. If the initial views straddle $1/2$, the consensus lies on the side of the view farther from $1/2$:

$$u^\star > \frac{1}{2} \iff u_1(0) - \frac{1}{2} > \frac{1}{2} - u_2(0).$$

(b) ***DGN.*** *The views meet at the same $t^\star$, but*

$$\Phi(u^\star) = \frac{\Phi_1^0 + \Phi_2^0}{2} + a \int_0^{t^\star} g(t)\, dt, \tag{23}$$

where

$$0 \le a \int_0^{t^\star} g(t)\, dt \le \frac{a}{2} \left(\Phi_1^0 - \Phi_2^0\right). \tag{24}$$

Thus DGN tilts the resolution toward the urgent side only while disagreement exists. In particular, if both views start on the same side of the boundary, their consensus remains on that side. After they agree, $g = 0$ and the directional motion stops.

(c) ***OQT.*** *Before the meeting time, the dynamics have the same form as part (b) with $g \equiv 1$. After agreement, however,*

$$\frac{d}{dt}\Phi(u_v) = a, \tag{25}$$

for both views. Consequently the common prediction continues to move toward the urgent extreme rather than becoming stationary.

Proof. Since $u_v = \sigma(\zeta_v)$,

$$\dot{u}_v = u_v(1 - u_v)\dot{\zeta}_v, \qquad \frac{\widetilde{\partial}\mathcal{R}_a}{\partial \zeta_v} = u_v(1 - u_v)\frac{\widetilde{\partial}\mathcal{R}_a}{\partial u_v}.$$

Together with Eq. 20,

$$\frac{d}{dt}\Phi(u_v) = \frac{\dot{\zeta}_v}{u_v(1 - u_v)} = -\lambda \frac{\widetilde{\partial}\mathcal{R}_a}{\partial u_v}.$$

Let

$$\chi = \operatorname{sign}(u_1 - u_2).$$

Lemma G.1, restricted to the single boundary, gives

$$\frac{\widetilde{\partial}\mathcal{R}_a}{\partial u_1} = \frac{\omega_{k^\star}}{2}(\chi - ag), \qquad \frac{\widetilde{\partial}\mathcal{R}_a}{\partial u_2} = \frac{\omega_{k^\star}}{2}(-\chi - ag).$$

After the stated time rescaling,

$$\frac{d}{dt}\Phi(u_1) = -\chi + ag, \qquad \frac{d}{dt}\Phi(u_2) = \chi + ag. \tag{26}$$

Before the two views meet, $u_1 > u_2$ and hence $\chi = 1$. Therefore

$$\frac{d}{dt}(\Phi(u_1) - \Phi(u_2)) = -2, \tag{27}$$

$$\frac{d}{dt}(\Phi(u_1) + \Phi(u_2)) = 2ag. \tag{28}$$

The potential difference reaches zero at

$$t^\star = \frac{\Phi_1^0 - \Phi_2^0}{2}.$$

Since Φ is strictly increasing, equality of the two potentials is equivalent to $u_1 = u_2 = u^\star$. Integrating their sum yields Eq. 23. Because $0 \le g \le 1$, Eq. 24 follows immediately.

For BWOT, $a = 0$, so Eq. 22 follows. If the views straddle $1/2$, then

$$u^\star > \frac{1}{2} \iff \Phi_1^0 + \Phi_2^0 > 0 \tag{29}$$

$$\iff \Phi(u_1(0)) > \Phi(1 - u_2(0)) \tag{30}$$

$$\iff u_1(0) > 1 - u_2(0), \tag{31}$$

which is exactly the stated confidence condition.

For DGN, if both initial views are below $1/2$, then $\Phi_1^0, \Phi_2^0 < 0$ and

$$\Phi(u^\star) \le \frac{1}{2}\left[(1+a)\Phi_1^0 + (1-a)\Phi_2^0\right] < 0$$

because $a < 1$. If both are above $1/2$, the unshifted average is already positive and the directional correction is nonnegative. Hence DGN cannot move a same-side pair across the cut in this idealized flow. At agreement, $g = \gamma(0) = 0$, so Eq. 26 becomes stationary.

For OQT, $g \equiv 1$. At agreement $\chi = 0$, and Eq. 26 gives $d\Phi(u_v)/dt = a$ for both views. Thus the predictions remain equal but their common value continues to move. □

The preceding proposition removes supervision to expose the regularizer's resolution rule. A complementary calculation restores cross-entropy and asks where persistent direction places an already-agreeing prediction.

Corollary G.5 (Persistent direction changes the agreement fixed point)**.** *Consider a collection of inputs that the model represents identically. Let their label distribution be $\pi = (\pi_1, \dots, \pi_K)$ with $\pi_j > 0$, and suppose both views output the same distribution p. At agreement, fixed-direction OQT has the same expected pseudo-gradient, up to the common factor $1/2$, as the simplex objective*

$$\Psi(p) = -\sum_{j=1}^{K} \pi_j \log p_j - \lambda a \sum_{k=1}^{K-1} \omega_k F_k(p). \tag{32}$$

Its unique minimizer is

$$p_j^\star = \frac{\pi_j}{\mu - c_j}, \qquad c_j = \lambda a \sum_{k=j}^{K-1} \omega_k, \tag{33}$$

where an empty sum is zero and $\mu > \max_j c_j$ is the unique value satisfying $\sum_j p_j^\star = 1$. For DGN, $g_k = 0$ at agreement, so the corresponding minimizer is simply $p^\star = \pi$.

Proof. At agreement, $d_{vk} = 0$. The absolute transport term therefore has zero pseudo-gradient under our automatic-differentiation convention. For OQT, $g_k = 1$, and the directional component contributes $-\lambda a \omega_k/2$ to the pseudo-gradient with respect to F_{vk}. The expected view-averaged cross-entropy contributes one half of the gradient of $-\sum_j \pi_j \log p_j$. This proves the pseudo-gradient identity with Eq. 32.

Because

$$F_k(p) = \sum_{j=1}^{k} p_j,$$

we may rewrite

$$\lambda a \sum_{k=1}^{K-1} \omega_k F_k(p) = \sum_{j=1}^{K} c_j p_j, \qquad c_j = \lambda a \sum_{k=j}^{K-1} \omega_k.$$

Hence

$$\Psi(p) = -\sum_j \pi_j \log p_j - \sum_j c_j p_j.$$

Since all $\pi_j > 0$, this objective is strictly convex on the open probability simplex and diverges at its relative boundary. Introducing a Lagrange multiplier μ for $\sum_j p_j = 1$ gives

$$-\frac{\pi_j}{p_j} - c_j + \mu = 0,$$

and therefore

$$p_j = \frac{\pi_j}{\mu - c_j}.$$

Positivity requires $\mu > \max_j c_j$. The function

$$h(\mu) = \sum_j \frac{\pi_j}{\mu - c_j}$$

is continuous and strictly decreasing on this interval, diverges as μ approaches $\max_j c_j$ from above, and tends to zero as $\mu \to \infty$. It therefore has exactly one root $h(\mu) = 1$.

For DGN, agreement implies $s_k = 0$ and therefore $g_k = 0$. The directional pseudo-gradient disappears, leaving ordinary cross-entropy, whose population minimizer is $p = \pi$. □

Scope of the flow analysis. Proposition G.4 and Corollary G.5 isolate mechanisms rather than describe the full LoRA optimization trajectory. They remove parameter sharing or replace a heterogeneous representation by an identical-feature cell. Their purpose is therefore qualitative: symmetric transport contains no risk-aware tie-break, DGN tilts a disagreement only while it exists, and fixed-direction OQT continues to act on an already-agreeing prediction.

G.3 Post-freeze sensitivity of the transport coordinates

Protocol and scope. This analysis was initiated only after all main-text configurations and reported comparisons had been frozen. We vary one transport hyperparameter at a time on MIMIC with the 4B backbone under the training setup of Appendix C. Accordingly, this section is a post-freeze diagnostic rather than an additional independently sealed evaluation: no setting identified here was used to alter a main-text configuration, rerun Table 1, or select a reported main-text result.

The DGN default ($\lambda = 3$, $a = 0.5$, $q = 0.9$) was retrained once for this diagnostic and reused across the DGN λ-, a-, and q-sweeps. Because this is an independent training batch, its default row differs numerically from the frozen five-seed DGN row in Table 1; we leave the main-table result unchanged. The BWOT sweep was trained in a separate post-freeze batch, except for its $\lambda = 3$ point, which is not retrained but reuses the frozen five-seed BWOT result of Table 1 so that the sweep contains the main-comparison operating point. Rows whose configurations coincide across batches therefore differ only by batch-to-batch variation: the two $\lambda = 0$ rows both reduce to CE and differ from the CE row of Table 1 by at most 2.2 points in U, and the DGN-a row at $a = 0$ reduces to BWOT and differs from the reused BWOT point by 0.45 points. Apart from the reused BWOT point, all comparisons below are made within a single batch.

From the exact decomposition to a network-level diagnostic. Theorem G.2 is an exact identity for the implemented two-view transport pseudo-gradient in cumulative-probability coordinates. After multiplication by the consistency strength λ, the symmetric disagreement component is scaled by λ, whereas the directional common-mode component is scaled by $\lambda a g_k$. The theorem does *not* imply that the held-out decisions of a shared-parameter network trained jointly with cross-entropy must obey the same coordinate separation: parameter sharing couples examples and modes, and discrete decisions are obtained only after the nonlinear network, softmax, and argmax. We therefore use the post-freeze sweeps below to ask whether the same organization is nevertheless visible empirically.

Table 12: **Post-freeze transport sensitivity on MIMIC-4B.** All entries are shown as $100\times$ the metric. BWOT varies λ with $a = 0$; DGN-λ varies λ at $a = 0.5$; and DGN-a varies a at $\lambda = 3$. The two rows marked † are the same independently retrained DGN default and are repeated only to make both sweeps explicit. The BWOT row marked * is not retrained: it reuses the frozen five-seed BWOT result of Table 1. All other rows are post-freeze retraining. At $\lambda = 0$, the transport term vanishes.

Setting	λa	$U\downarrow$	O	Flip↓	Cross.↓
BWOT: vary λ with $a = 0$					
$\lambda = 0$	0	46.91	15.03	24.65	11.46
$\lambda = 0.5$	0	47.47	15.07	20.54	9.74
$\lambda = 1$	0	48.42	13.97	17.73	7.75
$\lambda = 2$	0	49.51	13.37	14.94	6.03
$\lambda = 3^*$	0	49.96	13.59	10.54	4.41
$\lambda = 6$	0	51.63	13.44	7.55	3.38
DGN-λ: vary λ with $a = 0.5$					
$\lambda = 0$	0	46.64	14.19	22.86	10.37
$\lambda = 0.5$	0.25	45.55	15.69	20.51	9.44
$\lambda = 1$	0.50	44.83	16.37	15.97	7.19
$\lambda = 2$	1.00	40.35	19.23	14.61	6.06
$\lambda = 3^\dagger$	1.50	35.20	23.09	12.99	5.60
$\lambda = 6$	3.00	10.66	65.18	10.64	7.19
DGN-a: vary a with $\lambda = 3$					
$a = 0$	0	49.51	13.87	11.80	5.33
$a = 0.1$	0.30	46.67	15.39	11.40	5.27
$a = 0.25$	0.75	42.94	17.41	12.52	5.20
$a = 0.5^\dagger$	1.50	35.20	23.09	12.99	5.60
$a = 0.75$	2.25	24.27	33.35	15.01	7.06
$a = 0.9$	2.70	16.37	47.39	14.65	8.42

Symmetric strength is not a tuning artifact. The BWOT sweep in Table 12 strengthens only symmetric contraction. Across the full grid, mean Flip falls from 24.65% to 7.55% and critical-boundary crossing from 11.46% to 3.38%, while mean under-triage moves in the opposite direction, from 46.91% to 51.63%. Among the retrained points, Flip is non-increasing with λ for every seed, and under-triage is higher at $\lambda = 6$ than at $\lambda = 0$ for every seed; the reused $\lambda = 3$ point lies between its retrained neighbors in both mean Flip and mean under-triage. Moreover, no $\lambda > 0$ on the grid lowers mean under-triage below the unregularized endpoint. Thus the BWOT operating point used in the main comparison is not explained by an unfavorable choice of symmetric consistency strength: increasing λ reliably buys agreement on this grid, but does not produce a risk-preferred landing side.

Directional relocation is consistent with the λa coordinate. The two DGN sweeps change the directional scale in different ways. At fixed $a = 0.5$, increasing λ moves under-triage from 46.64% to 10.66% while over-triage rises from 14.19% to 65.18%. At fixed $\lambda = 3$, increasing a produces the same ordering of the risk operating point, moving under-triage from 49.51% to 16.37% and over-triage from 13.87% to 47.39%.

The two grids were not designed to provide exact product-matched pairs, so we do not infer an exact λa collapse from these runs. Nevertheless, their locations align when indexed by the directional scale. For example, the DGN-λ operating point at $\lambda a = 0.5$ lies between the neighboring DGN-a points at $\lambda a = 0.3$ and 0.75 in both under- and over-triage; the same bracketing holds for the DGN-λ point at $\lambda a = 1.0$ between the DGN-a points at 0.75 and 1.5. Meanwhile, the fixed-λ a-sweep moves the risk operating point by more than thirty under-triage points while Flip remains in the much narrower 11.40–15.01% range. These observations are consistent with the coordinate interpretation of Theorem G.2: λ supplies contraction pressure, while the product λa, modulated by disagreement gating, sets the scale of directional common-mode transport. They should be read as empirical alignment rather than a claim of exact parameter-space orthogonality.

The gate quantile is not an isolated favorable setting. Finally, we vary the normalization quantile q from 0.50 to 0.95 while holding $\lambda = 3$ and $a = 0.5$ fixed. Across this range, BA_2 remains within 70.80–71.55%, Flip within 11.96–13.72%, and critical-boundary crossing within 4.44–6.00%; the default $q = 0.9$ is therefore not an isolated optimum. The corresponding training-time gate statistics change smoothly: the mean gate decreases from 0.642 to 0.462, the saturation rate from 41.1% to 18.8%, and the observed effective coefficient $a\bar{g}$ from 0.321 to 0.231. As this ef-

fective directional scale decreases, under-triage rises monotonically from 30.28% to 35.74%, while over-triage falls monotonically from 27.05% to 22.05%. Empirically, changing q therefore behaves as a mild rescaling of the directional coordinate rather than exposing a narrow hyperparameter optimum.

Takeaway. The post-freeze sweeps address two distinct sensitivity concerns. First, the main BWOT comparison is not rescued by increasing or decreasing its symmetric consistency strength over the evaluated grid. Second, DGN's risk relocation changes smoothly with the directional scale and remains stable across a broad range of gate normalizations. Together, these diagnostics connect the exact function-space decomposition to held-out network behavior without treating that correspondence as a theorem about the full shared-parameter optimization trajectory.

G.4 Relative disagreement unlocks directional selection

What this ablation isolates. DGN gates its directional component using disagreement relative to an adaptive reference scale,

$$g_{ik} = \min\left(1, \frac{s_{ik}}{m_t}\right).$$

We ask whether disagreement dependence alone is sufficient, or whether the relative scale normalization is itself functionally important. To isolate this choice, the raw-gate ablation keeps the teacher, transport weights, optimization, $\lambda = 3$, and $a = 0.5$ unchanged, and replaces the normalized gate by

$$g_{ik}^{\mathrm{raw}} = \min(1, s_{ik}).$$

This is not another no-gate control: unlike fixed-direction OQT ($g \equiv 1$), the raw variant remains disagreement-conditioned and its directional force still vanishes at exact agreement. The intervention removes only the adaptive disagreement scale.

Why relative scale can matter. The mode decomposition in Theorem G.2 gives, for nonzero disagreement,

$$\left|\frac{\widetilde{\partial}\mathcal{R}_i/\partial \bar{F}_{ik}}{\widetilde{\partial}\mathcal{R}_i/\partial \delta_{ik}}\right| = a g_{ik}.$$

Because $0 \le s_{ik} \le 1$, the raw gate is simply $g_{ik}^{\mathrm{raw}} = s_{ik}$, so its relative directional strength is $a s_{ik}$ and therefore shrinks directly with the absolute orbit spread. By contrast, the normalized gate has relative strength

$$a \min\left(1, \frac{s_{ik}}{m_t}\right),$$

so a global contraction of disagreement need not by itself extinguish the directional component relative to the symmetric contraction force. Normalization therefore preserves a disagreement-relative directional scale while retaining DGN's defining property that direction disappears when agreement is reached.

Raw gating preserves contraction but largely removes relocation. Figure 3 shows a consistent three-stage pattern across the clinical datasets. Relative to BWOT, the raw gate changes under-triage by only -0.02, -0.42, and -1.17 percentage points on MIMIC, EMS-M, and ER-Reason, respectively. Its Flip changes by only $+0.79$, -0.98, and $+0.03$ points. The same near-BWOT behavior appears in the other primary risk endpoints: relative to BWOT, raw gating changes NPV by at most 0.53 points and $100 \times \mathrm{LR}^-$ by at most 1.59 points. Thus, making direction depend on absolute disagreement is sufficient to retain the contraction regime, but under this frozen configuration it does not produce substantial shared-solution relocation.

Normalization restores the directional control. Restoring the adaptive scale changes the operating point sharply. Relative to the raw gate, normalized DGN lowers under-triage by 9.41, 2.47, and 12.49 points on MIMIC, EMS-M, and ER-Reason, respectively. NPV increases by 2.18, 1.16, and 3.24 points, while $100 \times \mathrm{LR}^-$ decreases by 7.22, 2.37, and 9.29 points. The corresponding Flip changes are much smaller ($+2.88$, -0.10, and $+1.20$ points), so the intervention predominantly moves risk location rather than undoing contraction.

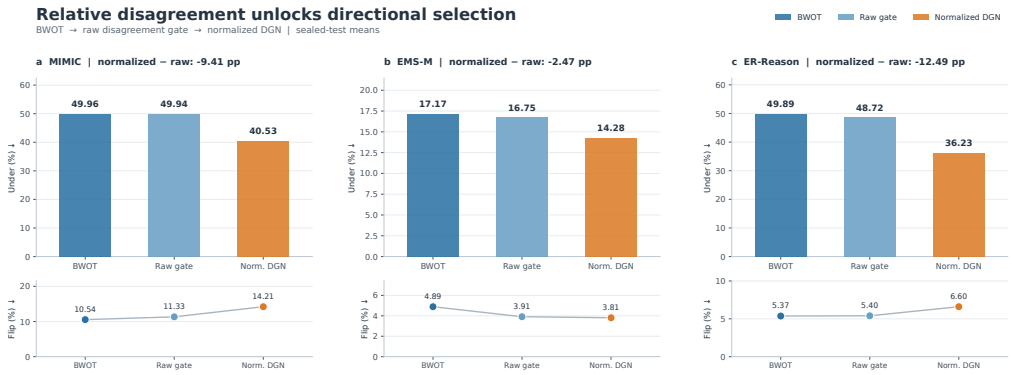


Figure 3: **Relative disagreement unlocks directional selection.** We compare symmetric BWOT, a raw disagreement gate $g^{\text{raw}} = \min(1, s)$, and normalized DGN under the frozen $\lambda = 3$, $a = 0.5$ configuration. Top panels report row-weighted under-triage and bottom panels report permutation Flip. Raw gating leaves the risk operating point close to symmetric BWOT on all three clinical datasets, whereas restoring relative disagreement normalization produces substantial shared-solution relocation without a comparable change in Flip.

This pattern is consistent with the force decomposition above: the raw gate introduces an absolute-spread factor into the common-mode directional force, whereas normalization expresses that force relative to the disagreement scale currently present during training. The result should not be read as a claim that this particular normalization is mathematically necessary for every task; rather, across all three clinical datasets, removing the adaptive scale strongly attenuates directional selection under the frozen DGN configuration. The same failure mode appears independently on CIFAR-10-LT (Appendix F.5), where replacing normalized disagreement by raw spread likewise suppresses target-side relocation. Together, the clinical and image ablations show that DGN's normalization is a functional part of the selection mechanism rather than a cosmetic numerical stabilization.

G.5 Critical-boundary crossing and landing location

The preceding results concern the training objective. We next give an algebraic separation directly in terms of discrete decisions.

For logits $z \in \mathbb{R}^K$, define the high-acuity decision margin

$$r(z) = \max_{j \le 2} z_j - \max_{j \ge 3} z_j. \tag{34}$$

Ignoring exact ties, the predicted label lies on the high-acuity side of the ESI-2/3 boundary if and only if $r(z) > 0$.

For a two-view orbit i, write

$$m_i = \frac{r(z_{i1}) + r(z_{i2})}{2}, \qquad \Delta_i = \frac{|r(z_{i1}) - r(z_{i2})|}{2}, \tag{35}$$

where m_i is the decision landing and Δ_i is the decision-level half-width. Let

$$X_i = \mathbf{1}\left[r(z_{i1})r(z_{i2}) < 0\right] \tag{36}$$

be the indicator that the two views cross the critical boundary.

For a collection of N two-view orbits, define

$$P = I - \frac{1}{K}\mathbf{1}\mathbf{1}^\top, \qquad \bar{z}_i^P = \frac{1}{2}\left(Pz_{i1} + Pz_{i2}\right), \tag{37}$$

and the projected orbit energy

$$J = \frac{1}{2NK}\sum_{i=1}^{N}\sum_{v=1}^{2}\left\|Pz_{iv} - \bar{z}_i^P\right\|_2^2. \tag{38}$$

This is the $V = 2$ specialization of the dispersion quantity constrained by OBCO.

Theorem G.6 (Separation of crossing and landing risk)**.** *For every two-view orbit, and ignoring exact decision ties,*

(i)

$$X_i = \mathbf{1}\left[|m_i| < \Delta_i\right]; \tag{39}$$

(ii) if $\mathcal{H}$ denotes the set of genuinely high-acuity cases, U is row-level under-triage over their two views, and

$$U^{\text{land}} = \Pr_{i\in\mathcal{H}}(m_i < 0), \tag{40}$$

then

$$\left|U - U^{\text{land}}\right| \le \frac{1}{2}\Pr_{i\in\mathcal{H}}(X_i = 1); \tag{41}$$

(iii) for every $t > 0$,

$$\Pr(X = 1) \le \Pr(|m| < t) + \frac{\mathbb{E}[\Delta^2]}{t^2}, \tag{42}$$

and, when J is evaluated on the same two-view orbits,

$$\mathbb{E}[\Delta^2] \le 4KJ. \tag{43}$$

Proof. Up to exchanging the two views,

$$r(z_{i1}) = m_i + \Delta_i, \qquad r(z_{i2}) = m_i - \Delta_i.$$

Their product is therefore

$$r(z_{i1})r(z_{i2}) = m_i^2 - \Delta_i^2,$$

which is negative exactly when $|m_i| < \Delta_i$. This proves part (i).

For an urgent case i, let

$$f_i = \frac{1}{2}\sum_{v=1}^{2}\mathbf{1}[r(z_{iv}) < 0]$$

be the fraction of its two rows that are under-triaged. If $X_i = 0$, both margins have the sign of m_i, so

$$f_i = \mathbf{1}[m_i < 0].$$

If $X_i = 1$, one view lies on each side of the boundary and hence $f_i = 1/2$. Thus

$$|f_i - \mathbf{1}[m_i < 0]| \le \frac{1}{2}X_i.$$

Averaging over $\mathcal{H}$ proves Eq. 41.

For part (iii), if $X_i = 1$ and $|m_i| \ge t$, then part (i) implies $\Delta_i > |m_i| \ge t$. Therefore

$$X_i \le \mathbf{1}[|m_i| < t] + \mathbf{1}[\Delta_i > t].$$

Taking expectations and applying Markov's inequality to Δ_i^2 gives Eq. 42.

It remains to relate Δ_i to J. The margin is invariant to a common logit shift, so $r(z) = r(Pz)$. Furthermore, each maximum is 1-Lipschitz under $\|\cdot\|_\infty$, and hence

$$|r(z) - r(z')| \le 2\|P(z - z')\|_\infty \le 2\|P(z - z')\|_2.$$

Consequently,

$$\Delta_i \le \|P(z_{i1} - z_{i2})\|_2.$$

For $V = 2$,

$$Pz_{iv} - \bar{z}_i^P = \pm\frac{1}{2}P(z_{i1} - z_{i2}),$$

and therefore

$$J = \frac{1}{4NK}\sum_{i=1}^{N}\|P(z_{i1} - z_{i2})\|_2^2.$$

Combining the last two displays gives

$$\mathbb{E}[\Delta^2] \le \frac{1}{N}\sum_i \|P(z_{i1} - z_{i2})\|_2^2 = 4KJ,$$

which proves Eq. 43. □

Interpretation. Theorem G.6 separates two ingredients of a risk-relevant consistency failure. The orbit width Δ_i determines how far equivalent views spread, while the landing m_i determines where the orbit sits relative to the clinical boundary. A boundary crossing requires the width to exceed the distance of the landing from the boundary. Conversely, reducing width alone does not specify the sign of m_i. Critical-boundary crossing is a subset of any-label Flip, so this theorem applies to the risk-relevant component of the reported permutation instability rather than to every within-side class change.

Equation 43 also clarifies the role of OBCO's orbit budget. Constraining J controls a continuous measure of orbit width on the features where the constraint is evaluated, but it does not fix the landing and does not by itself guarantee a discrete test-time Flip rate. This is why the OBCO constraint is interpreted as a training-representation budget rather than a sealed-test invariance guarantee.

G.6 Convexity and uniqueness of OBCO

We finally record a useful property of the frozen-readout intervention. Recall Eq. 4 from the main text:

$$\min_{W} \quad U_{\mathrm{sur}}(W) + \frac{\rho}{2}\|W - W_0\|_F^2$$

subject to

$$C(W) \le C(W_0), \qquad O_{\mathrm{sur}}(W) \le O_{\mathrm{sur}}(W_0), \qquad J(W) \le J(W_0).$$

Proposition G.7 (OBCO is a feasible convex program with a unique solution)**.** *Assume the frozen features are finite and $\rho > 0$. Then the feasible set of Eq. 4 is nonempty and convex, and the optimization problem has a unique global minimizer.*

Proof. For every frozen feature h_{iv},

$$z_{iv}(W) = W h_{iv}$$

is affine in W. The boundary surrogate used in the main text has the form

$$\ell_S(z, y) = \frac{1}{\log 2} \log\left(1 + \sum_{k \in S} e^{z_k - z_y}\right).$$

This is a log-sum-exp of affine functions of W, and is therefore convex in W. Averages preserve convexity, so both $U_{\mathrm{sur}}(W)$ and $O_{\mathrm{sur}}(W)$ are convex. The multiclass cross-entropy $C(W)$ is likewise an average of log-sum-exp functions and is convex.

For the orbit term, let

$$\bar{h}_i = \frac{1}{V}\sum_{u=1}^{V} h_{iu}.$$

Because P is linear,

$$P z_{iv}(W) - \bar{z}_i^P(W) = PW(h_{iv} - \bar{h}_i),$$

which is linear in W. Hence

$$J(W) = \frac{1}{NVK}\sum_{i,v} \left\|PW(h_{iv} - \bar{h}_i)\right\|_2^2$$

is a convex quadratic function.

A sublevel set of a convex function is convex. Therefore each of the three OBCO constraints defines a convex set, and their intersection is convex. Moreover, W_0 satisfies all three constraints with equality, so the feasible set is nonempty.

Finally,

$$\frac{\rho}{2}\|W - W_0\|_F^2$$

is ρ-strongly convex for $\rho > 0$. Adding the convex function $U_{\mathrm{sur}}(W)$ preserves strong convexity. The objective is also coercive because of the quadratic term, while the feasible set is closed and nonempty. A minimizer therefore exists, and strong convexity implies that it is unique. Any feasible stationary solution returned by an exact convex solver is thus the global OBCO solution. □

What these results do and do not establish. Theorem G.2 is exact for the implemented two-view, stop-gradient transport objective in cumulative-probability coordinates. Proposition G.4 and Corollary G.5 are mechanism-isolating idealizations and are not theorems about the complete shared-parameter LoRA dynamics. Theorem G.6 is an exact algebraic statement for the evaluated two-view margins, but an orbit-energy constraint measured on training features does not imply the same bound on unseen test features. Proposition G.7 guarantees the geometry of the frozen-readout optimization problem, not that its unique solution improves any particular test metric. Together, the results establish the narrower claim needed by the paper: agreement contraction and shared-solution location are distinct objects in the objective and decision geometry, and DGN and OBCO expose complementary ways of controlling the latter.