A THEORETICAL ANALYSIS OF SPARSE MIXTURE-OF-EXPERTS ROUTING VIA COMPATIBILITY OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse Mixture-of-Experts (MoE) architectures play an important role in scaling large language models with sparse computation. Yet it remains unclear why an abrupt change of expert can leave a prediction unchanged, while another equally small routing change causes substantial loss. To understand this, we formulate sparse expert routing as an optimal transport problem with downstream predictive risk as the cost and expert capacity as a constraint, which we call *compatibility optimal transport* (COT). As a theoretical contribution, we derive a testable optimality condition for capacity-constrained routing, along with an exact identity for its regularized regret. We show that, even at the same switching frequency, whether routing boundaries align with expert indifference boundaries determines whether regret grows linearly or quadratically. Building on these results, we explain how capacity coordination, continuous gates, and sparse gradient estimates affect predictive risk, and identify the conditions each explanation requires. Our experiments validate the predicted mechanism and find that boundary alignment governs how regret scales. Compared with switching frequency alone, incorporating expert costs improves switching-cost prediction from $R^2 = 0.33$–0.47 to $R^2 = 0.86$–0.88.

1 INTRODUCTION

Sparse Mixture-of-Experts (MoE) networks play an important role in scaling large models: they expand the available expert capacity while activating only a small subset of experts for each state (Shazeer et al., 2017; Fedus et al., 2022; Riquelme et al., 2021). Their top-k gates determine which computational modules process each state, providing a concrete setting in which to study how routing affects prediction. Is it possible for an abrupt change of expert to leave a prediction unchanged while another equally small routing change causes substantial loss, and if so, why? Geometric analysis shows where these assignments become discontinuous and motivates smoothing near switching boundaries (Huu et al., 2026). But a visible switch is not itself a prediction error. Treating every boundary as harmful can waste computation on interchangeable experts while leaving costly assignments untouched. We therefore need to connect routing changes to downstream predictions.

We develop that connection through *interface compatibility*: the conditional risk obtained by forcing a complete sparse action while holding the remaining network fixed. This separates three geometries: the hidden-state distribution, the router partition, and the predictive-cost partition. Expert capacity couples assignments across states. Their joint optimization is a semi-discrete optimal transport problem for top-1 routing and a moment-constrained finite-action transport problem for top-k. Its dual prices convert predictive cost into *reduced cost*, accounting for shared-capacity competition.

Assignment and Sinkhorn routers optimize learned affinities under expert-load constraints (Lewis et al., 2021; Liu et al., 2023b; Nguyen et al., 2025). We ask when those scores represent downstream risk. COT supplies an independently measurable cost, pairwise log-odds restriction, and exact excess-objective identity. It distinguishes fitting an assignment from identifying the objective that makes it useful. Combining duality with differential geometry makes this distinction operational through tests of reduced-cost alignment and downstream loss-weighted boundary risk.

The argument has two parts. ***Which changes matter?*** A weighted coarea law links boundary risk to density, transverse router variation, and loss jumps; Fisher expansion explains the jump, while

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.

weighted local time captures repeated encounters under continuous perturbations. ***How should harmful changes be corrected?*** Capacity coordination changes which states receive each expert, while continuous gates change the executed output and sparse gradient estimates change how selection is trained. We compare these effects with expert weights fixed. We test this chain with observable candidate costs and interventions in exact transport, learned expert banks, a compact causal language model, and a four-layer image model. Following Huu et al. (2026), these experiments isolate routing mechanisms through paired comparisons of complete predictions at fixed expert weights.

Contributions. **(1)** A predictive-risk formulation of sparse assignment, with dual identification and exact regularized router regret (Section 3). **(2)** Cost-weighted boundary laws separating geometric frequency from predictive harm and connecting local changes to accumulated loss (Section 4). **(3)** Conditions explaining the predictive effects of capacity coordination, continuous gates, and sparse gradient estimates (Section 5). **(4)** Layered tests of scaling, curvature, and full-forward repair: compatibility weighting explains switching cost better than boundary frequency alone (Section 6).

2 Related Work

Sparse routing under expert-load constraints. Mixture-of-experts models originated as collections of local experts with a learned gate (Jacobs et al., 1991); modern sparse layers increase parameter capacity while activating only a small subset of experts for each input (Shazeer et al., 2017). This makes expert-load allocation essential. GShard and Switch Transformers use load-balancing losses and capacity controls, while BASE and Expert Choice impose balance more directly in token-to-expert assignment (Lepikhin et al., 2021; Fedus et al., 2022; Lewis et al., 2021; Zhou et al., 2022). For a frozen model, this leaves a complementary question: among feasible assignments under the same expert-load target, which ones yield lower downstream predictive risk?

Routing discontinuities and predictive risk. Small input changes near routing boundaries can change the selected expert set. Soft MoE replaces hard token-to-expert selection with soft mixtures of input tokens (Puigcerver et al., 2024), whereas ReMoE uses ReLU routing to make the routing map continuous (Wang et al., 2025). Huu et al. (2026) analyze boundary geometry and study local smoothing rules. Boundary proximity alone does not determine predictive cost, which also depends on the selected experts and the frozen continuation. We therefore weight boundary neighborhoods by predictive cost to distinguish frequent switches from costly downstream assignments.

Predictive routing interventions. Recent analyses relate routing patterns to hidden-state geometry (Wang et al., 2026). Other methods use successful neighbors to improve routing (Li et al., 2026) or virtual ablations to evaluate counterfactual expert choices (Hu et al., 2026). SparseMixer and GRIN approximate gradients through discrete expert selection (Liu et al., 2023a; 2024a). We evaluate a complete sparse action by forcing its expert set and gate weights while holding the remaining network fixed. Independent choices can overload shared experts, requiring joint load-constrained optimization.

Transport and assignment objectives. Optimal transport provides a framework for constrained assignment and dual capacity prices (Cuturi, 2013; Peyré & Cuturi, 2019); transport-based MoE routers apply this machinery to learned routing scores (Liu et al., 2023b; 2024b; Nguyen et al., 2025). Those scores need not equal the conditional risk of forcing a candidate action. COT uses that predictive risk as the assignment cost, uses dual prices to represent shared capacity, and tests whether the router represents the resulting constrained objective. Our matched affinity-transport comparison isolates this cost choice with fixed expert weights and the same number of active experts per state.

3 Compatibility Optimal Transport

Actions and predictive cost. We compare routing choices at one interface with the remaining network fixed. Let $H \sim \mu$ contain the context needed by this frozen continuation, and let Y be the prediction target. Each action a in a finite family $\mathcal{A}$ selects a set S_a of k out of E experts and gate weights $w_a(h) \in \mathbb{R}^k$, producing interface output $\Phi_a(h)$. For the predictive distribution P_a, define

$$C_a(h) = \mathbb{E}[-\log P_a(Y \mid h) \mid H = h], \qquad g_i(a) = \mathbf{1}\{i \in S_a\}, \tag{1}$$

The observed loss $\ell_a(h, y) = -\log P_a(y \mid h)$ supervises C_a but is not itself the conditional risk.

From action costs to constrained assignment. A policy $q_a(h)$ assigns action probabilities, with average risk $L(q) = \mathbb{E}_\mu \sum_a q_a C_a$. Choosing the cheapest action at every state may overload shared experts, so we also track expert use through $g(a)$ and expected load $\rho(q) = \mathbb{E}_\mu \sum_a q_a g(a)$. We impose a feasible target $\rho(q) = \rho$, where $\sum_i \rho_i = k$. For top-1, $\pi(dh, a) = \mu(dh)q(a \mid h)$ couples states to the discrete expert-load measure $\sum_i \rho_i \delta_i$; for top-k, it satisfies the moment constraint $\int g(a) d\pi = \rho$. Joint action costs need not be additive across experts, since the continuation can couple their outputs even when all expert loads agree (Appendices B and I).

To characterize the optimal policy, we add a KL penalty relative to a reference policy ν, with strength $\tau > 0$. The following theorem shows how expert prices adjust predictive costs to meet the load target. Predictive losses and the temperature τ are measured in nats; division by $\log 2$ converts them to bits.

Theorem 1 (Compatibility transport and expert prices)**.** *Let $\mathcal{A}$ be finite, g bounded, C_a and $\log \nu_a$ integrable, and $\nu_a(h) > 0$ a reference probability kernel. Write $\mathrm{KL}(q\|\nu) = \sum_a q_a \log(q_a/\nu_a)$, with $0 \log 0 = 0$. If ρ lies in the relative interior of $\mathrm{conv}\{g(a) : a \in \mathcal{A}\}$ and $\tau > 0$, minimizing*

$$\mathcal{J}_\nu(q) = L(q) + \tau \mathbb{E}_\mu \mathrm{KL}(q(\cdot \mid H)\|\nu(\cdot \mid H)) \tag{2}$$

under $\rho(q) = \rho$ has the unique solution

$$q^\star_a(h) = \frac{\nu_a(h) \exp[-D_a(h)/\tau]}{Z_\alpha(h)}, \qquad D_a(h) = C_a(h) + \alpha^\top g(a). \tag{3}$$

Here $Z_\alpha = \sum_a \nu_a \exp(-D_a/\tau)$ and the expert-price vector $\alpha \in \mathbb{R}^E$ enforces the load target. Prices are unique modulo vectors orthogonal to load differences; dual-attaining hard optima minimize reduced cost among the declared actions at almost every state under the fixed data measure.

The classical entropic-transport Gibbs form (Cuturi, 2013; Peyré & Cuturi, 2019) specifies how predictive cost and expert load determine routing probabilities (Appendix B). An independent reference ν defines a target for testing the router; using the native router as ν defines a proximal correction without establishing native optimality. For upper capacities $\rho_i(q) \le b_i$, prices satisfy $\alpha_i \ge 0$ and complementary slackness. Batch capacity requires checks of realized expert loads.

Testing an existing router. We can test a router against this optimum. Pairwise odds characterize optimality, and divergence from $q^\star$ gives a feasible policy's regularized regret (Appendix B).

Theorem 2 (Identification and exact regret)**.** *Under Theorem 1, a positive load-feasible policy p equals $q^\star$ if and only if some price vector α satisfies, almost surely,*

$$\tau \log \frac{p_a(h)\nu_b(h)}{p_b(h)\nu_a(h)} = -(C_a - C_b)(h) - \alpha^\top (g(a) - g(b)) \quad \textit{for all } a, b. \tag{4}$$

Every load-feasible q also satisfies

$$\mathcal{J}_\nu(q) - \mathcal{J}_\nu(q^\star) = \tau \mathbb{E}_\mu \mathrm{KL}(q(\cdot \mid H)\|q^\star(\cdot \mid H)). \tag{5}$$

Equation (4) can be tested on held-out states using costs estimated independently of router scores and prices fitted on separate data. For a hard router, replacing active expert d with inactive expert j changes reduced cost by $\Delta C_{d\to j} + \alpha_j - \alpha_d$. All such differences must be nonnegative at a candidate optimum, although checking single swaps alone does not establish optimality over all expert combinations. For policy q_0 at a different load, the predictive-risk difference is

$$L(q) - L(q_0) = \mathbb{E}_q D - \mathbb{E}_{q_0} D - \alpha^\top \{\rho(q) - \rho(q_0)\}. \tag{6}$$

With matching loads, the price term vanishes: reduced-cost and predictive improvements coincide.

Decision-relevant cost errors. The preceding results use true action costs, which must be estimated in practice. We identify the errors that can move the constrained policy. Some estimation errors leave the optimal assignment unchanged: for an integrable state offset $c(h)$ and a vector v, consider

$$C'_a(h) = C_a(h) + c(h) + v^\top g(a), \qquad \alpha' = \alpha - v \tag{7}$$

Every feasible objective changes by the same constant $\mathbb{E}c + v^\top \rho$, preserving its minimizers. The next theorem removes these directions and quantifies regret after refitting prices.

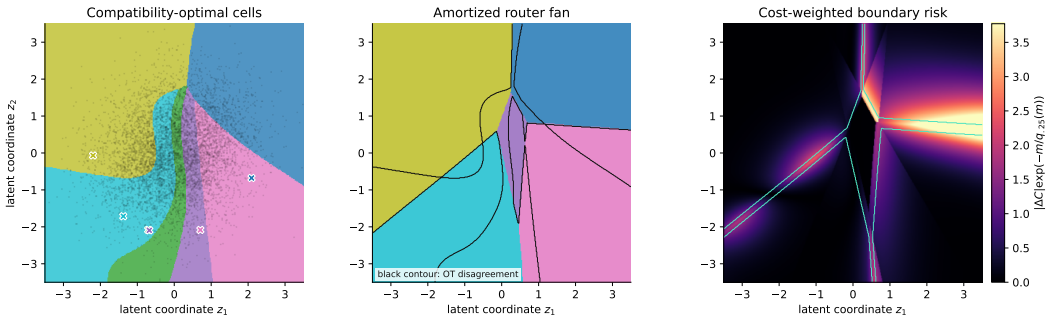

Figure 1: Three geometries in a controlled planar problem. Left: the compatibility-optimal partition. Center: the linear router partition and its disagreement with the transport optimum (black contour). Right: boundary proximity weighted by the cost of exchanging the proposed experts. The same visible boundary can carry different predictive risks under a common state distribution.

Theorem 3 (Decision-relevant cost error)**.** *Assume bounded costs and perturbations, $\nu_a \geq c > 0$, $\tau > 0$, and a relative-interior target. Let q_t optimize costs $C + te$, where $e_a(h)$ is cost error and $t \in \mathbb{R}$ its magnitude. Define $\mathcal{J}_C(q) = \mathbb{E}_{\mu q} C + \tau \mathbb{E}_\mu \mathrm{KL}(q\|\nu)$ and $\widetilde{z}_a = z_a - \sum_b q_{0,b} z_b$ for $z = e, g$. Set $K = \mathbb{E}_{\mu q_0} \widetilde{g}\widetilde{g}^\top$, $b_{ge} = \mathbb{E}_{\mu q_0} \widetilde{g}\widetilde{e}$; invert K only on $V = \mathrm{span}\{g(a) - g(b)\}$. Then*

$$\mathcal{J}_C(q_t) - \mathcal{J}_C(q_0) = \frac{t^2}{2\tau} \mathbb{E}_{\mu q_0} (\widetilde{e} - \widetilde{g}^\top K^{-1} b_{ge})^2 + O(t^3). \tag{8}$$

The squared capacity-projected error determines the leading regularized policy regret.

In Equation (8), centering removes statewise offsets, while the projection onto $\widetilde{g}$ removes errors absorbed by capacity prices (Appendix G). This coefficient converges to its hard-routing counterpart under the assumptions in Appendix H. Gradients of probabilistic action losses optimize centered costs, while gradients through mixed interfaces measure tangent output sensitivity (Appendix B).

4 Predictive Effects of Routing Changes

Figure 1 shows that router and predictive-cost partitions need not share boundaries. We quantify risk near these boundaries under the state distribution. Costly assignments can also occur inside routing cells, so the correction in Appendix M uses estimated gains both near and away from boundaries.

Data-weighted boundary risk. Let $s_i(h)$ be expert i's router score and $s_{(j)}$ the jth largest score. Only ties at the selection threshold change the expert set; where experts i, j exchange ranks k and $k+1$, write $u_{ij} = s_i - s_j$. For margin $m = s_{(k)} - s_{(k+1)}$, radius $\epsilon > 0$, and weight $W \geq 0$, define

$$\mathcal{B}_\epsilon(W) = \int_{\{m(h) \leq \epsilon\}} W(h) d\mu(h). \tag{9}$$

With $W = 1$, this is the probability mass near a boundary; $W = |C_a - C_b|$ instead weights it by the predictive-cost jump. For native action a_s, the weight $D_{a_s} - \min_a D_a$ measures regret relative to the best capacity-adjusted choice. The following law gives their small-ϵ behavior.

Theorem 4 (Cost-weighted tube law)**.** *Suppose μ has density f on a smooth r-dimensional manifold $\mathcal{M}$. In a compact regular boundary neighborhood, assume $u \in C^2$, $|\nabla_{\mathcal{M}} u| \geq c > 0$, and Wf is Lipschitz with compact support. For $\Gamma = \{u = 0\}$ and Hausdorff measure $\mathcal{H}^{r-1}$,*

$$\int_{|u| \leq \epsilon} W \, d\mu = 2\epsilon \int_\Gamma \frac{W(h) f(h)}{|\nabla_{\mathcal{M}} u(h)|} \, d\mathcal{H}^{r-1}(h) + O(\epsilon^2). \tag{10}$$

For a finite union of active exchange pieces with transverse intersections, the margin tube has the sum of their one-sided coefficients. For m_0 independent simultaneous tie constraints with uniformly full-rank Jacobian on the compact support, weighted intersections are $O(\epsilon^{m_0})$.

The coarea formula (Federer, 1969) separates three contributions: state density, inverse transverse variation of the score gap, and the chosen risk weight (Appendix C). For jump weights, interchangeable interfaces contribute nothing even when their boundary is frequently encountered. A rare higher-order intersection can matter at a fixed radius if its cost is large, although lower-order terms remain asymptotically dominant whenever all risk weights are uniformly bounded.

Alignment removes first-order boundary regret. A switch can occur even when both neighboring actions are optimal at the boundary. To distinguish such switches from costly assignments, fix the prices and define $R_s(h) = D_{a_s(h)}(h) - \min_a D_a(h)$. The next corollary gives the order of $\mathcal{B}_\epsilon(R_s)$.

Corollary 1 (Boundary alignment and vanishing first-order regret)**.** *Under the regular, active-piece assumptions of Theorem 4, suppose the finitely many reduced costs are locally Lipschitz. If the actions on both sides minimize reduced cost on each exchange boundary, then $\mathcal{B}_\epsilon(R_s) = O(\epsilon^2)$. If instead their nonnegative one-sided regret traces have a strictly positive weighted surface integral, then $\mathcal{B}_\epsilon(R_s) = c\epsilon + O(\epsilon^2)$, with the positive coefficient c determined by those boundary regret traces.*

Thus comparable switching frequencies can coexist with different leading orders of boundary regret. At a capacity-constrained optimum, neighboring actions tie in reduced cost; their raw predictive costs may still differ by the expert-price difference. This conclusion concerns boundaries, not all adjacent states. Even at equal load, tube regret can be zero, quadratic, or linear (Appendix C.2).

From cost error to boundary displacement. Estimated costs can displace an otherwise optimal boundary, linking this geometry to the estimation error in Theorem 3. Under the compactness, smoothness, and transverse-interface assumptions in Appendix H, take a two-expert top-1 optimal interface Γ, with $d\omega = f|\nabla_{\mathcal{M}}(C_2 - C_1)|^{-1} d\mathcal{H}^{r-1}$ and $W_0 = \omega(\Gamma) > 0$. Let q_t be the hard optimizer for $C + te$ at the original load. Then

$$L(q_t) - L(q_0) = \frac{t^2 W_0}{2} \operatorname{Var}_{\omega/W_0}(e_2 - e_1) + O(t^3). \tag{11}$$

Refitting prices centers $e_2 - e_1$; its weighted variance determines the leading risk increase. The multi-action coefficient is the hard-routing limit of Theorem 3 (Appendix H).

Determinants of the predictive-cost jump. The boundary laws use differences in predictive cost. We now ask where this predictive-cost jump comes from in the model's output distribution. Categorical logits yield a risk expansion separating miscalibration from curvature.

Theorem 5 (Predictive risk and Fisher curvature)**.** *Let interfaces induce $P_\eta(y) \propto \exp(\eta^\top T(y))$, with sufficient statistic T, log-partition A, and natural parameter η. The true conditional mean of T is $\bar{T}_\star$. For actions i, j, put $\delta = \eta_j - \eta_i$ and assume A has bounded third derivative along this segment. Then*

$$C_j - C_i = (\bar{T}_i - \bar{T}_\star)^\top \delta + \tfrac{1}{2}\delta^\top F(\eta_i)\delta + O(\|\delta\|^3), \tag{12}$$

where $\bar{T}_i = \nabla A(\eta_i)$ and $F = \nabla^2 A$ is Fisher information. At a calibrated reference $P_\star = P_i$, the linear term vanishes: the exact excess risk is the predictive divergence $\mathrm{KL}(P_i \| P_j)$.

At a calibrated reference, predictive divergence determines the excess risk; away from calibration, the linear term can change both its magnitude and sign. For a nonlinear continuation $\eta(\Phi)$, the risk Hessian also contains map curvature weighted by $\bar{T}_i - \bar{T}_\star$. The resulting signed risk change differs from the nonnegative magnitude used to weight a routing switch (Appendix C).

Repeated boundary encounters. The preceding laws average over states. A continuous perturbation can encounter the same boundary repeatedly, so we now replace state mass by a pathwise occupation measure. For a perturbation H_t on $[0, T_{\mathrm{end}}]$, let $X_t = u(H_t)$; here t parameterizes the perturbation, not the network's discrete layers. The next identity expresses weighted boundary occupation through local time, with quadratic variation specifying the integration measure.

Theorem 6 (Weighted occupation and local time)**.** *For a continuous semimartingale X with jointly continuous local time L_t^x and bounded continuous adapted $W(H_t)$, if $\mathbb{E}\sup_{|x|\le\epsilon_0} L^x_{T_{\mathrm{end}}} < \infty$,*

$$\lim_{\epsilon \downarrow 0} \frac{1}{2\epsilon} \mathbb{E} \int_0^{T_{\mathrm{end}}} W(H_t) \mathbf{1}\{|X_t| \le \epsilon\}\, d\langle X \rangle_t = \mathbb{E} \int_0^{T_{\mathrm{end}}} W(H_t)\, dL_t^0. \tag{13}$$

For $dH_t = b(H_t)dt + \Sigma(H_t)dB_t$, with drift b, diffusion Σ, and Brownian motion B_t, this measure is $d\langle X\rangle_t = \|\Sigma(H_t)^\top \nabla u(H_t)\|^2 dt$ (Appendix D). Relating occupation (Revuz & Yor, 1999) to switching loss requires the predictive-cost jumps as well as boundary visit counts.

From local actions to sequence loss. Multiple routing interventions introduce a further issue: an early change can alter the states seen at later stages. For S stages with Markov states, let π_0, π be native and modified policies, J_{seq} expected total loss, and d_ℓ^π the stage-ℓ state law. Define $A_\ell^{\pi_0}$ as forced continuation loss, including all remaining stages, minus its native-policy average. The performance-difference identity (Kakade & Langford, 2002) then gives

$$J_{\text{seq}}(\pi) - J_{\text{seq}}(\pi_0) = \sum_{\ell=1}^{S} \mathbb{E}_{H_\ell \sim d_\ell^\pi,\, a\sim\pi} A_\ell^{\pi_0}(H_\ell, a). \tag{14}$$

Because this expectation uses the modified state distribution, improvements measured only on native states do not establish an end-to-end gain. Full-forward reruns test that link, with advantage-estimation error and stagewise candidate restriction entering the bound separately (Appendix D).

5 Interpreting Routing Methods

COT benchmarks capacity allocation. We also analyze changes to outputs and selection gradients.

Capacity coordination. BASE, sparse OT, and Loss-Free Balancing coordinate competing uses of shared experts (Lewis et al., 2021; Liu et al., 2023b; Wang et al., 2024). With frozen scores and shared biases, Loss-Free Balancing's top-k rule is score-optimal at its realized load if no tokens are dropped or reassigned (Appendix N). This verifies the premise below at $\tau = \delta = 0$ for negative score sums, without assuming prescribed-load feasibility. More generally, a δ-optimal $\widehat{q}$ for costs $\widehat{C}$, under the same state law, actions, reference, temperature, and loads, satisfies

$$0 \le \mathcal{J}_C(\widehat{q}) - \mathcal{J}_C(q^\star) \le \delta + \inf_v \mathbb{E}_\mu \operatorname{osc}_a\{e_a - v^\top g(a)\}. \tag{15}$$

Here $q^\star$ minimizes $\mathcal{J}_C$, $e = \widehat{C} - C$, and osc_a is the range over actions. Errors absorbed by capacity prices vanish from the bound (Equation (7)). Under the local assumptions, residual errors determine regularized and hard-routing regret (Theorem 3; Equation (11)); regret can differ at equal MSE.

Continuous gates. A continuous gate changes the output path; predictive loss depends on that path. Write $\ell_h(z)$ for conditional loss after the complete frozen continuation and let $\phi(s) = \ell_h(\Phi_a + s(\Phi_b - \Phi_a))$. For $\phi \in C^2$ and $\theta \in [0,1]$, the gain from mixing over sampling these endpoints is

$$I_h(\theta) = (1-\theta)\phi(0) + \theta\phi(1) - \phi(\theta) = \int_0^1 K_\theta(s)\phi''(s)\, ds, \tag{16}$$

where $K_\theta(s) = \min(s,\theta)\{1 - \max(s,\theta)\} \ge 0$. Thus convexity along the complete output path favors mixing over endpoint sampling. On a compact regular boundary, assume the chart density and path losses are uniformly Lipschitz in the normal direction. For a fixed symmetric transition of width ϵ, the signed endpoint-selection effect is quadratic, while interpolation can contribute at first order:

$$R_{\text{sel}} - R_0 = O(\epsilon^2), \qquad R_{\text{mix}} - R_0 = -\epsilon B + O(\epsilon^2). \tag{17}$$

Here R_0, R_{sel}, and R_{mix} are the hard-routing, endpoint-sampling, and affine-mixture risks; the latter two use the same endpoint probabilities. The coefficient B integrates I_h over the boundary and transition profile, so a first-order gain requires $B > 0$ (Appendix N). SmoothSMoE follows this chord for isolated top-1 exchanges (Huu et al., 2026); top-k normalization also changes common-expert weights. That extra change can reverse the benefit even for convex loss. At a top-2 tie, take a common expert of output 0 and unnormalized weight 2, and two exchanging experts of output 1 and weight 1. The hard outputs both equal $1/3$, but activating both alternatives gives $1/2$; for target $1/3$, the squared loss increases to $1/36$. Additional expert calls count toward the compute budget.

ReMoE has a different mechanism: a newly active expert output e enters as $[s]_+ e$, vanishing at $s = 0$ (Wang et al., 2025). If the complete loss has bounded sensitivity and the boundary is regular, removing activations with $0 < s < \epsilon$ changes predictive risk by $O(\epsilon^2)$, while the changed-support mass is $O(\epsilon)$ (Appendix N). For a smooth continuation, a negative directional loss derivative makes sufficiently small activation useful. The cost of each active expert call remains discrete.

Sparse gradient estimates. SparseMixer and GRIN use local derivatives to approximate the selection part of a specified routing gradient (Liu et al., 2023a; 2024a). At a frozen single-expert interface, let $\phi(s) = \ell_h(sz)$ for weighted output z with fixed residual context. For a piecewise C^1 loss ϕ with integrable derivatives, finite one-sided traces, no endpoint jumps, and finitely many signed jumps $J_j = \phi(s_j^+) - \phi(s_j^-)$,

$$\phi(1) - \phi(0) = \int_0^1 \phi'(s)\, ds + \sum_j J_j. \tag{18}$$

On a smooth path, endpoint, midpoint and Heun evaluations approximate the integral; their errors depend on higher derivatives of the path loss. For example, the endpoint error is at most $\frac{1}{2} \sup_s |\phi''(s)|$. This explains why ordinary branch gradients can already agree with finite route comparisons when the path is nearly linear, and why a downstream route change can add information these derivatives miss. The gradient target also matters: GRIN's v2 stops the output-amplitude derivative, whereas v2* includes it. An accurate selection estimate alone need not give a descent direction when output amplitudes change. Both methods retain an endpoint rule on the argmax branch, so a higher-order quadrature bound alone does not bound their full estimator bias. Joint top-k rules require their sampling law and complete action costs, since expert choices can interact (Appendix N).

A cost-aware application. The analysis also yields a correction that ranks candidates by estimated reduced cost and filters them by calibrated gain and batch load (Appendix M). At fixed weights, experiments test boundary and loss predictions and the correction under compute and load constraints.

6 Experiments

We test the theory on 500 exact transport problems, 100 learned expert banks, 90 compact causal MoEs, five CIFAR-10 models, and a frozen 5.61B-parameter vision-language MoE. Independent units are problem seeds, checkpoints, or distinct images (Table 1); Appendix E specifies architectures, splits, candidates, and fitting. These are mechanism tests at fixed expert weights. Below, COT denotes corrections based on predictive costs and capacity prices. We test whether the laws hold for known costs, estimated costs support useful routing changes, and the same predictive quantities remain informative in trained image and vision-language models. Image smoothing can activate extra experts on the same inputs, so its budget is reported separately from native-count substitutions.

6.1 Boundary frequency versus predictive harm

Exact problems vary nonlinearity, capacity, and state anisotropy. First-order tube exponents are 0.993–1.001; independent simultaneous ties give $1.00, 1.99, 3.02, 3.94$ for orders one to four (Figure 2a). Compatibility-weighted occupation explains switching cost with $R^2 = 0.86$–0.88, versus 0.33–0.47 for unweighted occupation across uniform-load, skewed-load, and anisotropic settings (Figure 2b).

In learned banks, the cost head reaches held-out $R^2 = 0.962$. COT reduces test cost by 0.165 under uniform load and 0.240 under skewed load, with normalized load TV below 0.010. The exact sufficient-label control has zero hard regret (Proposition 1): retaining a common minimizing action suffices for optimal decisions. This separates identifying the best action from predicting its cost.

System	Independent units	Quantity isolated
Exact transport	5×100 problems	Known costs and optima; tube and occupation laws
Learned expert banks	2×50 banks	Estimated cost under uniform or skewed capacity
Compact causal MoE	3×30 models	Context-dependent risk; complete sequence repair
Four-layer image MoE	5 models	Same-weight continuity and predictive jump on CIFAR-10
Vision-language MoE	209 image clusters	Complete top-2 action costs at image and answer positions

Table 1: Validation units and isolated quantities. All paired comparisons fix expert weights.

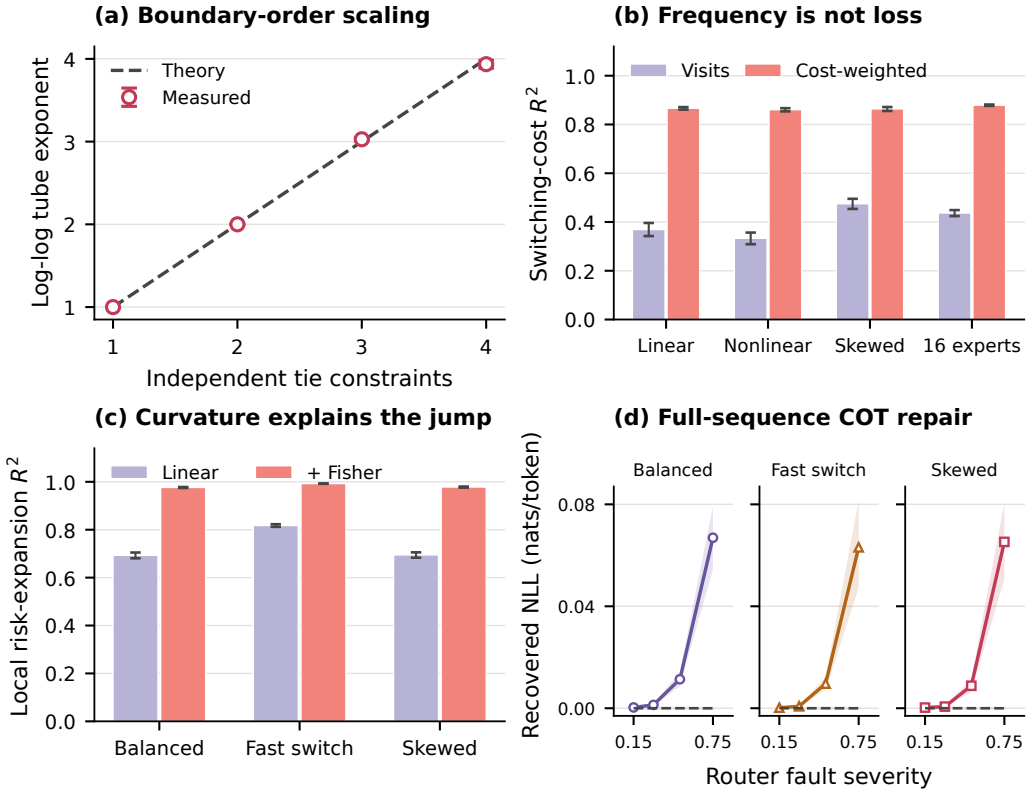

Figure 2: From geometry to predictive risk. **(a)** Tube exponents follow the number of independent tie constraints. **(b)** Cost weighting improves prediction of switching loss under stochastic perturbations. **(c)** Adding Fisher curvature improves the signed risk expansion at interpolation scale $t = 0.20$. **(d)** Shared-scale views show COT recovery; dashed lines repeat the pooled margin comparator. Panels (a,b) use 100 problems per setting; (c,d) use 30 independently trained checkpoints per causal process.

6.2 Learned causal interfaces and full-sequence repair

The causal MoE (157,320 parameters, eight experts) learns its experts and router on three hidden-regime processes varying persistence and balance. All tokens occur in all regimes; the generating regime supplies audit reference costs but is never a router or head input. We evaluate all eight actions and replay corrections through complete observed sequences. Sequence replay captures changes in later states that summing gains along the native trajectory would miss. Native compatibility percentiles are 0.977–0.986, with oracle headroom 0.053–0.089 nats/token. At logit interpolation scale $t = 0.20$, adding Fisher curvature raises R^2 from 0.693–0.817 to 0.976–0.992 (Figure 2c); remainder orders change from 2.10–2.14 to 3.02–3.07. After injecting a routing fault of severity 0.75, COT recovers 0.063–0.067 nats/token (Figure 2d) with load TV 0.0023–0.0026, while margin-based recovery is statistically indistinguishable from zero at comparable load. Appendix F reports the full severity sweep, condition-specific estimates, and candidate-oracle comparison. Repair preserves expert weights and active expert count to isolate routing effects on sequence loss.

6.3 Same-weight image boundaries and cost controls

The four-layer MoE has 2.51M parameters, 32 experts, and top-4 dispatch (Appendix E.3), using the official CIFAR-10 split (Krizhevsky, 2009). Five checkpoints use 40,000/5,000/5,000 training/cost-fitting/calibration images and 10,000 test images. At the third layer, we compare native routing, the smoothstep rule of Huu et al. (2026), and cost/affinity corrections using five joint actions from the top-eight experts. Fixed checkpoints isolate routing interventions from retraining effects. Continuity tests probe interfaces near ties; held-out loss measures the downstream effect of expert-set changes.

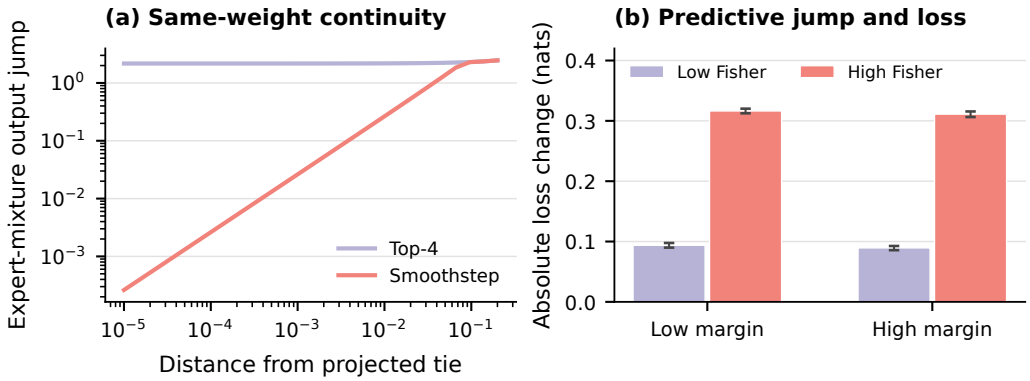

Figure 3: Same-weight CIFAR-10 boundary tests. **(a)** Symmetric normal perturbations of the third-layer gate, averaged over 128 test states per seed. **(b)** Absolute held-out loss changes, split at each checkpoint's median margin and median Fisher displacement. Both panels aggregate five independently trained checkpoints evaluated on the same untouched test images.

At projected fourth/fifth-score ties, hard dispatch retains a finite jump; smoothstep has local exponent 1.000 across checkpoints. Fisher displacement correlates with absolute candidate loss change at Spearman 0.707, versus -0.027 for margin. High-Fisher alternatives produce larger changes in both margin strata (Figure 3); we average absolute changes over four alternatives per test image and split at within-checkpoint medians. Calibrated correction changes 4.28% of assignments with load TV 0.00239 and four active experts; Appendix F.4 reports paired policy quality and full-model runtimes.

A pretrained vision-language test. For MoE-LLaVA (5.61B parameters, four experts, top-2), we audit all six expert pairs at three layers and two positions on 256 AI2D questions from 209 diagrams (Lin et al., 2026; Kembhavi et al., 2016). Fisher displacement correlates with absolute answer-loss changes at $0.915/0.940$ for image/answer positions, versus $0.051/-0.081$ for margin. All 1,536 identity checks pass; each intervention recomputes the remaining decoder at fixed weights and expert count (Appendix F.5). Appendix L extends this audit across datasets and backbones.

Route derivatives and complete loss differences. We compare complete route losses with rank-2/rank-3 exchange derivatives in a six-layer, eight-expert language model to probe downstream support changes. These probes follow an expert-exchange path, while SparseMixer and GRIN use zero-to-output paths. Six checkpoints come from two training seeds. Each layer uses $1{,}024$ probes and $16{,}384$ held-out tokens per checkpoint. Updates use eight independently reset groups; the uncertainty intervals resample both these groups and validation documents while holding the checkpoint fixed. At the final MoE layer, loss-weighted directional disagreement is 0.37–1.18%; at layer 3 it is 3.87–9.90%. Fixing downstream expert IDs on the same layer-3 probes, while recomputing states and weights, reduces it to 0.57–2.30%. This control isolates a contribution from later support changes. At the prespecified displacement of 0.001 times the router parameter norm, all six conditional endpoint-versus-pairwise-gradient validation intervals include zero at each layer. Conditional on these checkpoints, extra complete-route evaluations establish no consistent advantage in validation loss over the matched pairwise-gradient updates (Appendix N).

7 Conclusion

A routing boundary matters through the predictive cost it carries, not simply because it is discontinuous. Compatibility optimal transport makes this distinction precise: downstream risk defines the assignment cost, expert capacity supplies dual prices, and reduced costs give testable optimality conditions. Capacity projection identifies decision-relevant cost error, whose smooth regret coefficient converges to a weighted hard-boundary variance. Fisher curvature and weighted occupation connect interface changes to predictive harm. These quantities explain how capacity coordination, continuous gates, and sparse gradient estimates affect prediction under explicit compute and load constraints.

Ethics Statement

This work studies how routing decisions affect predictive loss in frozen mixture-of-experts models. A reduction in average predictive loss does not establish fairness, safety, or reliability for particular user groups or deployment settings. Routing corrections can retain biases and harmful behavior inherited from the underlying models and data. Applications therefore require separate evaluations of these properties, together with compliance with the usage conditions of the underlying models and datasets.

Reproducibility Statement

The theoretical assumptions and proofs are provided in Appendices B–D and G–J. Appendices E–F describe the experimental protocols and additional checks, including data splits, independent evaluation units, and training settings. The pretrained vision-language audit specifies the checkpoint revision and intervention procedure in Appendix F.5. Appendix K brings together cost fitting, price estimation, calibration, and finite-batch execution to connect evaluation with the theoretical quantities.

AI Use Statement

We used AI assistants to help translate and revise manuscript text, improve the organization and presentation of theoretical statements and proofs, and provide feedback on the exposition of the research methodology. We also used these tools to locate and check references, inspect terminology and cross-references, and assist with LaTeX formatting and compilation. These uses supported manuscript preparation and review; they do not replace author verification of mathematical claims, references, or experimental evidence. The authors take responsibility for the paper's final content.

REFERENCES

Ravindra K. Ahuja, Thomas L. Magnanti, and James B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, 1993. URL https://mitmgmtfaculty.mit.edu/jorlin/network-flows/.

Jason M. Altschuler, Jonathan Niles-Weed, and Austin J. Stromme. Asymptotics for semi-discrete entropic optimal transport. *SIAM Journal on Mathematical Analysis*, 54(2):1718–1741, 2022. doi: 10.1137/21M1440165. URL https://arxiv.org/abs/2106.11862.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101. URL https://arxiv.org/abs/2107.07511.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013. URL https://papers.neurips.cc/paper_files/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

Herbert Federer. *Geometric Measure Theory*, volume 153 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, 1969.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. URL https://www.jmlr.org/papers/v23/21-0998.html.

X. Y. Han and Yuan Zhong. A theoretical framework for auxiliary-loss-free load balancing of sparse mixture-of-experts in large-scale AI models, 2026. URL https://arxiv.org/abs/2512.03915v3. Version 3, April 2026; first posted December 2025.

Wentao Hu, Yanbo Zhai, Xiaohui Hu, Mingkuan Zhao, Shanhong Yu, Xue Liu, Kaidong Yu, Shuangyong Song, and Xuelong Li. Awakening dormant experts: Counterfactual routing to mitigate MoE hallucinations, 2026. URL https://arxiv.org/abs/2604.14246.

Tho Tran Huu, Huu-Tuan Nguyen, Thien-Hai Nguyen, Nhat-Tri Ho, Viet-Hoang Tran, Tho Quan, and Tan Minh Nguyen. Geometric and stochastic analysis of discontinuities in sparse mixture-of-experts, 2026. URL https://arxiv.org/abs/2606.19036.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79. URL https://www.cs.toronto.edu/~hinton/absps/jjnh91.pdf.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 267–274, 2002. URL https://www.cs.cmu.edu/~jcl/papers/aoarl/Final.pdf.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision*, pp. 235–251, 2016. doi: 10.1007/978-3-319-46493-0_15.

Jun Kitagawa, Quentin Mérigot, and Boris Thibert. Convergence of a Newton algorithm for semi-discrete optimal transport. *Journal of the European Mathematical Society*, 21(9):2603–2651, 2019. doi: 10.4171/JEMS/889.

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

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2006.16668.

Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. BASE layers: Simplifying training of large, sparse models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6265–6274. PMLR, 2021. URL https://proceedings.mlr.press/v139/lewis21a.html.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023. doi: 10.18653/v1/2023.emnlp-main.20.

Zhongyang Li, Ziyue Li, and Tianyi Zhou. Routing manifold alignment improves generalization of mixture-of-experts LLMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/249b8d8f41970822651435629e68a6e1-Abstract-Conference.html.

Bin Lin, Zhenyu Tang, Yang Ye, Jinfa Huang, Junwu Zhang, Yatian Pang, Peng Jin, Munan Ning, Jiebo Luo, and Li Yuan. MoE-LLaVA: Mixture of experts for large vision-language models. *IEEE Transactions on Multimedia*, 28:4408–4419, 2026. doi: 10.1109/TMM.2026.3654458. URL https://arxiv.org/abs/2401.15947.

Liyuan Liu, Jianfeng Gao, and Weizhu Chen. Sparse backpropagation for MoE training, 2023a. URL https://arxiv.org/abs/2310.00811v1. Version 1.

Liyuan Liu, Young Jin Kim, Shuohang Wang, Chen Liang, Yelong Shen, Hao Cheng, Xiaodong Liu, Masahiro Tanaka, Xiaoxia Wu, Wenxiang Hu, Vishrav Chaudhary, Zeqi Lin, Chenruidong Zhang, Jilong Xue, Hany Awadalla, Jianfeng Gao, and Weizhu Chen. GRIN: GRadient-INformed MoE, 2024a. URL https://arxiv.org/abs/2409.12136v1. Version 1.

Tianlin Liu, Joan Puigcerver, and Mathieu Blondel. Sparsity-constrained optimal transport. In *International Conference on Learning Representations*, 2023b. URL https://arxiv.org/abs/2209.15466.

Tianlin Liu, Mathieu Blondel, Carlos Riquelme, and Joan Puigcerver. Routers in vision mixture of experts: An empirical study. *Transactions on Machine Learning Research*, 2024b. URL https://arxiv.org/abs/2401.15969.

Duc Anh Nguyen, Huu Binh Ta, Duc Nhuan Le, Tan Minh Nguyen, and Toan Tran. Selective sinkhorn routing for improved sparse mixture of experts, 2025. URL https://arxiv.org/abs/2511.08972v2. arXiv:2511.08972, version 2, revised June 2026.

Gabriel Peyré and Marco Cuturi. Computational optimal transport with applications to data sciences. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. doi: 10.1561/2200000073. URL https://optimaltransport.github.io/book/.

Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/79fea214543ba263952ac3f4e5452b14-Abstract-Conference.html.

Daniel Revuz and Marc Yor. *Continuous Martingales and Brownian Motion*. Springer, 3 edition, 1999. doi: 10.1007/978-3-662-06400-9.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8583–8595, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/48237d9f2dea8c74c2a72126cf63d933-Abstract.html.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candes. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/5103c3584b063c431bd1268e9b5e76fb-Abstract.html.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1701.06538.

Lean Wang, Huazuo Gao, Chenggang Zhao, Xu Sun, and Damai Dai. Auxiliary-loss-free load balancing strategy for mixture-of-experts, 2024. URL https://arxiv.org/abs/2408.15664v1. Version 1.

Xi Wang, Soufiane Hayou, and Eric Nalisnick. The myth of expert specialization in MoEs: Why routing reflects geometry, not necessarily domain expertise, 2026. URL https://arxiv.org/abs/2604.09780.

Ziteng Wang, Jun Zhu, and Jianfei Chen. ReMoE: Fully differentiable mixture-of-experts with ReLU routing. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/94dc604e115237a7f4a758b3146cd976-Abstract-Conference.html.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. URL https://arxiv.org/abs/2412.10302.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Y. Zhao, Andrew M. Dai, Zhifeng Chen, Quoc V. Le, and James Laudon. Mixture-of-experts with expert choice routing. In *Advances in Neural Information Processing Systems*, volume 35, pp. 7103–7114, 2022. doi: 10.52202/068431-0515.

Appendix

Appendix contents

A. Notation . . . 14
B. Compatibility transport and identification proofs . . . 15
C. Cost-weighted geometry proofs . . . 17
D. Stability and correction proofs . . . 19
E. Experimental protocols . . . 22
F. Controlled results and additional checks . . . 24
G. Capacity-constrained cost error and routing response . . . 31
H. Boundary displacement and the hard-routing limit . . . 35
I. Joint actions and finite-batch certificates . . . 40
J. Capacity response to distribution changes . . . 43
K. From cost estimation to calibrated execution . . . 46
L. Additional datasets, architectures, and routing controls . . . 49
M. Correcting costly assignments . . . 52
N. Interpreting existing routing methods . . . 53

The appendices follow the same order as the argument in the main text. We first establish the transport problem and identify the quantities that determine a decision (Appendix B). We then study how those quantities behave near routing boundaries (Appendix C) and under cost or continuation error (Appendix D). The remaining appendices turn the results into reproducible protocols: the controlled experiments come first (Appendices E–F), followed by capacity and distribution-shift analyses (Appendices G–J), and the end-to-end calibration contract (Appendix K).

A Notation

Symbol	Meaning
H, h, μ	Complete routing state, its realization, and its probability law
$\mathcal{A}, a, S_a$	Finite action family, action index, and active expert set
k, E, M	Active experts, total experts, and candidate actions
Φ_a, P_a, C_a	Expert interface, downstream prediction, and conditional risk
$g(a), \rho(q), \rho$	Expert dispatch vector, expected dispatch, and target dispatch
$\nu, p, q, q^\star$	Reference kernel, native policy, candidate policy, and transport optimum
τ, α, D_a	Temperature, capacity prices, and reduced cost
$\mathcal{J}_\nu, L, J_{\mathrm{seq}}$	Regularized objective, unregularized risk, and sequence loss
s_i, u_{ij}, m, τ_r	Gate score, pairwise gap, threshold margin, and router temperature
$\mathcal{M}, r, f$	State manifold, intrinsic dimension, and density
$\Gamma, W, \mathcal{B}_\epsilon$	Active exchange boundary, risk weight, and weighted tube mass
$\eta, A, F, \bar{T}_\star$	Natural parameter, log-partition, Fisher matrix, and target mean

Table 2: Important notation, first part.

Symbol	Meaning
$X_t, L_t^x, \langle X \rangle$	Boundary-gap process, local time, and quadratic variation
$d_\ell^\pi, A_\ell^{\pi_0}, S$	Induced state law, continuation-loss advantage, and stage count
$\widehat{C}, \lambda, \widehat{G}, G$	Learned cost, fitted prices, predicted gain, and true gain
u_δ, ξ	Group-calibrated residual quantile and allowed load deviation
ℓ_a, G^{obs}	Observed log loss and realized reduced-cost improvement
B, d_f, d_c	Routed states in a batch, input dimension, and hidden width
$e, \widetilde{e}, r_a$	Cost perturbation, state-centered error, and projected error
V, K, b_{ge}	Load-difference span, load covariance, and load–error covariance
$\omega_{ab}, G_\partial, b_\partial$	Weighted interface measure, boundary Gram matrix, and error moment
z, d, μ_t	Density perturbation, induced demand change, and perturbed state law

Table 3: Important notation, second part. Costs and temperatures are in nats; binary dispatch satisfies $\sum_i g_i(a) = k$.

B Compatibility Transport and Identification Proofs

We begin with the population problem that underlies every later certificate. Let $(\mathcal{H}, \mathfrak{H}, \mu)$ be a probability space. All kernels below are measurable and the action set is finite. Action weights may depend on state through a fixed rule $w_a(h)$; the selected set and dispatch vector $g(a)$ are fixed for that action index. This includes joint sparse interfaces without an additive expert-cost assumption.

B.1 Transport representation and existence of prices

We first express a routing kernel as a coupling. This representation makes the load constraint explicit and leads to the dual problem used to construct the prices. For top-1, set $\pi(B \times \{i\}) = \int_B q_i(h) d\mu(h)$. Its marginals are μ and ρ. Conversely the measures $\pi(\cdot, \{i\})$ are dominated by μ; their Radon–Nikodym derivatives form a probability kernel whose expected load is ρ. Moreover $\int C_i(h) d\pi = \mathbb{E}_\mu \sum_i q_i C_i$. This proves the semi-discrete transport representation. For top-k, use the same disintegration on $\mathcal{H} \times \mathcal{A}$ and impose $\int g(a) d\pi = \rho$. This is a finite-codomain transport problem with linear moments, not in general transport to individual experts.

Proof of Theorem 1. Define the convex dual minimization function

$$\Psi(\alpha) = \tau \mathbb{E}_\mu \log \sum_a \nu_a(H) e^{-(C_a(H) + \alpha^\top g(a))/\tau} + \alpha^\top \rho.$$

It is finite and continuous: the log-sum-exp is bounded above and below by integrable finite maxima of $|C_a| + \tau |\log \nu_a|$, plus a constant depending on $\|\alpha\|$. Let $V = \mathrm{span}\{g(a) - g(b) : a, b \in \mathcal{A}\}$. On $V^\perp$, $\alpha^\top g(a)$ is constant in a and equals $\alpha^\top \rho$, so Ψ and the Gibbs kernel are invariant. On V, relative interior supplies a constant $c > 0$ such that

$$\max_a \alpha^\top (\rho - g(a)) \geq c \|\alpha\|.$$

Indeed a relative ball of radius c centered at ρ lies in the marginal polytope. Taking its support function in direction $-\alpha$ proves the inequality. The lower log-sum-exp bound then gives

$$\Psi(\alpha) \geq c\|\alpha\| - \mathbb{E} \max_a |C_a| - \tau \mathbb{E} \max_a |\log \nu_a|.$$

Thus Ψ is coercive on finite-dimensional V and attains a minimum.

Differentiating under the integral is justified by bounded g:

$$\nabla \Psi(\alpha) = \rho - \mathbb{E}_\mu \sum_a q_a^\alpha g(a), \qquad \nabla^2 \Psi(\alpha) = \tau^{-1} \mathbb{E}_\mu \, \mathrm{Cov}_{q^\alpha}(g).$$

The gradient belongs to V, so stationarity at a minimizer on V gives the required marginal constraint. For nonzero $v \in V$, the values $v^\top g(a)$ cannot all coincide. Positivity of every q_a^α makes $v^\top \nabla^2 \Psi v > 0$, proving uniqueness of the price representative on V.

For any feasible q, direct expansion of the finite-action entropy gives

$$\mathcal{J}_\nu(q) = -\Psi(\alpha) + \tau \mathbb{E}_\mu \mathrm{KL}(q(\cdot \mid H) \| q^\alpha(\cdot \mid H))$$

when α is a feasible Gibbs price. The divergence is nonnegative and vanishes only for $q = q^\alpha$ almost surely. Together with the minimizing price constructed above, this identity proves existence, optimality, and uniqueness of the routing kernel. Integrability is legitimate because $q_a \log q_a$ is bounded on the finite simplex and $C_a, \log \nu_a$ are integrable.

For the unregularized problem with an attained finite dual price, subtracting the dual bound from primal cost yields

$$\mathbb{E}_\mu \sum_a q_a(h)\{C_a(h) + \alpha^\top g(a) - \min_b (C_b(h) + \alpha^\top g(b))\} = 0.$$

Every summand is nonnegative, so an optimum assigns mass only to minimizing actions. Conversely a feasible kernel with this support attains the dual bound. For a zero-temperature sequence whose price representatives converge, ratios between a strictly suboptimal action and a minimizing action tend to zero pointwise by the exponential formula. This limiting statement is conditional on price convergence and keeps the action family fixed. □

For upper capacities, the Lagrangian adds $\alpha^\top(\rho(q) - b)$ with $\alpha \geq 0$. Under a suitable strict-feasibility condition, minimizing over q gives the same Gibbs kernel and the dual includes $-\alpha^\top b$. Complementary slackness reads $\alpha_i(\rho_i(q^\star) - b_i) = 0$. For fixed-target equality constraints, prices need not be nonnegative. At boundary targets, some actions can be forced to have zero mass; the strictly positive, finite-price theorem is then applied on the admissible face; it is not asserted on the full action family.

B.2 Identification, compensation, and gradients

The previous result characterizes the optimizer. We next record the converse identities that let us recognize that optimizer from observed routing probabilities, compare policies with different loads, and differentiate the two objectives used in the paper.

Proof of Theorem 2. Divide the two Gibbs probabilities in Equation (3); the normalizer cancels, giving Equation (4). Conversely, fix any reference action b. The pairwise restriction implies p_a/ν_a is proportional to $e^{-(C_a + \alpha^\top g(a))/\tau}$, with an action-independent factor. Normalization determines that factor as $1/Z_\alpha$. The assumed load feasibility then makes p the unique optimum.

To establish the exact identity for any feasible q, expand

$$\mathbb{E}\mathrm{KL}(q\|\nu) - \mathbb{E}\mathrm{KL}(q^\star\|\nu) = \mathbb{E}\mathrm{KL}(q\|q^\star) + \mathbb{E}\sum_a (q_a - q_a^\star)\log(q_a^\star/\nu_a).$$

Stationarity gives $\tau \log(q_a^\star/\nu_a) = -C_a - \alpha^\top g(a) - \tau \log Z_\alpha$. Pointwise normalization removes the normalizer term, while matched expected load removes the price term. The remaining term cancels $L(q) - L(q^\star)$, giving Equation (5). □

The same cancellation also shows what changes when loads differ. For arbitrary policies, add and subtract the price inside the cost expectation: $\mathbb{E}_q D = L(q) + \alpha^\top \rho(q)$. Subtracting the corresponding expression for q_0 proves Equation (6). A swap of d for j changes binary dispatch by $e_j - e_d$, so it has reduced-cost gap $\Delta C_{d\to j} + \alpha_j - \alpha_d$. Nonnegative gaps are necessary and sufficient only for optimality on the declared native-plus-swap family.

For router temperature $\tau_r > 0$, let $q_b = e^{s_b/\tau_r}/\sum_c e^{s_c/\tau_r}$. Then $\partial q_b/\partial s_a = q_b(\mathbf{1}\{a = b\} - q_a)/\tau_r$. Multiplication by D_b and summation give

$$\partial_{s_a} \mathbb{E}_q D = q_a(D_a - \mathbb{E}_q D)/\tau_r. \tag{19}$$

For a differentiable mixture $\bar{\Phi} = \sum_a q_a \Phi_a$ with q-independent interfaces,

$$\frac{\partial \ell(\bar{\Phi})}{\partial s_a} = \frac{q_a}{\tau_r}\langle \nabla_\Phi \ell, \Phi_a - \bar{\Phi}\rangle.$$

Differentiating $\alpha^\top \sum_b q_b g(b)$ supplies the centered load term. The first derivative averages complete-action costs; the second differentiates the loss of a mixed interface. Keeping this distinction explicit determines which objective the update below optimizes. To see the associated mirror step, minimize $\sum_a q_a D_a + \eta^{-1}\mathrm{KL}(q\|q^{\mathrm{old}})$ on the simplex. Stationarity and normalization give $q_a^+ \propto q_a^{\mathrm{old}} e^{-\eta D_a}$. For the mixture loss, replace D_a by its first-order interface derivative; the resulting step minimizes the linearized objective, not the full nonlinear loss.

B.3 Label sufficiency as a decision-theoretic consequence

The preceding identities concern complete costs. The following proposition isolates a simpler question: when can a coarsened state label still support an optimal action? This separates information loss from the additional coupling introduced by capacity constraints.

Proposition 1 (Information sufficient for an optimal action)**.** *For a sub-σ-algebra $\mathfrak{A} \subseteq \sigma(H)$ and finite actions with integrable costs, define the best $\mathfrak{A}$-measurable decision risk $R(\mathfrak{A})$. Then*

$$R(\mathfrak{A}) = \mathbb{E} \min_a \mathbb{E}[C_a \mid \mathfrak{A}].$$

If $\mathfrak{A} \subseteq \mathfrak{B}$, then $R(\mathfrak{A}) \geq R(\mathfrak{B})$. Equality holds exactly when an $\mathfrak{A}$-measurable selector attains the $\mathfrak{B}$-conditional minimum almost surely.

Proof. For an $\mathfrak{A}$-measurable selector $a(H)$, iterated expectation gives $\mathbb{E}C_{a(H)} = \mathbb{E}\sum_j \mathbf{1}\{a(H) = j\}\mathbb{E}[C_j \mid \mathfrak{A}]$. Choose the smallest-index minimizer of the finitely many conditional costs; it is measurable and attains the displayed minimum. Every coarse rule is also fine-measurable, proving monotonicity. If equality holds, this coarse minimizer has zero expected gap above the fine conditional minimum. The gap is nonnegative, so it is zero almost surely. Conversely, any coarse selector with this property attains the fine risk. □

Preserving the entire cost vector is sufficient but not necessary: preserving a common minimizing action suffices. Costs can vary arbitrarily within a label cell when the same action remains optimal. This is an unconstrained decision result; capacity-coupled decisions require the reduced-cost and feasibility conditions as well. The sufficient-label control in our exact suite tests the case in which this information reduction preserves an optimal action.

A related invariance concerns the costs themselves: some changes to their numerical values leave every feasible comparison unchanged. For Equation (7), pointwise normalization implies $\mathbb{E}_q c(H) = \mathbb{E}_\mu c(H)$ for every policy. Load feasibility implies $\mathbb{E}_q v^\top g = v^\top \rho$. Thus the transformed objective differs by a policy-independent constant. Substituting $\alpha' = \alpha - v$ into its Gibbs form leaves only the common factor $e^{-c(h)/\tau}$, which cancels in normalization. This proves both optimizer invariance and the stated price transformation.

C Cost-Weighted Geometry Proofs

C.1 Regular boundary tubes and their intersections

We now move from optimization to geometry. The relevant mass is the mass of states that can actually change action, so we first derive the tube expansion for one regular tie and then explain how active intersections alter its order.

Proof of the pairwise part of Theorem 4. Write $b = Wf$. By coarea on $\mathcal{M}$,

$$\int_{|u|\leq\epsilon} b\, d\,\mathrm{vol}_{\mathcal{M}} = \int_{-\epsilon}^{\epsilon} \mathcal{I}(t)dt, \qquad \mathcal{I}(t) = \int_{u^{-1}(t)} \frac{b}{|\nabla_{\mathcal{M}} u|}\, d\mathcal{H}^{r-1}.$$

On the compact regular neighborhood the vector field $V = \nabla_{\mathcal{M}} u/|\nabla_{\mathcal{M}} u|^2$ is continuously differentiable with bounded derivative. Its local flow Ξ_t satisfies $u(\Xi_t(h)) = u(h) + t$. Compactness yields a common small time interval and bounded surface Jacobians. Pulling $\mathcal{I}(t)$ back to the zero level gives

$$\mathcal{I}(t) = \int_\Gamma \frac{b(\Xi_t(h))}{|\nabla_{\mathcal{M}} u(\Xi_t(h))|} J_{r-1}\Xi_t(h)\, d\mathcal{H}^{r-1}(h).$$

A finite chart cover and partition of unity may be used if the neighborhood is not a single chart. The Lipschitz numerator, positive gradient lower bound, and C^1 flow imply $|\mathcal{I}(t) - \mathcal{I}(0)| \le K|t|$. Integrating over the symmetric interval proves Equation (10), with remainder at most $K\epsilon^2$. The constant K depends on the surface measure in the finite chart cover, the Lipschitz norm of Wf, the C^2 bounds of u, and the positive lower gradient bound. In particular, the factor $|\nabla_{\mathcal{M}} u|^{-1}$ remains inside the surface integral; bounded density alone does not give this remainder. □

To pass from one tie to the top-k margin, retain only boundaries that can change the selected set. Index these active exchange pieces Γ_γ. Away from their intersections, exactly two experts straddle the top-k threshold, and $m = |u_\gamma|$. If W has bounded one-sided Lipschitz extensions $W_\gamma^\pm$, apply the preceding argument separately on the two sides. Their combined coefficient is

$$\mathcal{B}_\epsilon(W) = \epsilon \sum_\gamma \int_{\Gamma_\gamma} \frac{(W_\gamma^+ + W_\gamma^-)f}{|\nabla_{\mathcal{M}} u_\gamma|}\, d\mathcal{H}^{r-1} + O(\epsilon^2).$$

Finitely many regular pieces, compact support, and transverse codimension-two intersections ensure that overlapping strips have weighted mass $O(\epsilon^2)$; inclusion–exclusion changes only the remainder. The active-rank restriction is essential: ties between two already selected experts, or between two excluded experts, do not necessarily alter the route.

At a multiple tie, the tube order is determined by the number of independent transverse coordinates. For the higher-order claim, let $U = (u_1, \dots, u_{m_0})$ be independent tie constraints and let $J_U = \sqrt{\det(D_{\mathcal{M}} U\, D_{\mathcal{M}} U^\top)}$. Assume $m_0 \le r$ and the rank is uniformly m_0 in the compact neighborhood. Vector coarea gives

$$\int_{\|U\|_\infty \le \epsilon} W\, d\mu = \int_{[-\epsilon,\epsilon]^{m_0}} \left(\int_{U^{-1}(t)} \frac{Wf}{J_U}\, d\mathcal{H}^{r-m_0} \right) dt.$$

A finite collection of constant-rank charts bounds the inner integrals uniformly. The cube has volume $(2\epsilon)^{m_0}$, proving the stated order. Counting all pairwise equalities in a multiple tie overcounts codimension; only independent constraints determine the exponent. For example, three equal scores impose two independent constraints, not three.

The same local expansion applies to a nonnegative reduced-cost regret weight when its one-sided regularity assumptions hold. Neither linear empirical mass over an arbitrary range nor concentration of global model loss near low margins follows without further distributional evidence.

C.2 Boundary alignment and the equal-load example

The tube law gives a surface-integral coefficient. We now identify when its linear term vanishes. The next corollary identifies when the linear term disappears and then gives an equal-load construction showing why visible boundary size alone does not determine predictive harm.

Proof of Corollary 1. The minimum of finitely many locally Lipschitz functions is locally Lipschitz. Hence $D_a - \min_b D_b$ has Lipschitz extensions on the two sides of each regular active piece. If both selected actions minimize D on that piece, both regret traces are zero. Substituting these traces into the one-sided coarea expansion makes its linear coefficient vanish. The bounded remainder and transverse intersection terms are $O(\epsilon^2)$. Equivalently, normal flow from a boundary point changes each regret by at most a constant times $|u|$, and integrating $|u|$ over a regular tube has order ϵ^2. If the weighted surface integral of the traces is positive, the same expansion has a positive linear coefficient. Nonnegative regret precludes cancellation between pieces or sides. □

To see the distinction explicitly, take $H = (X, Z)$ uniform on $[-1, 1]^2$, target $Y \mid H \sim \mathcal{N}(X, 1)$, and predictions $\mathcal{N}(+1, 1)$ and $\mathcal{N}(-1, 1)$. Compare the optimal router, its reversed orientation, and the tilted router selecting $+$ for $X - bZ > 0$, $0 < b < 1$. Every policy has equal load and small margin-tube mass ϵ, whereas

$$\mathcal{B}_\epsilon(R_s) = \begin{cases} 0, & \text{optimal orientation,} \\ \epsilon^2, & \text{reversed orientation,} \\ \frac{b}{2}\epsilon + O(\epsilon^2), & \text{tilted boundary.} \end{cases} \tag{20}$$

Gaussian cross-entropy has the form $C_{\pm}(x,z) = \kappa + \frac{1}{2}(x \mp 1)^2$, where $\kappa = \frac{1}{2}\log(2\pi) + \frac{1}{2}$. Thus the optimal unconstrained action is $+$ for $x > 0$, and this action already has the required load $1/2$. Zero prices solve the equality-constrained problem. Central symmetry $(x,z) \mapsto (-x,-z)$ makes all three compared policies load-feasible. Use score differences $u = x$, $u = -x$, and $u = x - bz$, respectively; the corresponding top-1 margin is $|u|$.

For $\epsilon < 1 - b$, the tilted strip has conditional x-length 2ϵ for every z, so its probability is ϵ, just like the two vertical strips. The correctly oriented policy has zero regret throughout the square. For the reversed policy, regret is $2|x|$, and therefore

$$\mathcal{B}_\epsilon(R_{\text{reversed}}) = \frac{1}{2}\int_{-\epsilon}^{\epsilon} 2|x|dx = \epsilon^2.$$

For the tilted policy and $z > 0$, the wrong-action interval is $0 < x < bz$, where regret is $2x$. Inside its margin strip this becomes $\max(0, bz - \epsilon) < x < bz$. By symmetry, the negative-z half contributes identically. With the joint density $1/4$, for $0 < \epsilon < \min(b, 1-b)$ we obtain

$$\begin{aligned}\mathcal{B}_\epsilon(R_{\text{tilted}}) &= \int_0^{\epsilon/b} \frac{b^2 z^2}{2}\,dz + \int_{\epsilon/b}^{1}\left(bz\epsilon - \frac{\epsilon^2}{2}\right)dz \\ &= \frac{b}{2}\epsilon - \frac{1}{2}\epsilon^2 + \frac{\epsilon^3}{6b}.\end{aligned}$$

This proves Equation (20) with an explicit remainder. Because dispatch is matched throughout, the different tube costs arise from the routing decisions themselves.

C.3 Fisher expansion and nonlinear downstream maps

The boundary calculation is agnostic about the predictive model. To connect it to measurable loss, we next expand the downstream risk itself and separate Fisher curvature from the extra term created by a nonlinear interface map.

Proof of Theorem 5. For a regular exponential family, $\log P_\eta(y) = \eta^\top T(y) - A(\eta) + b(y)$. Assuming finite conditional expectations, risk is $C(\eta) = A(\eta) - \eta^\top \bar{T}_\star - \mathbb{E}_\star b(Y)$. With $\bar{T}(\eta) = \nabla A(\eta)$ and $F_i = F(\eta_i)$, its first two derivatives are $\bar{T}(\eta) - \bar{T}_\star$ and $F(\eta)$. Taylor's theorem with integral remainder yields

$$C(\eta_i + \delta) - C(\eta_i) = (\bar{T}_i - \bar{T}_\star)^\top \delta + \tfrac{1}{2}\delta^\top F_i \delta + \tfrac{1}{2}\int_0^1 (1-t)^2 \nabla^3 A(\eta_i + t\delta)[\delta,\delta,\delta]dt.$$

If the third derivative has operator norm at most K, the remainder magnitude is at most $K\|\delta\|^3/6$. If $P_\star = P_i$, the linear term is zero and subtracting the two cross-entropies gives $\mathrm{KL}(P_i\|P_j)$ exactly. For arbitrary $P_\star$, it instead gives $\mathrm{KL}(P_\star\|P_j) - \mathrm{KL}(P_\star\|P_i)$ whenever these divergences are finite. □

For categorical logits, $F = \mathrm{diag}(p) - pp^\top$ and $\frac{1}{2}\delta^\top F\delta = \frac{1}{2}\{\sum_v p_v\delta_v^2 - (\sum_v p_v\delta_v)^2\}$. The expansion remains valid despite the redundant all-ones logit direction because it never inverts the positive semidefinite Fisher matrix. Write $\eta = \eta(\Phi)$ with Jacobian J_η. The chain rule gives

$$\nabla_\Phi C = J_\eta^\top(\bar{T} - \bar{T}_\star), \qquad \nabla_\Phi^2 C = J_\eta^\top F J_\eta + \sum_v (\bar{T}_v - \bar{T}_{\star,v})\nabla_\Phi^2 \eta_v.$$

Thus a Fisher pullback alone is the loss Hessian only if the map is affine or the second term vanishes. On a C^3 interface segment with bounded third derivative, expansion in $d = \Phi_j - \Phi_i$ is $\nabla_\Phi C^\top d + \frac{1}{2}d^\top \nabla_\Phi^2 C d + O(\|d\|^3)$. Thus predictive-logit interpolation and expert-output interpolation use different second-order terms whenever the downstream map contributes curvature.

D Stability and Correction Proofs

D.1 Weighted local time

The preceding section describes where a router can change. This section describes how those changes accumulate along a trajectory, beginning with the occupation measure of a trajectory near a routing boundary.

Proof of Theorem 6. Write $W_t = W(H_t)$. For a continuous semimartingale X, the occupation formula on an interval $[a, b]$ is

$$\int_a^b f(X_t)d\langle X\rangle_t = \int_{\mathbb{R}} f(x)\{L_b^x - L_a^x\}dx.$$

It also holds multiplied by any bounded random variable constant in time on this interval, on the common pathwise event where the formula holds. Approximate each continuous path W_t uniformly on $[0, T_{\text{end}}]$ by its left-endpoint step values on successively finer deterministic partitions. For step weights the desired limit follows from continuity of finitely many $L_b^x - L_a^x$ in x. The approximation error on the normalized occupation integral is bounded by

$$\|W - W^{(n)}\|_\infty \frac{1}{2\epsilon}\int_{-\epsilon}^{\epsilon} L_{T_{\text{end}}}^x dx \le \|W - W^{(n)}\|_\infty \sup_{|x|\le\epsilon_0} L_{T_{\text{end}}}^x.$$

The corresponding zero-level Stieltjes integral has the same bound with $L_{T_{\text{end}}}^0$. First taking the partition fine and then ϵ small proves pathwise convergence. A deterministic bound on $|W_t|$ and the assumed integrable local-time envelope permit dominated convergence in expectation. □

If $dH_t = b_t dt + \Sigma_t dB_t$ and u is C^2, Itô's formula makes the martingale part of $u(H_t)$ equal to $\nabla u(H_t)^\top \Sigma_t dB_t$, giving the quadratic-variation factor in the main text. This factor explains why replacing quadratic-variation time by ordinary clock time requires either fixing it appropriately or including it in the weight. The theorem neither identifies a diffusion law from discrete hidden states nor proves that occupation by itself causes a prediction error.

D.2 Continuation loss and candidate restriction

A local gate change is useful only through the continuation it induces. We therefore formulate the pathwise loss identity before discussing what is lost when the candidate action family is restricted. Let the stage index order all interventions in a finite computation. Include residual streams, attention caches, position, and any relevant prefix in the state so that the transition kernel is Markov. Let c_ℓ be stage loss and $Q_\ell^{\pi_0}(h, a) = \mathbb{E}[c_\ell + V_{\ell+1}^{\pi_0}(H_{\ell+1}) \mid h, a]$, $V_\ell^{\pi_0} = \mathbb{E}_{\pi_0} Q_\ell^{\pi_0}$, and $A_\ell^{\pi_0} = Q_\ell^{\pi_0} - V_\ell^{\pi_0}$. Assume the relevant expectations are integrable (bounded losses suffice). Taking expectation under π and summing the advantages cancels consecutive V terms by the tower property. Only the initial and terminal values remain: the terminal value is zero and the initial reference value is $J_{\text{seq}}(\pi_0)$, giving Equation (14).

If π minimizes an estimate $\widehat{A}_\ell$ on a subset $\mathcal{S}_\ell(h)$ with uniform error at most ε_ℓ on states visited by π, then

$$A_\ell^{\pi_0}(h, a_\pi) \le \min_{a\in\mathcal{S}_\ell(h)} A_\ell^{\pi_0}(h, a) + 2\varepsilon_\ell = \min_{a\in\mathcal{A}_\ell(h)} A_\ell^{\pi_0}(h, a) + b_\ell(h) + 2\varepsilon_\ell.$$

Here $b_\ell(h)$ is the difference between the candidate-restricted and unrestricted minima of $A_\ell^{\pi_0}(h, \cdot)$. The global minimum is nonpositive because π_0 averages this advantage to zero. Substitute into Equation (14) to obtain the stated bound. The required error bound therefore concerns states visited by the changed policy; a bound under the native distribution alone is insufficient.

For nested candidate families whose policy feasible sets are also nested, minimization of the same exact objective cannot worsen as support grows. For raw hard-action costs, the truncation gap is $\min_{a\in\mathcal{S}} D_a - \min_{a\in\mathcal{A}} D_a \ge 0$. This monotonicity need not hold after fitting different heads or renormalizing different priors: these are changed objectives or changed approximations.

D.3 Surrogate ranking and group calibration

We next return to a fixed candidate comparison. A uniform cost-error bound controls its ranking error; an independent calibration sample then controls the probability of accepting a harmful proposal.

Proof of Proposition 7. Suppose $|(\widehat{D}_a - \widehat{D}_{a_0}) - (D_a^\lambda - D_{a_0}^\lambda)| \le \varepsilon$ simultaneously for all candidates. Let $\widehat{a}$ minimize the estimated cost and a^M the true candidate cost. Then

$$D_{\widehat{a}}^\lambda - D_{a_0}^\lambda \le \widehat{D}_{\widehat{a}} - \widehat{D}_{a_0} + \varepsilon \le \widehat{D}_{a^M} - \widehat{D}_{a_0} + \varepsilon \le D_{a^M}^\lambda - D_{a_0}^\lambda + 2\varepsilon.$$

Also $G = D^{\lambda}_{a_0} - D^{\lambda}_{\widehat{a}} \geq \widehat{G} - \varepsilon$. This proves both deterministic conclusions. A full-action regret bound adds the nonnegative candidate-truncation gap.

For calibration, condition on every fitted object: head, price, candidate generation, hyperparameters, and proposal rule. These objects must not depend on calibration outcomes. For each exchangeable group j, let $R_j = \max_{h \in j}(\widehat{G}(h) - G(h))$ over the fixed proposals. The n calibration residuals and a new residual are exchangeable. A randomized tie-breaking rank is uniform on $1, \ldots, n+1$; using the non-random upper order statistic is conservative. For $r = \lceil (n+1)(1-\delta) \rceil$ and $u_\delta = R_{(r)}$ (or $+\infty$ if $r = n+1$), $\Pr\{R_{\text{new}} > u_\delta\} \leq \delta$. On the complementary event every accepted proposal in the new group obeys $G \geq \widehat{G} - u_\delta > 0$. The joint event of any accepted harmful proposal is therefore contained in the rank-failure event. Reverting any subset cannot enlarge this event. The statement is marginal over the new group and calibration data; division by an arbitrarily small acceptance probability is not justified. □

Calibration permits reverting proposals, but load feasibility still needs a separate check. Algorithm 1 scans all relevant thresholds because reverting more moves can initially increase load deviation when those moves previously canceled each other. At the infinite threshold every route is native, so deviation relative to native dispatch is zero and the finite scan terminates. With a fixed external load target, termination requires native feasibility at the allowed tolerance. Empirical total variation does not impose individual hard capacities.

Which gain is calibrated? The deterministic statement concerns conditional risk $G(h) = C_{a_0}(h) - C_{a_1}(h) + \lambda^\top(g(a_0) - g(a_1))$. If calibration instead uses one observed target per state, it observes

$$G^{\text{obs}}(h, y) = \ell_{a_0}(h, y) - \ell_{a_1}(h, y) + \lambda^\top(g(a_0) - g(a_1)).$$

Conditional on $H = h$, its mean is $G(h)$, but the two random variables are not equal. Replacing G by G^{obs} throughout the rank proof gives exactly a realized-loss guarantee under exchangeability of the labeled groups. It does not turn a single-label observation into conditional-risk coverage. The CIFAR-10 calibration uses this observed-target interpretation; the exact-cost controls can calibrate conditional risk directly. Neither version assumes that the native router minimizes the transport objective.

Proposition 2 (Calibration from repeated noisy risk measurements)**.** *Fix the head, prices, and proposal rule. Suppose each of n exchangeable calibration groups contains at most s states and each state's conditional gain is estimated by r independent target draws. Assume every observed loss lies in $[0, L_{\max}]$. Let $\bar{G}$ be the sample-mean gain and set*

$$\bar{R}_j = \max_{h \in j}(\widehat{G}(h) - \bar{G}(h)), \qquad b_r = L_{\max}\sqrt{\frac{2\log(2ns/\eta)}{r}}, \quad 0 < \eta < 1.$$

At rank $m = \lceil (n+1)(1-\delta) \rceil$, use threshold $\bar{R}_{(m)} + b_r$, with $+\infty$ when $m > n$. The probability that any accepted proposal in a new exchangeable group has negative conditional gain is at most $\delta + \eta$.

Proof. The paired observed loss difference has range length at most $2L_{\max}$; the fixed price term changes no sampling error. Conditional Hoeffding bounds and a union bound over at most ns calibration states give $\Pr\{\max|\bar{G} - G| > b_r\} \leq \eta$. On the complementary event, $R_j \leq \bar{R}_j + b_r$ for every group, hence $R_{(m)} \leq \bar{R}_{(m)} + b_r$. The true group residuals $R_1, \ldots, R_n$ and R_{new} are exchangeable, so the rank argument above bounds $\Pr\{R_{\text{new}} > R_{(m)}\}$ by δ. The harmful-acceptance event is contained in the union of this event and the calibration-estimation failure. Adding their probabilities proves the result. Independence between the two events is not needed. □

The repeated observations and bounded-loss assumption determine the scope of this certificate. Since ordinary log loss is unbounded, an application must specify both its sampling budget and either a bounded-loss target or an appropriate tail bound.

For an estimated probabilistic optimum $\widehat{q}$ feasible at the same load as a comparator q^o, if $|\widehat{C}_a - C_a| \leq \varepsilon$ uniformly, optimality of its regularized objective gives

$$L(\widehat{q}) - L(q^o) \leq 2\varepsilon + \tau\mathbb{E}\text{KL}(q^o\|\nu) - \tau\mathbb{E}\text{KL}(\widehat{q}\|\nu) \leq 2\varepsilon + \tau\mathbb{E}\text{KL}(q^o\|\nu).$$

System	Independent units	Quantity isolated
Exact transport	5×100 problems	Known costs and optima; tube and occupation laws
Learned expert banks	2×50 banks	Estimated cost under uniform or skewed capacity
Compact causal MoE	3×30 models	Context-dependent risk; complete sequence repair
Four-layer image MoE	5 models	Same-weight continuity and predictive jump on CIFAR-10
Vision-language MoE	209 image clusters	Complete top-2 action costs at image and answer positions

Table 4: Levels of validation. Each level adds learned computation or natural data while testing a specific theoretical object. Comparisons keep each checkpoint's expert weights fixed.

An approximate solve adds its objective suboptimality. An imperfectly matched load contributes exactly the price charge in Equation (6), bounded in magnitude by $\|\lambda\|_\infty \|\rho(\widehat{q}) - \rho(q^o)\|_1$. For hard maximization with nonconstant prior, the additional ranking perturbation has magnitude at most $\tau\beta \log(p_{\max}/p_{\min})$ on the declared positive support. A raw-risk guarantee must therefore account for optimization error, load mismatch, and any prior-induced ranking change.

D.4 Price estimation and refresh

For $\Psi(\lambda) = \tau \mathbb{E} \log Z_\lambda + \lambda^\top \rho$, the gradient and Hessian were calculated in the existence proof. If $\|g(a)\| \le M_g$, then $\|\nabla^2 \Psi\| \le M_g^2/\tau$. A deterministic gradient step of size at most τ/M_g^2 therefore gives the usual smooth convex descent bound; strict positivity gives uniqueness on V, not a uniform strong-convexity constant over all distributions. Empirical prices replace expectations by development averages. Residual load, stopping tolerance, iteration count, and the final target are saved along with each fitted policy. Repeatedly updating prices after calibration changes the proposal and invalidates that calibration contract. State-dependent candidate supports require feasibility of the restricted moment problem and dual attainment; native inclusion alone ensures only that the native empirical load remains attainable.

E Experimental Protocols

The experiments progressively replace known costs with learned computation and natural data. Their statistical units follow the question being tested: independent transport problems test analytic identities, independent learned checkpoints test mechanism estimates, and held-out images or sequences measure behavior. Controlled-model uncertainty resamples complete problem seeds or checkpoints. Route-gradient intervals instead resample paired update groups and validation documents conditional on each checkpoint.

E.1 Exact transport and learned nonlinear interfaces

We begin with a model whose action costs are known. This lets us test the transport and boundary identities before introducing cost-estimation error. Draw $v \sim \mathcal{N}(0, \mathrm{diag}(\sigma^2))$ with RMS-normalized coordinate scales; set $z_1 = v_1 + 0.25 \tanh(v_r)$ and $z_j = v_j$ for $j > 1$. Use $h = Az + 0.03 \sin(3Az)$ and $\psi(z) = z + \gamma \sin(Bz) + 0.15\gamma \sin(2Bz)$. Here $A \in \mathbb{R}^{d \times r}$ has orthonormal columns from Gaussian QR, $B_{ij} \sim \mathcal{N}(0, 1/r)$, and expert centers $e_i \in \mathbb{R}^r$ are Gaussian vectors normalized to length 2.2. Expert costs are $C_i(h) = \frac{1}{2}\|\psi(z) - e_i\|^2$. There are 100 independent problems in each of five settings: linear/uniform, nonlinear/uniform, nonlinear/skewed, anisotropic with 16 experts, and a sufficient-label positive control. The first, second, third, and fifth use latent/ambient dimension $4/48$ and eight experts; the anisotropic condition uses $8/96$ and 16 experts. Nonlinearity is 0.65 except in the linear case (zero) and anisotropic case (0.55).

Each problem uses 262,144 training, 65,536 validation, and 131,072 test states, with 2,400 router steps, 1,800 cost-head steps, batch size 16,384, and $\tau = 0.18$. Tube statistics use 1,048,576 fresh states. Stochastic tests use 8,192 simulated paths with 512 time steps. Switching cost sums the current top-two absolute cost gap at steps where the top-ranked action changes. Occupation sums the boundary-band indicator times the time step, with or without this cost weight; it uses clock time. The

reported R^2 is the in-sample fit across paths, using a separate affine regression for each occupation statistic. Higher-order scaling is independently checked using one through four independent Gaussian tie coordinates. The sufficient-label rule preserves the hard minimizing action.

We then replace the fixed interfaces with learned expert banks. The nonlinear-bank extension trains eight two-layer SiLU experts of width 72 against a nonlinear teacher, then trains a linear router and a width-128 cost head without using the teacher assignment as an input. Uniform and skewed capacity conditions each have 50 seeds, with 131,072/32,768/65,536 train/validation/test states. Experts and routers train for 2,000 steps, the head for 1,400 steps, batch size 8,192, and $\tau = 0.15$. Exact-transport and learned-bank corrections reuse development data for price fitting, proposal selection, and empirical gain-residual thresholds. They test empirical repair, not the independent split-conformal coverage statement of Proposition 7.

E.2 Compact causal sequence model

The next experiment introduces a causal continuation, so a routing change can affect later predictions. We retain a known data generator to distinguish local conditional costs from the measured loss of a complete sequence. The *compact causal MoE* has 157,320 trainable parameters: width 64, four attention heads in one causal attention block, eight width-128 expert networks, top-1 dispatch, 64 vocabulary symbols, and length-48 contexts. Token embedding and output weights are tied. A straight-through Gumbel-softmax gate trains experts and router jointly; evaluation uses hard argmax.

The data generator is a hidden-regime Markov process. Four regimes add increments $(1, 5, 11, 17)$ modulo 64 to the current token, with probability 0.12 of a uniform random next token. Balanced and skewed regimes have persistence 0.94; fast switching uses 0.72. Regime probabilities are uniform except for the skewed condition $(0.50, 0.25, 0.15, 0.10)$. Every token appears in every regime. Training uses 3,000 steps with batch 128 and load weight 0.03. Each of 30 seeds per condition has 2,048/1,024/2,048 independently generated sequences for head training, development, and test; the cost head has width 256 and trains for 2,500 steps.

The exact conditional costs in this generator are conditioned on the *generating regime as well as the observed prefix*. The router and surrogate see only model hidden states, never that regime. These costs are not the analytically marginalized Bayes risk conditional only on the observed prefix. Realized next-token negative log-likelihood is also measured by complete-sequence forwards and is the behavioral endpoint. Candidates enumerate all eight experts at the same state; identity forwards reproduce the native loss to numerical precision. The reported native percentile is the fraction of candidate costs no lower than native, including native itself. With eight distinct costs, random selection has mean percentile $9/16$ under this inclusive ranking convention.

The load-calibrated fault adds $\eta(\sigma_s/\sigma_W)Wh$ to gate scores at severities $\eta = 0.15, 0.30, 0.50, 0.75$, then fits compensating expert prices on development data. One Gaussian matrix W is drawn per seed; σ_s, σ_W are development standard deviations of native and perturbation scores. The native model weights are unchanged. Margin repair and COT use the same development split and held-out evaluation; repair strength is selected by development cost under a load tolerance. The endpoint is recovery relative to the faulted model, not improvement of the unperturbed model. Full-sequence replay fixes the observed token prefix (teacher forcing); it is neither free-running generation nor an average of local costs relabeled as a forward run.

A separate differential audit checks the local expansion underlying this comparison. It compares native and runner-up predictive logits. For each candidate pair it interpolates logits by $t \in \{0.05, 0.10, 0.20, 0.50, 1\}$ and calculates the exact conditional cross-entropy difference, linear gradient term, and Fisher quadratic. Remainder exponents fit the first three scales, fixed before inspection. Because the interpolation is performed in predictive logits, the measured remainder does not establish second-order accuracy for a unit-length intervention in nonlinear hidden-state coordinates.

E.3 Four-layer image boundary experiment

We adapt the CIFAR-10 boundary-normal design of Huu et al. (2026), using a four-layer sparse residual network, 32 experts and top-4 routing per layer. A 3072-to-128 image stem precedes width-128 blocks with width-64 GELU experts and a ten-class head. We specify this compact architecture

independently because the reference paper does not identify its CIFAR backbone. The dataset is the official 50,000 training/10,000 test split (Krizhevsky, 2009). A seeded permutation divides training images into 40,000 model-training, 5,000 cost-fitting/price-fitting, and 5,000 calibration images. Five checkpoint seeds are fixed at 9100–9104.

The model trains for 12 epochs with AdamW, learning rate $8\,10^{-4}$, weight decay 0.01, batch 256, gradient clipping 1, and auxiliary load weight 0.01. Normalization uses mean $(0.4914, 0.4822, 0.4465)$ and standard deviation $(0.2470, 0.2435, 0.2616)$. There is no test-based checkpoint selection. The third block is audited. Its top-eight experts define five complete actions: native top-four, or replacing its fourth expert by ranks five through eight. Each route applies softmax to its selected experts' original gate logits.

The width-128 cost head receives hidden state, all router log probabilities, and the candidate dispatch vector. It learns relative loss by squared error for 1,200 AdamW steps, batch 512, learning rate 0.001. Price fitting uses temperature 0.2 and 300 updates with step $2/\sqrt{1 + t/50}$, after which head and prices are frozen. Residual quantiles use the 5,000 independent calibration images at level 0.9, with the finite-sample rank correction; load filtering allows normalized dispatch TV 0.01 relative to native test dispatch. No test label enters deployed candidate selection. Test candidate losses are computed separately for audit and checked against fresh full-model forwards.

The affinity-OT control uses negative mean selected router score instead of predicted risk, with the same action family, price updates, and load filter. It isolates cost choice, not the complete training algorithm of Nguyen et al. (2025). The SmoothSMoE boundary rule follows Huu et al. (2026) with smoothstep exponents $a = b = 1$: multiply each expert's unnormalized softmax weight by $\mathrm{clip}((s_i - s_{(k)} + \epsilon)/\epsilon, 0, 1)$. Its ϵ is the development 20th-percentile margin. It may activate additional experts; the measured active count is reported; we do not describe it as equal-compute hard routing.

For boundary-normal traces, project each of the first 128 held-out normalized interface states to its affine fourth/fifth-expert tie hyperplane and evaluate symmetric distances from 10^{-5} to 0.2. We retain every prespecified state, instead of selecting the visually largest discontinuity. Projection can encounter a different active tie; the statistic is the average over this declared set. Both policies share model weights. Outputs, candidates, per-image losses, dispatch, runtimes, and checkpoint metadata are retained.

F Controlled Results and Additional Checks

F.1 Exact transport, tube, and local-time controls

We first check the identities under known costs, using the exact-transport protocol above. Table 5 and Figure 4 report all settings before we turn to learned experts and sequence-level behavior.

Setting	Router regret	Coarse-label regret	Thresholded COT gain	Weighted R^2
Linear	0.027	0.870	0.009	0.866
Nonlinear	0.072	0.951	0.032	0.861
Skewed	0.107	0.884	0.162	0.864
Anisotropic	0.070	2.432	0.048	0.879
Sufficient-label control	0.072	–	0.032	0.861

Table 5: Ground-truth benchmark over 100 seeds per row. Tabulated regret is excess regularized transport objective. Gain is held-out cost reduction relative to the trained router. Weighted R^2 predicts stochastic switching cost from compatibility-weighted occupation. The constructed hard-label policy has exactly zero unregularized decision regret in every sufficient-label seed.

Unweighted occupation has $R^2 = 0.369, 0.333, 0.475, 0.437$, and 0.333 in the same row order, compared with 0.861–0.879 after compatibility weighting. The validation-oracle margin rule is slightly harmful in every setting, with mean gains from -0.00023 to -0.00043. Thus COT's advantage is not explained by tuning a more favorable low-margin threshold.

The first-order margin-tube exponent lies between 0.993 and 1.001 across the five scenarios. Pooling seed-level estimates by tie order gives means 1.00, 1.99, 3.02, and 3.94. At fourth order, shrinking the fitted epsilon window moves the estimate toward four while increasing sampling variance.

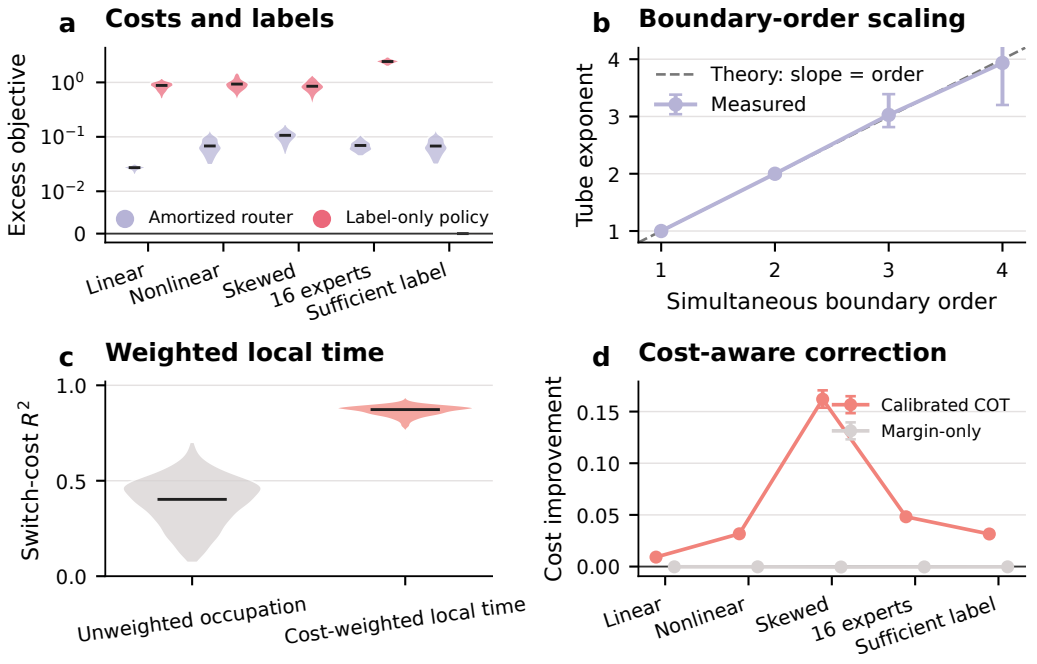

Figure 4: Complete ground-truth compatibility-transport benchmark. **(a)** Router and label-only excess objectives. **(b)** Tube exponents for one through four transverse ties. **(c)** Prediction of diffusion switching cost. **(d)** Thresholded COT versus the validation-oracle margin rule.

F.2 Learned experts and end-to-end routing

We next ask whether estimated compatibility costs remain useful with learned experts. Table 6 reports expert reconstruction by mean squared error (MSE), held-out cost-head fit, and test-cost reduction under the native expert budget. Table 7 then tests recovery through a complete causal continuation; Figure 5 gives the full sweeps.

Expert load	Expert MSE	Cost-head R^2	Thresholded COT gain	Load TV
Uniform	0.00341	0.962	0.165	0.0084
Skewed	0.00342	0.962	0.240	0.0093

Table 6: Learned nonlinear expert banks, 50 seeds per row. Gain is test-cost reduction relative to the native learned router.

Compatibility action agreement compares the learned router argmax with the row-wise argmax of the test-cost entropic OT solution at the target load and temperature 0.15; it is 0.832 under uniform load and 0.809 under skewed load. Residual candidate-oracle headroom is 0.384 and 0.457 after correction, measured against the per-state minimum without a load constraint. Margin-only selection is slightly harmful in both settings. Thus the learned surrogate reduces cost while leaving substantial room for a better candidate selector.

Process	Native percentile	Oracle gap	COT recovery	Oracle recovery
Balanced	0.984	0.062	0.067	0.092
Fast switching	0.977	0.089	0.063	0.085
Skewed	0.986	0.053	0.065	0.088

Table 7: Nonlinear sequence control and full-sequence recovery at router-fault severity 0.75. Costs and recoveries are mean nats per predicted token over 30 independently trained checkpoints per process. Oracle gap compares native cost with the per-state minimum without a load constraint.

The COT policy changes load by total variation 0.00260, 0.00232, and 0.00241 for the three rows. Margin repair changes load by a similar amount but recovers −0.00008, −0.00003, and +0.00003 nats, without a consistent positive recovery. This isolates compatibility ranking from extra route churn and capacity drift.

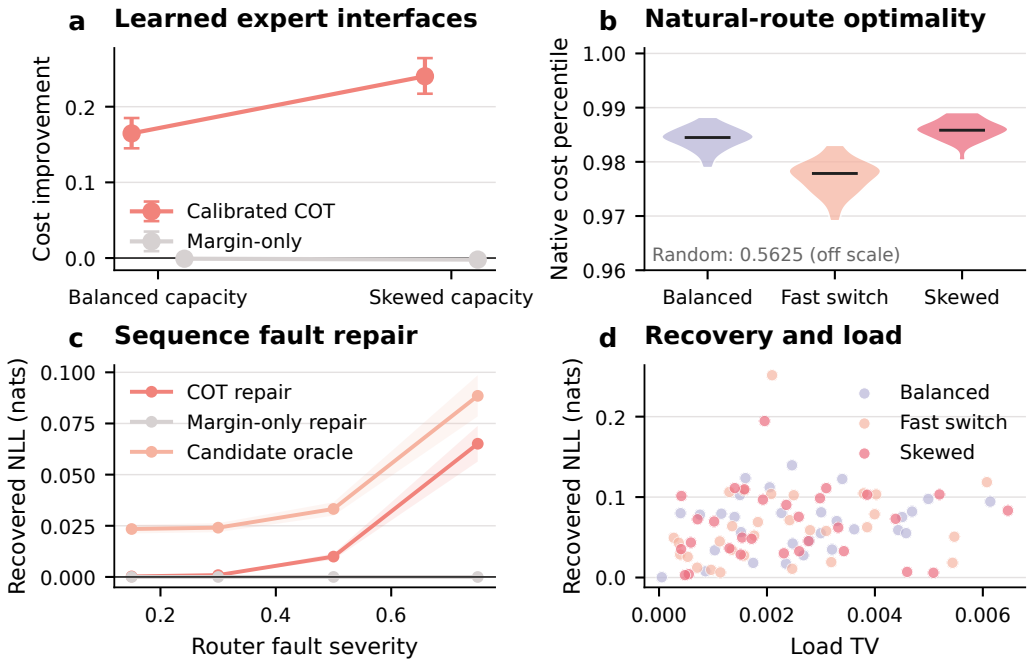

Figure 5: Complete learned-interface and causal-MoE results. **(a)** Thresholded COT on learned nonlinear expert banks. **(b)** Native expert compatibility percentile. **(c)** Full-sequence recovery across router-fault severities. **(d)** Recovery versus expert-load deviation at severity 0.75.

F.3 Fisher expansion across displacement scales

The recovery results compare complete interventions. To examine the local loss expansion separately, we now vary the displacement scale. Table 8 compares the linear and quadratic fits, and Figure 6 shows their remainders across the full sweep.

Process	Scale	Linear R^2	Linear+Fisher R^2	Remainder orders
Balanced	0.05	0.986	0.99991	2.137/3.066
Balanced	0.20	0.693	0.976	2.137/3.066
Fast switching	0.05	0.991	0.99997	2.095/3.024
Fast switching	0.20	0.817	0.992	2.095/3.024
Skewed	0.05	0.986	0.99991	2.133/3.061
Skewed	0.20	0.705	0.978	2.133/3.061

Table 8: Differential fit over 30 checkpoints per process. The final column gives fitted orders after the linear and quadratic approximations.

At full displacement, signed Spearman correlation is 0.451, 0.712, and 0.462 for the first-order term and 0.843, 0.883, and 0.860 after adding Fisher curvature. Fisher magnitude alone correlates 0.763, 0.638, and 0.785 with absolute candidate cost. These are out-of-sample token-state statistics within each independently trained checkpoint, followed by checkpoint-level inference.

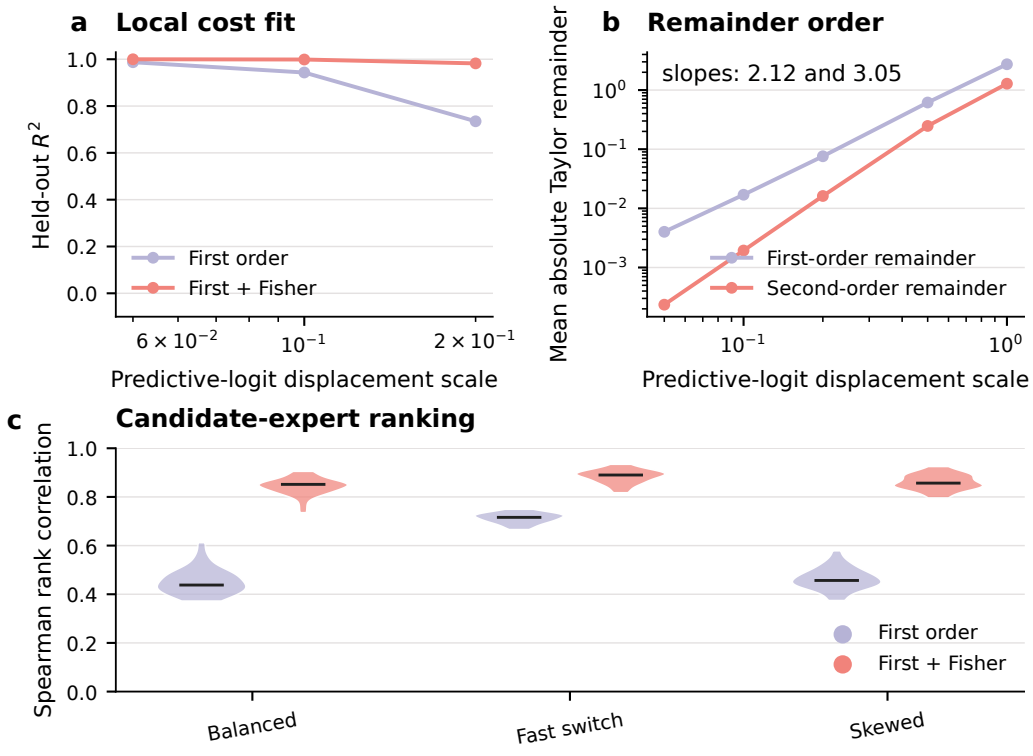

Figure 6: Complete differential audit. **(a)** First- and second-order held-out cost fit. **(b)** Taylor remainders over the interpolation range. **(c)** Candidate rank correlation at complete displacement.

F.4 Natural-image boundary and policy controls

The image experiment separates output continuity from predictive harm. We use a boundary-normal intervention for the first question and a candidate audit for the second; Table 9 reports the associated policy controls. The former approaches a declared tie in normalized gate-input coordinates and tests continuity of the expert-mixture output. The latter keeps the unperturbed test state and measures each complete action through the remaining network. Candidate loss is the observed-label negative log-likelihood, not an exactly known conditional risk. Likewise, the image calibration residual concerns a realized paired loss difference. The Fisher quantity measures predictive displacement; its correlation with absolute loss change does not supply a signed improvement guarantee.

Policy	NLL (nats)	Accuracy (%)	Load TV	Active experts
Native top-4	2.22269	50.590	0	4
Unfiltered cost head	2.22583	50.596	0.014995	4
Calibrated COT	2.22256	50.582	0.002390	4
Affinity transport control	2.22286	50.588	0.009995	4
Candidate oracle	2.04518	53.436	0.010890	4
Boundary smoothstep	2.21934	50.648	—	4.228

Table 9: Complete same-checkpoint policy controls, averaged over five seeds. Active experts refer to the corrected layer; the other three layers retain top-4. The oracle uses test targets and is a candidate reference, not a deployable policy. Smoothstep admits additional experts and is not an equal-compute hard-routing comparison.

The paired native-minus-COT NLL change is 0.00013 nats. The small difference does not establish a reliable classification improvement. Here the main evidence concerns predictive jump sensitivity; the controlled sequence experiment provides the intervention evidence for loss repair. COT changes the complete action for 4.28% of all test images, compared with 79.62% for the unfiltered head, and its load TV is 0.00239. Candidate-oracle headroom is 0.1775 nats, evaluated without a matched-load restriction. The complete per-seed metrics and candidate arrays are included in the artifact; images from one checkpoint are not treated as independent training replications.

Execution and accounting. Hard policies execute exactly the four chosen expert networks. The smoothstep reference uses a dense implementation of the published weight rule; its nonzero expert count describes the mathematical support, not measured sparse kernel speed. Online timing replays both policies from image input, recomputes the COT head and proposals without test labels or cached costs, and includes the load filter. One logical batch contains the 10,000 test images, with expert-execution chunks of 256. This measures batched throughput, not single-image serving latency. The first timing pass warms up the kernels; five synchronized passes per checkpoint are retained. The original `rerun_seconds` field times execution of already chosen actions only; the separate `online_rerun` field includes selection. On a single NVIDIA L20, mean checkpoint-median runtime is 36.37 ms for native execution and 38.61 ms for COT, a 6.16% increase under this batched protocol. These measurements retain the original expert count; they quantify the additional prediction and filtering work; equal dispatch does not imply identical runtime.

F.5 A 5.61B-PARAMETER VISION-LANGUAGE CHECKPOINT

We finally test the same diagnostic in a public pretrained vision-language mixture. Here visual and linguistic representations pass through the same sparse decoder, allowing us to ask whether predictive displacement still distinguishes consequential routing changes at both types of position. Table 10 summarizes the complete-action audit, and Figure 7 compares displacement with margin. All comparisons use frozen weights; the experiment evaluates the mechanism; it does not evaluate a new training procedure.

Checkpoint and architecture. We use `LanguageBind/MoE-LLaVA-Phi2-2.7B-4e` from Lin et al. (2026), pinned to revision `114cedb30ff903600555dce5ed7a22b1012b63c7`. The loaded model has exactly 5,609,680,896 parameters, including the vision encoder and image projector. The decoder has 32 layers, hidden width 2,560, feedforward width 10,240, and vocabulary size 51,200. Sixteen alternating layers, indexed $0, 2, \ldots, 30$, have four experts each. The audited policy selects two experts and normalizes their original gate logits. The visual front end is CLIP ViT-L/14 at 336×336 pixels; its penultimate hidden layer supplies 576 patch tokens after removing the class token. A two-layer GELU projector maps width 1,024 visual features to the decoder width.

This gives a public checkpoint in the same approximate parameter class and four-expert/top-2 regime as the visual-language experiment described by Huu et al. (2026). We adopt that scale and the same-weight intervention principle; we do not claim to recover an unspecified checkpoint or reproduce its multimodal fine-tuning data. All parameters of the downloaded checkpoint remain frozen in every reported comparison.

Implementation correspondence. The implementation uses the Hugging Face Phi and CLIP computations for attention, positional embeddings, normalization, and visual encoding. Checkpoint names for the DeepSpeed expert blocks are remapped to explicit expert modules. All 1,055 state-dictionary keys are covered, with shape checking during loading and no unmatched checkpoint or model parameters. The language prediction head has no bias, matching the released architecture; the image projector uses the released spatial-projection weights. Nonpersistent rotary and position-index buffers are initialized from the checkpoint configuration and are not treated as trainable weights.

For repeatable same-state action comparisons, the execution rule is *deterministic top-2 without token dropping*. This disables the stochastic second-expert selection used by the original DeepSpeed path. The comparison therefore concerns this deterministic policy at the released weights; benchmark equivalence to the original stochastic implementation is not assumed. Gate scores are computed in float32; expert weights and activations use bfloat16; predictive logits and losses are evaluated in float32. Selected expert outputs are combined with the softmax restricted to that expert pair. The attention implementation is scaled dot-product attention in PyTorch 2.6.0 with Transformers 4.53.3.

Data, prompts, and the statistical unit. We use the AI2D diagram question-answering test split (Kembhavi et al., 2016), distributed as `lmms-lab/ai2d`. A permutation generated with NumPy seed 202709 selects 256 questions without using model outputs or candidate costs. They contain 209 distinct image-byte hashes. Questions sharing an image are assigned the same image cluster, so multiple questions about the same diagram are not counted as independent images. All selected question indices and image hashes are retained with the raw candidate groups.

Images are converted to RGB, padded to a square with the CLIP mean color, and processed by the checkpoint's CLIP image processor. The conversational prefix follows the released Phi conversation format. The image tokens are followed by the question, lettered answer choices, and a request for the option letter only. The prediction target is the first answer-letter token with a leading space. Every answer option is checked to tokenize to one token; all prompts remain within the 2,048-token context limit. The loss is negative log-probability under the *full vocabulary*, not a softmax renormalized over answer options. Option-restricted accuracy is stored as an execution check but is not the response variable of the compatibility analysis.

Same-state complete-action sweep. We preselect routed decoder layers 8, 16, and 24 (zero-based), and inspect two positions per question: the middle visual patch and the last prompt token predicting the answer. Each question therefore contributes six candidate groups, for 1,536 groups overall. With four experts and top-2 there are exactly $\binom{4}{2} = 6$ unordered complete actions. Every action is evaluated, including the native pair, giving 9,216 forced forward passes. Softmax is applied to the original logits restricted to each candidate pair; no donor state, donor weight, or external route pool is used.

For one candidate evaluation, only the pair at the designated layer and position is forced. All other router decisions remain endogenous to their current hidden states. Each forward pass runs from the image and question to the actual answer logits, so downstream changes are included. The measured loss therefore includes the complete continuation after the forced joint action. Forcing the original pair must reproduce the unforced answer loss within 10^{-4} nats. All 1,536 groups pass, with maximum observed discrepancy exactly zero in the recorded float32 losses.

Predictive displacement and summary statistics. Let z_0 and z_a be full-vocabulary answer logits under the native and forced complete action, $p_0 = \mathrm{softmax}(z_0)$, $p_a = \mathrm{softmax}(z_a)$, and $\delta_a = z_a - z_0$. We compute

$$F_a = \tfrac{1}{2}\left\{\sum_v p_{0v}\delta_{av}^2 - \left(\sum_v p_{0v}\delta_{av}\right)^2\right\}, \qquad K_a = \tfrac{1}{2}\{\mathrm{KL}(p_0\|p_a) + \mathrm{KL}(p_a\|p_0)\}. \tag{21}$$

Both diagnostics require evaluating candidate outputs, whose cost must be counted in deployment. They require no target labels; their association with loss changes is evaluated using the observed answers. Their purpose is to test the connection between interface displacement and loss magnitude from Theorem 5 in an actual multimodal continuation.

The five non-native actions enter the correlation analysis. Identity actions are excluded from these correlations by design, since their zero cost and zero displacement would create an artificial concen-

tration at the origin; they remain in the identity checks and raw data. Define the observed answer loss $\ell_a = -\log p_a(y)$ and $\ell_0 = -\log p_0(y)$. For each modality position we correlate F_a, K_a, and the native second/third logit margin with $|\ell_a - \ell_0|$. An additional within-state concordance compares all ten pairs of non-native alternatives in each group. It scores one for an agreement between Fisher and absolute-cost ordering, zero for disagreement, and one half for a tie in either quantity. This retains all groups and controls the image, question, position, and layer within each comparison. Margin is constant across alternatives in a group and thus has concordance exactly 0.5 under this definition.

Random-alternative gap is the mean signed difference over the five non-native actions. Candidate headroom is $\max(0, -\min_{a \neq a_0}(\ell_a - \ell_0))$. Natural percentile is the fraction of alternatives with larger loss, counting equal losses as one half. These definitions separate a useful harm diagnostic from a claim that the audited policy minimizes all local candidate costs. The cost-response theory compares well-specified constrained optimizers; it does not assume that every pretrained router already satisfies that optimality condition.

Measure	Image position	Answer position
Fisher correlation	0.915	0.940
Margin correlation	0.051	−0.081
Predictive KL correlation	0.916	0.941
Within-state concordance	0.824	0.844
Absolute cost change (nats)	0.026	0.025
Random alternative gap (nats)	−0.002	−0.002
Candidate headroom (nats)	0.031	0.029
Natural percentile	0.478	0.449

Table 10: Frozen 5.61B vision-language candidate audit. Entries are estimates over 209 image clusters, retaining every question, layer, and candidate. Correlation and concordance measure loss magnitude; the signed random gap and oracle headroom answer different questions.

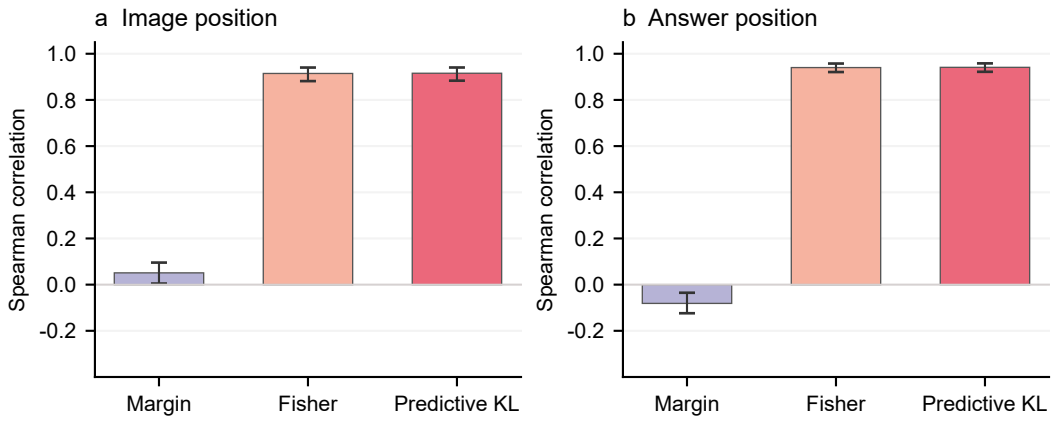

Figure 7: Predictive displacement versus routing margin in a frozen vision-language MoE. Bars show Spearman correlation with absolute answer-loss change across all non-native complete actions, separately at image and answer positions, using 209 distinct diagrams and 256 questions. The same candidate outputs determine Fisher and predictive KL diagnostics; they are not independent experiments.

Interpretation and execution budget. Fisher and symmetric predictive KL track loss magnitude at both modalities, whereas the router margin carries substantially less ordering information. The paired Fisher-minus-margin correlation contrast is positive for both positions. The within-state comparison tests this distinction without relying on between-layer or between-question differences. These observations support predictive weighting of routing changes at the scale of a pretrained multimodal MoE. They do not turn a nonnegative curvature statistic into a signed improvement criterion: the linear target-dependent term in the risk expansion remains necessary for choosing a beneficial action.

The full 256-question sweep takes 744 seconds after checkpoint loading on one NVIDIA L20, with a peak allocated tensor footprint of 11.31 GB. It uses the same two active experts for every complete candidate; exhaustive candidate supervision is an offline audit expense. No cost head is fitted and no additional fine-tuning is performed; answer targets are used only to evaluate candidate losses. The experiment extends the controlled comparisons to a released vision-language model while keeping the same frozen-action evaluation principle.

G Capacity-Constrained Cost Error and Routing Response

A compatibility predictor can be inaccurate without changing the optimal routing decision. A statewise cost shift leaves every action comparison unchanged, while a linear expert-load shift can be absorbed by refitting prices. Raw squared error counts both effects, so it need not measure decision error. This section isolates the error that matters under a fixed load constraint and proves the bounds used in Sections 3 and 5. We characterize the remaining error exactly, first at finite amplitude and then to second order. The latter calculation supplies the soft-routing side of the boundary limit in Appendix H.

G.1 Standing assumptions and the error that decisions can observe

Let $\mathcal{A}$ be finite, and write $\mathcal{Q}_\rho$ for the probability kernels with $\mathbb{E}\sum_a q_a g(a) = \rho$. In this section g is bounded, the reference kernel ν is strictly positive, and costs are integrable. The same action family, reference, temperature, state distribution, and target load are used in every comparison. Changing any of them would change the optimization problem, not merely perturb its cost estimate.

To remove these two irrelevant components, first subtract a global load term and then measure the remaining range across actions at each state. For a perturbation $e = (e_a)$, this gives the capacity-adjusted oscillation

$$\|e\|_{\rho,\mathrm{osc}} := \inf_{v\in\mathbb{R}^E} \mathbb{E}_\mu \left[\max_a\{e_a(H) - v^\top g(a)\} - \min_a\{e_a(H) - v^\top g(a)\}\right]. \tag{22}$$

The notation ρ records the feasible comparison class. The expression itself depends on $\mu, g, \mathcal{A}$, not on the numerical value of a relative-interior target. It is a seminorm on integrable cost fields modulo state offsets and linear load terms. A minimizer v is not needed for the bounds: an infimum and an approximating sequence suffice.

Lemma 1 (Invariance and subadditivity)**.** *For integrable e, f, scalars t, an integrable state function c, and a constant vector w,*

$$\|e + c + w^\top g\|_{\rho,\mathrm{osc}} = \|e\|_{\rho,\mathrm{osc}}, \tag{23}$$

$$\|te\|_{\rho,\mathrm{osc}} = |t|\|e\|_{\rho,\mathrm{osc}}, \tag{24}$$

$$\|e + f\|_{\rho,\mathrm{osc}} \le \|e\|_{\rho,\mathrm{osc}} + \|f\|_{\rho,\mathrm{osc}}. \tag{25}$$

Proof. A state offset adds the same scalar to the maximum and minimum and cancels. Replacing v by $v - w$ proves invariance under a load term. For $t \neq 0$, substitute $v = tu$ and use $\mathrm{osc}(tx) = |t|\,\mathrm{osc}(x)$; the case $t = 0$ is immediate. For arbitrary u, v,

$$\mathrm{osc}\{e + f - (u + v)^\top g\} \le \mathrm{osc}(e - u^\top g) + \mathrm{osc}(f - v^\top g).$$

Integrate, then let u, v approach their separate infima. Nonnegativity is inherited from the range of a finite vector. These arguments establish a seminorm; they do not assume that every zero-seminorm field admits a particular pointwise decomposition on an arbitrary measure space. □

G.2 A finite-error guarantee that ignores absorbable expert bias

Use $\mathcal{J}_C(q) = \mathbb{E}_q C + \tau\mathbb{E}\mathrm{KL}(q\|\nu)$, allowing $\tau = 0$ whenever the required optimizers exist. Let q_C minimize this objective on $\mathcal{Q}_\rho$ and let $\widehat{q}$ minimize $\mathcal{J}_{C+e}$ on the same set.

Theorem 7 (Capacity-adjusted error bound)**.** *The excess true objective obeys*

$$0 \le \mathcal{J}_C(\widehat{q}) - \mathcal{J}_C(q_C) \le \|e\|_{\rho,\mathrm{osc}}. \tag{26}$$

If $\widehat{q}$ is only η-optimal for its estimated objective, the right-hand side becomes $\|e\|_{\rho,\mathrm{osc}} + \eta$.

Proof. Optimality for the estimated objective gives

$$\mathcal{J}_C(\widehat{q}) + \mathbb{E}_{\widehat{q}} e \le \mathcal{J}_C(q_C) + \mathbb{E}_{q_C} e + \eta, \tag{27}$$

$$\mathcal{J}_C(\widehat{q}) - \mathcal{J}_C(q_C) \le \mathbb{E}_{q_C} e - \mathbb{E}_{\widehat{q}} e + \eta. \tag{28}$$

For any constant v, equal loads imply $\mathbb{E}_{q_C} v^\top g = \mathbb{E}_{\widehat{q}} v^\top g = v^\top \rho$. Subtract these equal terms from the last line. At each state, the difference between two probability averages of $e_a - v^\top g(a)$ is at most its range. Integration proves the claim for every v; taking the infimum proves Equation (26). The lower bound follows from optimality of q_C for the true objective. □

The bound explains why raw prediction error can overstate routing error. Suppose $e_a(h) = c(h) + v^\top g(a) + r_a(h)$ with $|r_a| \le \epsilon$. Then the bound is 2ϵ, independently of the sizes of c and v. For example, a predictor can overestimate every use of expert 1 by one hundred nats and still recover the exact constrained optimum after its price is refitted. An unadjusted mean squared error can be arbitrarily large in this example while routing regret is identically zero.

Proposition 3 (Sharpness of the range coefficient)**.** *The coefficient one multiplying the oscillation cannot be decreased uniformly over unregularized finite-state problems.*

Proof. Take two equiprobable states x_1, x_2, two top-1 experts, and target load $(1/2, 1/2)$. Fix $0 < d < 1$. Give true costs

$$C(x_1) = (0, d), \qquad C(x_2) = (d, 0)$$

and errors $e(x_1) = (0, -1)$, $e(x_2) = (-1, 0)$. The true optimum routes the first state to expert 1 and the second to expert 2; the estimated optimum reverses both. Their excess risk is d. For a load-price difference $v_2 - v_1 = t$, the two residual ranges are $|1 + t|$ and $|1 - t|$. Their mean has infimum one, attained for $|t| \le 1$. Thus the ratio of excess risk to adjusted oscillation equals d and tends to one as $d \uparrow 1$. Both assignments satisfy the same exact load. □

G.3 Load mismatch prevents cancellation of expert-specific error

The preceding cancellation requires the same load. It is useful to state the correction when a hard implementation is only approximately feasible. Write $\Delta\rho = \rho(\widehat{q}) - \rho(q_C)$.

Proposition 4 (Error and capacity slack)**.** *If Equation (28) holds but loads differ, then for every v,*

$$\mathcal{J}_C(\widehat{q}) - \mathcal{J}_C(q_C) \le \mathbb{E}\,\mathrm{osc}(e - v^\top g) - v^\top \Delta\rho + \eta. \tag{29}$$

Consequently $-v^\top \Delta\rho$ may be upper-bounded by $\|v\|_\infty \|\Delta\rho\|_1$.

Proof. Substitute $e = (e - v^\top g) + v^\top g$ into the right-hand side of Equation (28). The two load terms now differ by $-v^\top \Delta\rho$ instead of canceling. Bound the remaining difference of probability averages by its statewise range. The last inequality is the duality between the ℓ_∞ and ℓ_1 norms. □

The premise is explicit because independently optimizing over two different load sets does not by itself imply estimated optimality relative to q_C. For an implementation with an objective certificate, the formula identifies the price of residual load error. In particular, an arbitrarily large expert-specific predictor bias is harmless at exact load but need not be harmless under even a small unconstrained capacity change.

G.4 Differentiating the constrained optimum

The finite-error bound applies without a small-perturbation approximation. For a small error at positive temperature, we can say more: differentiating the load constraint identifies exactly which part of the error changes the policy. Assume bounded costs and error e, $\nu_a \ge \nu_{\min} > 0$, $\tau > 0$, and a relative-interior target. Work on the linear space $V = \mathrm{span}\{g(a) - g(b) : a, b \in \mathcal{A}\}$, fixing the orthogonal price gauge. Set

$$C_t = C + te, \qquad q_t = \operatorname*{arg\,min}_{q \in \mathcal{Q}_\rho} \mathcal{J}_{C_t}(q), \qquad \alpha_t \in V.$$

At $t = 0$, denote $q = q_0$ and introduce statewise centered quantities

$$\bar{g}(h) = \sum_a q_a g(a), \quad \widetilde{g}_a = g(a) - \bar{g}(h), \quad \bar{e}(h) = \sum_a q_a e_a, \quad \widetilde{e}_a = e_a - \bar{e}(h). \tag{30}$$

Define a matrix and vector on V by

$$K = \mathbb{E} \sum_a q_a \widetilde{g}_a \widetilde{g}_a^\top, \qquad b_{ge} = \mathbb{E} \sum_a q_a \widetilde{g}_a \widetilde{e}_a. \tag{31}$$

Positivity of all action probabilities implies that K is positive definite on V. Indeed, $z^\top K z = 0$ implies that $z^\top g(a)$ is constant across all actions at almost every state, which is possible for $z \in V$ only when $z = 0$.

Theorem 8 (Price-projected routing response)**.** *The maps $t \mapsto \alpha_t, q_t$ are differentiable near zero. With*

$$r_a(h) = \widetilde{e}_a(h) - \widetilde{g}_a(h)^\top K^{-1} b_{ge}, \tag{32}$$

their derivatives are

$$\dot{\alpha}_0 = -K^{-1} b_{ge}, \qquad \dot{q}_{0,a}(h) = -\frac{q_a(h)}{\tau} r_a(h). \tag{33}$$

Moreover $\sum_a q_a r_a = 0$ statewise and $\mathbb{E} \sum_a q_a \widetilde{g}_a r_a = 0$.

Proof. The price equation is

$$\mathcal{R}(t, \alpha) := \mathbb{E} \sum_a g(a) \frac{\nu_a \exp[-(C_a + t e_a + \alpha^\top g(a))/\tau]}{\sum_b \nu_b \exp[-(C_b + t e_b + \alpha^\top g(b))/\tau]} - \rho = 0.$$

Boundedness and a positive reference permit differentiation under the integral in a neighborhood of $(0, \alpha_0)$. Direct differentiation gives $\partial_\alpha \mathcal{R} = -K/\tau$ and $\partial_t \mathcal{R} = -b_{ge}/\tau$. The first matrix is invertible on V, so the finite-dimensional implicit function theorem supplies a differentiable price curve. Differentiating $\mathcal{R}(t, \alpha_t) = 0$ yields $K\dot{\alpha}_0 = -b_{ge}$.

To differentiate the probabilities, use their log form:

$$\log q_{t,a} = \log \nu_a - (C_a + t e_a + \alpha_t^\top g(a))/\tau - \log Z_t.$$

The derivative of $\log Z_t$ equals $-\sum_b q_{t,b}(e_b + \dot{\alpha}_t^\top g(b))/\tau$. Subtracting this average and setting $t = 0$ produces $\partial_t \log q_{t,a}|_0 = -r_a/\tau$, which proves the probability derivative. Statewise centering proves the first orthogonality relation. The second is $b_{ge} - K K^{-1} b_{ge} = 0$ by definition. □

The residual is an orthogonal projection in the joint probability space $\mu(dh)q(a \mid h)$ after state centering. It separates the state baseline and the expert bias removed by prices from the residual that changes decisions. A single global price vector cannot remove that residual.

G.5 The exact second-order price of misspecification

Let $V(t) = \min_{q \in \mathcal{Q}_\rho} \mathcal{J}_{C+te}(q)$ and let $S(t) = \mathcal{J}_C(q_t) - \mathcal{J}_C(q_0)$ denote regret measured against the original objective, not the perturbed objective used to choose q_t.

Theorem 9 (Projected error determines local regret)**.** *Under the bounded assumptions above,*

$$V'(0) = \mathbb{E}_q e, \tag{34}$$

$$V''(0) = -\frac{1}{\tau} \mathbb{E} \sum_a q_a r_a^2, \tag{35}$$

$$S(t) = \frac{t^2}{2\tau} \mathbb{E} \sum_a q_a r_a^2 + O(t^3). \tag{36}$$

The coefficient vanishes precisely when $r = 0$ almost surely under μq. It is invariant to adding any bounded state offset or linear load term to e.

Proof. For small s, optimality at t and $t+s$ bounds the difference quotient of V between $\mathbb{E}_{q_{t+s}} e$ and $\mathbb{E}_{q_t} e$. Continuity gives $V'(t) = \mathbb{E}_{q_t} e$. Hence

$$V''(0) = \mathbb{E} \sum_a \dot{q}_{0,a} e_a = -\frac{1}{\tau} \mathbb{E} \sum_a q_a r_a e_a.$$

Subtract $\bar{e}$ using $\sum_a q_a r_a = 0$, then substitute $\widetilde{e} = r + \widetilde{g}^\top K^{-1} b_{ge}$. The load-orthogonality relation removes the second term, leaving the asserted squared residual.

For the final expansion, both q_t and q_0 have load ρ. The exact regret identity therefore gives $S(t) = \tau \mathbb{E}\mathrm{KL}(q_t \| q_0)$. Write $q_{t,a} = q_a + t\dot{q}_a + O(t^2)$ uniformly. A positive lower bound on q_a follows from bounded costs, the finite price at zero, and the lower bound on ν_a. Taylor expansion of the finite-dimensional KL divergence at q_0 is therefore uniform and yields

$$\mathrm{KL}(q_t \| q_0) = \frac{t^2}{2} \sum_a \frac{\dot{q}_a^2}{q_a} + O(t^3).$$

Insert Equation (33), integrate, and multiply by τ. Smoothness through third order and domination follow from the same bounded exponential-family derivatives. The zero condition follows from nonnegativity, and invariance follows directly from the projection. □

G.6 A COMPUTABLE SENSITIVITY DECOMPOSITION

The quadratic coefficient can be evaluated without differentiating a trained router. Expanding the squared residual gives

$$\mathbb{E} \sum_a q_a r_a^2 = \mathbb{E}\, \mathrm{Var}_{q(\cdot|H)}(e_a(H)) - b_{ge}^\top K^{-1} b_{ge}. \tag{37}$$

The first term is the conditional error variance before repricing; the second is the part explained by the load features. The proof below shows that the second cannot exceed the first. This decomposition concerns the declared constrained optimizer and does not give a mechanism interpretation to an arbitrary regression R^2.

Proof. For any $z \in V$,

$$\mathbb{E} \sum_a q_a (\widetilde{e}_a - \widetilde{g}_a^\top z)^2 = \mathbb{E}\, \mathrm{Var}_q(e) - 2 z^\top b_{ge} + z^\top K z.$$

The right side is a strictly convex quadratic with unique minimizer $z = K^{-1} b_{ge}$. Its minimum is Equation (37); a squared norm cannot be negative. This also proves that using any approximate price response z introduces the additional sensitivity $(z - K^{-1} b_{ge})^\top K (z - K^{-1} b_{ge})$. □

One may estimate K, b_{ge} using independent candidate groups, solve on V, and report the residual energy with group-level uncertainty. Regularizing the inverse changes the measured projection and must be reported. A small eigenvalue of K means that a price direction is poorly observed, not that the corresponding expert is intrinsically unimportant.

G.7 A FINITE-TEMPERATURE IDENTITY ALONG THE ENTIRE PERTURBATION PATH

The second-order coefficient has an exact integral counterpart. Let r_t be the residual formed with $q_t, K_t, b_{ge,t}$, and assume the bounded relative-interior conditions hold along $0 \le t \le T_{\mathrm{end}}$. Then

$$S(t) = \frac{1}{\tau} \int_0^t s\, \mathbb{E} \sum_a q_{s,a} r_{s,a}^2 \, ds. \tag{38}$$

Proof. The definition implies $S(t) = V(t) - tV'(t) - V(0)$. Differentiate to obtain $S'(t) = -tV''(t)$. Apply the response calculation at each point of the path, where the direction e is unchanged, giving $V''(t) = -\mathbb{E}_{q_t} r_t^2 / \tau$. Integrate from zero, using $S(0) = 0$. □

This identity extends the local calculation along a path, provided its assumptions continue to hold. If the residual energy on the path lies between σ_-^2 and σ_+^2, then

$$\frac{t^2\sigma_-^2}{2\tau} \le S(t) \le \frac{t^2\sigma_+^2}{2\tau}.$$

The upper bound quantifies robust correction only over the verified path. It does not substitute for a bound on distribution shift or on a change of candidate support, which are different perturbations.

G.8 How this extends the main mechanism test

The pairwise log-odds condition in the main text is an identification statement for a fixed objective. The present results ask what happens when that objective is misspecified. They yield three distinguishable outcomes: large raw error but zero projected error, nonzero projected error with quadratic regret, and a load mismatch that prevents cancellation of expert bias. These outcomes are testable with complete candidate costs. An evaluation of a compatibility head should therefore report its action differences, the target load, and whether prices are refitted, instead of equating a raw prediction metric with policy quality.

The argument uses standard implicit differentiation of constrained entropic transport. Its routing interpretation lies in the residual in Equation (32) and the associated loss coefficient. Appendix H shows how both quantities become boundary objects as temperature vanishes.

H Boundary Displacement, Capacity Projection, and the Hard-Routing Limit

This section proves the hard-routing counterpart of the cost-response theorem. A boundary tube counts states near a fixed router boundary. Here the boundary itself moves because the compatibility cost is perturbed and the capacity prices are refitted. The resulting excess predictive risk has a different mathematical structure: it is a quadratic form of the component of boundary error that no global expert price can remove. We derive this form, describe its null directions, and connect it to the positive-temperature response from Appendix G.

H.1 Regular cost partitions and orientation conventions

Let $\mathcal{M}$ be a compact smooth r-dimensional state manifold, or a compact region with all relevant interfaces away from its exterior boundary. The density f is C^2. Costs C_a and perturbations e_a are C^3 with uniformly bounded derivatives on the region. Let α_0 be an optimal finite price for a relative-interior target, and set

$$D_a = C_a + \alpha_0^\top g(a), \qquad A_a = \{h : D_a(h) < D_b(h) \text{ for every } b \neq a\}.$$

For an action map $a(h)$, write $L_C(a) = \int C_{a(h)}(h)d\mu(h)$. Ties have zero μ measure. An active interface Γ_{ab} is the regular part on which a, b are the only minimum-cost actions. Count each unordered pair once and fix its orientation from a to b. On this interface define

$$u_{ab} = D_b - D_a, \quad v_{ab} = g(b) - g(a), \quad d\omega_{ab} = \frac{f}{|\nabla_{\mathcal{M}} u_{ab}|} d\mathcal{H}^{r-1}. \tag{39}$$

The a cell is the positive side of u_{ab}. Changing orientation reverses u, v, and the error difference simultaneously, so all formulas below are orientation invariant.

We assume gradients of active cost differences are uniformly nonzero. The finitely many higher-order junctions have transverse local coordinates, with their codimension-two tubes of volume $O(t^2)$. The corresponding cost differences are $O(t)$ inside a tube of radius $O(t)$. These conditions ensure that junctions affect a second-order risk expansion only by $O(t^3)$. They also rule out the creation of a positive-measure tie plateau. The binary case needs no junction assumption.

The load-feature space is $V = \operatorname{span}\{g(a) - g(b)\}$. Define

$$G_\partial = \sum_{a<b} \int_{\Gamma_{ab}} v_{ab} v_{ab}^\top d\omega_{ab}, \qquad b_\partial = \sum_{a<b} \int_{\Gamma_{ab}} v_{ab}(e_b - e_a)\, d\omega_{ab}. \tag{40}$$

We require G_∂ to be positive definite on V; inverses throughout this section mean inverses on that space. For top-1 actions, G_∂ is the weighted graph Laplacian of the active-cell adjacency graph. For joint top-k actions it is a Gram matrix of dispatch differences, not necessarily an ordinary graph Laplacian on experts.

H.2 Deriving the derivative of a cell's load

Let $v \in V$ and perturb the reduced costs to $D_a + t(e_a + v^\top g(a))$. On an active interface write $d_{ab} = e_b - e_a + v^\top v_{ab}$. The new tie is

$$u_{ab}(h) + t d_{ab}(h) = 0.$$

Use $u = u_{ab}$ as a transverse coordinate, with tangential coordinate z. The implicit function theorem gives a perturbed threshold

$$u_t(z) = -t d_{ab}(z, 0) + O(t^2) \tag{41}$$

uniformly on each compact interface chart. The density and volume element in these coordinates have product $w(z, u) = w(z, 0) + O(|u|)$, with $w(z, 0)dz = d\omega_{ab}$.

For $d_{ab} > 0$, the a cell gains the strip $-t d_{ab} < u < 0$ from b. Its dispatch change is $g(a) - g(b) = -v_{ab}$. For $d_{ab} < 0$, the opposite strip is transferred and gives the same signed formula. Consequently

$$\frac{d}{dt}\rho(q_t)\bigg|_0 = -b_\partial - G_\partial v. \tag{42}$$

To justify the summation, remove small neighborhoods of junctions, integrate the signed strips on a finite chart cover with a partition of unity, and then bound junction volume by $O(t^2)$. Division by t makes the latter vanish. The derivative is locally Lipschitz in the perturbation and price parameters. In each regular chart its surface integrand and Jacobian vary Lipschitz-continuously by the C^2 density and C^3 cost assumptions. The boundary of an active facet moves by order equal to the parameter change: transversality gives a regular coordinate for each junction within that facet. The excluded surface strip therefore has measure of the same order, and its integrand is uniformly bounded. A finite chart cover bounds both the interior change and these moving-facet terms by a constant times the parameter change. Thus the derivative is locally Lipschitz, not just continuous, within the stated regularity class. Uniform invertibility on V makes the implicit price curve's derivative locally Lipschitz, justifying its order-t^2 remainder in the following risk expansion.

H.3 Capacity-neutral boundary error determines quadratic regret

For $C_t = C + te$, let α_t be the nearby price yielding the original load, and let $a_t(h)$ minimize $C_{t,a} + \alpha_t^\top g(a)$. Put

$$r^\partial_{ab}(h) = e_b(h) - e_a(h) - v_{ab}^\top G_\partial^{-1} b_\partial. \tag{43}$$

Theorem 10 (Capacity-neutral boundary regret)**.** *Under the preceding regularity and invertibility assumptions, a differentiable load-preserving price curve exists near zero and*

$$\dot\alpha_0 = -G_\partial^{-1} b_\partial, \tag{44}$$

$$L_C(a_t) - L_C(a_0) = \frac{t^2}{2} \sum_{a<b} \int_{\Gamma_{ab}} (r^\partial_{ab})^2 \, d\omega_{ab} + O(t^3). \tag{45}$$

The quadratic coefficient equals

$$\frac{1}{2}\left[\sum_{a<b} \int_{\Gamma_{ab}} (e_b - e_a)^2 d\omega_{ab} - b_\partial^\top G_\partial^{-1} b_\partial\right]. \tag{46}$$

Proof. By the preceding chart calculation, the cell-load map is continuously differentiable and its derivative with respect to price is $-G_\partial$ on V. The implicit function theorem applied to the equation $\rho(a_t) = \rho(a_0)$ gives a nearby price curve and $G_\partial \dot\alpha_0 = -b_\partial$. The second-order threshold term is uniformly bounded by the smooth chart assumptions and the inverse bound on G_∂.

Because both policies have identical load, the cost compensation identity allows us to replace C by the original reduced costs D:

$$L_C(a_t) - L_C(a_0) = \int [D_{a_t(h)}(h) - D_{a_0(h)}(h)] d\mu(h).$$

The integrand is zero where the selected action does not change. In a two-cell switching strip its value is exactly $|u_{ab}|$. Using Equation (41), its chart integral is

$$\int dz \int_{\min(0,u_t(z))}^{\max(0,u_t(z))} |u| w(z,u) du = \frac{1}{2} \int u_t(z)^2 w(z,0) dz + O(t^3) \tag{47}$$

$$= \frac{t^2}{2} \int (r_{ab}^{\partial})^2 d\omega_{ab} + O(t^3). \tag{48}$$

The first error follows from integrating $O(u^2)$ over a strip of width $O(t)$. Replacing u_t by its first-order term changes its squared value by $O(t^3)$. A partition of unity combines charts without double counting. At transverse junctions, the affected volume is $O(t^2)$ and all relevant cost differences are $O(t)$, hence their contribution is $O(t^3)$. Summing gives Equation (45).

Finally expand the square in Equation (43). The cross term is $-2b_{\partial}^{\top} G_{\partial}^{-1} b_{\partial}$, and the price-square term is $b_{\partial}^{\top} G_{\partial}^{-1} G_{\partial} G_{\partial}^{-1} b_{\partial}$. Their sum gives Equation (46). □

The expansion now assigns a predictive cost to the switching strips derived above. After capacity is restored, their leading contribution is the squared boundary residual. Large raw boundary error may produce no quadratic regret if it is a pure price shift. Conversely, errors with zero net load displacement can still have positive energy and incur loss while leaving every expert's load unchanged.

H.4 A BINARY FORMULA: WEIGHTED VARIANCE AFTER REPRICING

With two top-1 actions, write $\omega = \omega_{12}$ and $W_0 = \int_\Gamma d\omega > 0$. Let $d = e_2 - e_1$ on the boundary and $\bar{d} = W_0^{-1} \int d\, d\omega$. The theorem reduces to

$$L_C(a_t) - L_C(a_0) = \frac{t^2 W_0}{2} \operatorname{Var}_{\omega/W_0}(d) + O(t^3). \tag{49}$$

Proof. There is only one observable price difference. The load derivative is minus $\int (d + \dot{\alpha}_2 - \dot{\alpha}_1) d\omega$. Its vanishing gives $\dot{\alpha}_2 - \dot{\alpha}_1 = -\bar{d}$. Substitute $r = d - \bar{d}$ into Equation (45) and apply the definition of variance. □

The variance in this formula is taken over the active interface. Errors deep inside a cell do not enter its leading coefficient. A head that is accurate throughout cell interiors but varies incorrectly along the interface can be worse than a head with a large constant expert bias. This identifies the interface variation that local cost supervision needs to resolve.

H.5 AN EXACT PREDICTIVE EXAMPLE WITH FIXED EXPERT LOAD

Take $H = (X, Z)$ uniform on $[-1, 1]^2$ and the Gaussian predictive model from Appendix C.2: $C_- - C_+ = 2X$. The original optimum selects $+$ for $X > 0$ and has equal expert loads. Perturb the cost difference by $t\psi(Z)$, with bounded ψ and $|t| \|\psi - \mathbb{E}\psi\|_\infty < 2$. The equal-load price difference subtracts $t\mathbb{E}\psi$ exactly, because the conditional density of X is constant. The new decision is

$$2X + t\{\psi(Z) - \mathbb{E}\psi(Z)\} > 0.$$

The conditional boundary is $X_t(Z) = -t\{\psi(Z) - \mathbb{E}\psi(Z)\}/2$, whose mean is zero. The full excess risk, not merely a small-tube proxy, is

$$L_C(a_t) - L_C(a_0) = \frac{t^2}{8} \operatorname{Var}(\psi(Z)). \tag{50}$$

Proof. At fixed $Z = z$, only points between zero and $X_t(z)$ change action. Their true risk difference is $2|X|$. Since X has density $1/2$, the conditional excess equals $\int_0^{|X_t(z)|} 2x\,(1/2)dx = X_t(z)^2/2$. Average over Z and substitute the expression for X_t. There is no remainder because the cost gap and transverse density are exactly linear and constant, respectively. □

For $\psi(Z) = c$, even an arbitrarily large constant c disappears after repricing. For $\psi(Z) = Z$, the regret is $t^2/24$. For $\psi(Z) = \sqrt{3}Z$, it is $t^2/8$. These policies have the same expert loads. The difference is the interface variation of prediction error, not a failure of aggregate load balancing.

H.6 Price sensitivity and weakly connected expert interfaces

For top-1 routing, let $w_{ab} = \int_{\Gamma_{ab}} d\omega_{ab}$. The matrix G_∂ has entries $(G_\partial)_{aa} = \sum_{b\neq a} w_{ab}$ and $(G_\partial)_{ab} = -w_{ab}$ for $a \neq b$. If the adjacency graph is connected, its smallest positive eigenvalue $\lambda_2(G_\partial)$ satisfies

$$\|\dot{\alpha}_0\|_2 \leq \frac{\|b_\partial\|_2}{\lambda_2(G_\partial)}. \tag{51}$$

The estimate is sharp along an eigenvector for $\lambda_2(G_\partial)$. This identifies a different kind of instability from a visible routing flip: prices can be highly sensitive when capacity must be exchanged through an interface with little data mass.

Proof. On the zero-sum gauge space, diagonalize G_∂ in an orthonormal eigenbasis. Each coefficient of $\dot{\alpha}_0 = -G_\partial^{-1} b_\partial$ is divided by the corresponding positive eigenvalue. The largest such divisor is $1/\lambda_2(G_\partial)$. Equality holds when b_∂ is an eigenvector for the smallest positive eigenvalue. □

The Laplacian form of semi-discrete mass derivatives is established in optimal transport (Kitagawa et al., 2019). Our use of that derivative is to remove capacity-absorbable predictive error and compute the remaining loss coefficient. A small eigenvalue alone does not force large regret: it must be paired with an error direction that excites that mode.

H.7 Why perturbation risk is not the same as smoothing bias

Entropy smoothing changes the temperature of an otherwise fixed objective. Cost misspecification changes the objective while retaining its constraints. These operations have different expansions and must not be conflated. Second-order entropic cost asymptotics for semi-discrete transport are known (Altschuler et al., 2022); the calculation below instead takes a cost derivative at each fixed temperature and then studies the limit of that derivative as the temperature vanishes.

Let q_τ solve the original constrained entropic problem and let $K_\tau, b_{ge,\tau}$ be its covariance matrix and error-covariance vector from Equation (31). Assume its prices converge to α_0, the regular active interfaces above persist, and all inactive actions are uniformly separated on compact subsets away from those interfaces. Assume the reference is smooth and bounded away from zero. Define

$$E_\partial = \sum_{a<b} \int_{\Gamma_{ab}} (e_b - e_a)^2 d\omega_{ab}.$$

Theorem 11 (Soft-to-hard limit of predictive-error sensitivity)**.** *Under these assumptions,*

$$\tau^{-1} K_\tau \longrightarrow G_\partial, \qquad \tau^{-1} b_{ge,\tau} \longrightarrow b_\partial, \tag{52}$$

$$\tau^{-1} \mathbb{E}\,\mathrm{Var}_{q_\tau}(e) \longrightarrow E_\partial, \tag{53}$$

and therefore

$$\lim_{\tau\downarrow 0} \frac{1}{2\tau} \mathbb{E} \sum_a q_{\tau,a} r_{\tau,a}^2 = \frac{1}{2}(E_\partial - b_\partial^\top G_\partial^{-1} b_\partial). \tag{54}$$

The equality concerns the limit of local cost-response coefficients. It does not interchange arbitrary finite cost perturbations with the zero-temperature limit.

Proof. For any finite probability vector and action-valued quantities x_a, y_a,

$$\mathrm{Cov}_q(x, y) = \sum_{a<b} q_a q_b (x_a - x_b)(y_a - y_b)^\top. \tag{55}$$

This follows by expanding half of the double sum over ordered pairs. It suffices to prove convergence of the measures $\tau^{-1} q_{\tau,a} q_{\tau,b} \mu$ to ω_{ab} on each active interface, with zero contribution elsewhere.

On a chart with exactly two competitive actions, let u_τ be their price-adjusted cost gap. After including the smooth reference ratio, their probabilities have the logistic form up to exponentially small contributions from the separated actions. Write $\sigma(v) = (1+e^{-v})^{-1}$ and $k(v) = \sigma(v)(1-\sigma(v))$. The relevant kernel is $\tau^{-1} k(u_\tau/\tau + c(h))$, with a bounded smooth shift c from the reference. Its transverse integral tends to one:

$$\int_{\mathbb{R}} k(v + c) dv = \sigma(+\infty) - \sigma(-\infty) = 1.$$

Use a chart centered on the perturbed zero set $u_\tau = 0$; price convergence and regularity make its surface measure converge to that of $u = 0$. Coarea and dominated convergence then give, for continuous bounded φ,

$$\lim_{\tau \downarrow 0} \int \varphi(h) \frac{q_{\tau,a} q_{\tau,b}}{\tau} d\mu(h) = \int_{\Gamma_{ab}} \varphi(h) d\omega_{ab}.$$

For completeness, junctions can be excluded with a shrinking gap window $a_\tau = \tau K |\log \tau|$. A transverse simultaneous-tie region has volume $O(a_\tau^2)$, so the crude bound $q_a q_b/\tau \le 1/(4\tau)$ gives $O(\tau |\log \tau|^2) \to 0$ there. Outside this region and outside the two-action charts, at least one of the pair probabilities is bounded by a constant times $\exp(-a_\tau/\tau) = \tau^K$. Choose $K > 2$; the integral after division by τ vanishes. Compactness and a finite chart cover make all these estimates uniform.

Use these limiting measures in the covariance identities for (g, g), (g, e), and (e, e) to obtain Equation (53). Since G_∂ is positive definite on V, inverses of K_τ/τ converge to G_∂^{-1} there. Finally use Equation (37) and take the limit term by term. □

H.8 What the limit says about sparse expert interfaces

At positive temperature, predictive-error sensitivity is distributed over states according to conditional action variance. In the hard limit that variance concentrates on active compatibility boundaries. The limiting coefficient retains the capacity projection, giving price-adjusted prediction error the same role in the soft and hard formulations. The connection applies to the constrained optima studied here. Applying it to a pretrained router still requires checking whether that router follows the corresponding optimal cost partition.

For joint top-k routing, apply the theorem to complete actions and use $v_{ab} = g(b) - g(a)$. An action transition may exchange several experts. Its predictive loss remains the loss of the complete interface; replacing it by a sum of independent expert costs would change $e_b - e_a$ and in general change the boundary response. The covariance and surface formulas therefore preserve the nonadditive definition used in the main text.

H.9 Cases in which the quadratic expansion does not apply

The regularity assumptions explain when this local description can fail. Three examples separate the possible obstructions. First, at fixed prices, if a cost gap is $u = x^3$ in one dimension, its gradient vanishes on the tie, so a constant perturbation moves the boundary by $O(t^{1/3})$, not $O(t)$. Second, a positive-measure region with $D_a = D_b$ can change assignments at order one under an arbitrarily small perturbation. Third, a disconnected active adjacency graph makes some capacity transfers invisible to first-order price changes. None of these settings satisfies the invertibility and transversality hypotheses.

In the first example, uniform density near zero gives exact excess risk proportional to $\int_0^{t^{1/3}} x^3 dx = t^{4/3}/4$, not t^2. The different exponent follows from the degenerate cost gap at fixed prices. In a two-expert fixed-load problem, refitting prices instead cancels this constant cost difference (Equation (7)).

A quadratic certificate therefore requires checking cost-gap regularity before fitting a scaling window. The general weighted tube framework remains meaningful, but its exponent must match the actual transverse order.

I JOINT ACTIONS AND FINITE-BATCH CAPACITY CERTIFICATES

The preceding response laws enforce load in expectation. A finite batch adds an integer constraint: each token must select one complete action, and each expert has a fixed number of slots. We now ask which population-price arguments also certify this finite assignment. The optimization facts used below are classical linear programming and minimum-cost flow; see Ahuja et al. (1993). The routing conclusions concern what can, and cannot, be certified by independent single-token or single-expert comparisons.

I.1 AN EMPIRICAL ACTION PROBLEM

Let a batch contain n states $h_1, \dots, h_n$. Write $C_{ta} = C_a(h_t)$ and let $\mathcal{A}_t$ contain the permitted complete actions at state t. The relaxed batch problem is

$$\min_x \frac{1}{n} \sum_{t=1}^{n} \sum_{a \in \mathcal{A}_t} x_{ta} C_{ta}, \qquad \sum_a x_{ta} = 1, \quad x_{ta} \geq 0, \quad \sum_{t,a} x_{ta} g(a) = b. \tag{56}$$

Here b counts dispatches, not normalized load; necessarily $\sum_i b_i = nk$ for binary top-k actions. The integral problem adds $x_{ta} \in \{0, 1\}$. The states and cost table are fixed; the batch objective is additive across states. Interacting simultaneous interventions require complete group assignments as actions (Appendix N). Each state cost evaluates complete interfaces, including the declared weight normalization, without decomposing their losses into independent expert contributions.

For any prices α, define

$$m_t(\alpha) = \min_{a \in \mathcal{A}_t} \{C_{ta} + \alpha^\top g(a)\}.$$

Summing the pointwise lower bounds gives the empirical dual certificate

$$\frac{1}{n} \sum_{t,a} x_{ta} C_{ta} \geq \frac{1}{n} \left\{ \sum_t m_t(\alpha) - \alpha^\top b \right\}. \tag{57}$$

The gap is exactly the average nonnegative reduced-cost slack. This identity is useful even when solving the full assignment problem is too expensive: any feasible returned assignment and any prices give a computable upper bound on its excess cost relative to the relaxed optimum.

Proposition 5 (Complete-action certificate)**.** *Suppose an integral assignment a_t has load b. If there exists α such that*

$$C_{ta_t} + \alpha^\top g(a_t) \leq C_{ta} + \alpha^\top g(a) \quad (t \leq n,\ a \in \mathcal{A}_t),$$

then this assignment is optimal in both the relaxed and integral problems. More generally, if these inequalities hold with nonnegative slacks ϵ_t, its excess cost relative to either optimum is at most $n^{-1} \sum_t \epsilon_t$.

Proof. Subtract Equation (57) from the assignment cost. Feasibility cancels the price term, leaving

$$\frac{1}{n} \sum_t \{C_{ta_t} + \alpha^\top g(a_t) - m_t(\alpha)\}.$$

Each summand is at most ϵ_t. The dual value is a lower bound on both optima. Zero slacks give equality with that bound, proving optimality without any assumption that the joint-action relaxation is integral. □

This certificate is sufficient for optimality. Its converse can fail for integral top-k assignments because the relaxation may have an integrality gap. Thus failure to find certifying prices need not imply suboptimality. For top-1, the bipartite-flow structure removes this obstruction and yields the converse below.

I.2 EXACT-LOAD IMPROVEMENTS ARE CYCLES, NOT ISOLATED SWAPS

For top-1, write the current expert of token t as $i(t)$. Construct a directed multigraph on experts: every permitted reassignment $i(t) \to j$ is an edge carrying that token and weight

$$w_{tj} = C_{tj} - C_{t,i(t)}.$$

Different tokens assigned to the same expert give parallel edges. An edge alone changes load. A directed cycle changes one outgoing and one incoming dispatch at every visited expert and therefore preserves exact counts.

Theorem 12 (Cycle characterization of batch regret)**.** *For a feasible top-*1 *assignment with fixed integer expert counts, the following are equivalent:*

1. *no other permitted integral assignment has lower total cost;*
2. *its residual expert graph has no negative-weight directed cycle;*
3. *there are expert prices* α_i *with* $w_{tj} + \alpha_j - \alpha_{i(t)} \geq 0$ *for every edge.*

Every improving alternative decomposes into capacity-preserving cycles, and its total cost change is the sum of their edge weights.

Proof. A negative closed walk contains a negative simple cycle: repeatedly remove the segment between repeated vertices until the walk is a union of simple cycles; if their sum is negative, at least one is negative. On a simple cycle each source expert occurs once. Its edges therefore refer to distinct tokens, and applying all its reassignments simultaneously is feasible and preserves every expert count. Thus the first assertion implies the second.

Conversely, compare the present assignment to any other feasible assignment. For every changed token draw its old-to-new edge. Equality of expert counts means that each vertex has equal incoming and outgoing degree. The resulting finite balanced directed multigraph decomposes into cycles: follow outgoing edges until a vertex repeats, remove that cycle, and continue. The objective difference is the sum of the cycle costs. Without a negative cycle this sum cannot be negative, proving the reverse implication.

If prices exist, summing their inequalities around a cycle cancels the potentials, so that cycle cannot be negative. To prove existence in the other direction, add a source with zero-weight edges to every expert. Shortest-path distances d_i are finite in a finite graph without negative cycles and satisfy $d_j \leq d_i + w_{tj}$. Set $\alpha_i = -d_i$; the required inequality follows by rearrangement. Finally, the preceding decomposition establishes the stated additive accounting of any feasible improvement. □

This classical minimum-cost-flow certificate explains why independent routing checks can miss an improvement. Preserving exact counts may require changing several tokens together, even when no individual reassignment is feasible. Testing a route under an unchanged capacity constraint can therefore require coordinating several states instead of independently choosing their best experts.

A three-expert obstruction to pairwise auditing. Take three tokens with native experts $1, 2, 3$, one slot per expert, and the following cost matrix:

$$C = \begin{pmatrix} 1 & 0 & 3 \\ 3 & 1 & 0 \\ 0 & 3 & 1 \end{pmatrix}.$$

The native total cost is 3. Every two-token exchange changes the total from 2 to 3 and is rejected by a pairwise test. The three-cycle $1 \to 2 \to 3 \to 1$ instead has total cost 0. All costs are nonnegative, and multiplying the matrix by any constant makes the missed gain arbitrarily large. Pairwise exchange stability is therefore not an approximation guarantee for exact-load assignment, even in a three-token batch with complete support.

I.3 What candidate restriction makes feasible

Let S_t be the allowed experts for a top-1 token and let b_i be nonnegative integer capacities with $\sum_i b_i = n$. For a token subset T, write $N(T) = \bigcup_{t \in T} S_t$. Feasibility is equivalent to

$$|T| \le \sum_{i \in N(T)} b_i \qquad \text{for every } T \subseteq \{1, \dots, n\}. \tag{58}$$

This is the capacitated matching condition. To see it directly, create a network with source-to-token capacity 1, token-to-permitted-expert capacity $n+1$, and expert-to-sink capacity b_i. Any cut of capacity below n must retain all neighbors of its retained tokens. Its capacity is then $n - |T| + \sum_{i \in N(T)} b_i$, which is at least n precisely under Equation (58). The integral max-flow theorem gives a feasible integral assignment. Necessity follows by counting the slots available to T in any feasible assignment.

Including each native expert in S_t preserves feasibility for the native load vector. It does not preserve feasibility for an arbitrary desired load. For example, if every token permits only experts 1 and 2, a positive target for expert 3 is unattainable, regardless of how dual prices are updated. No price update can create a feasible assignment on that support.

For general complete actions the exact relaxed condition is instead

$$b \in \sum_{t=1}^{n} \operatorname{conv}\{g(a) : a \in \mathcal{A}_t\}, \tag{59}$$

where the sum is a Minkowski sum. Equivalently, for every $v \in \mathbb{R}^E$,

$$v^\top b \le \sum_t \max_{a \in \mathcal{A}_t} v^\top g(a).$$

The forward implication is immediate by applying a linear functional to each convex combination. The converse follows by separating a point from the compact convex Minkowski sum. Integral feasibility is stronger; the support-function condition characterizes only fractional dispatch.

I.4 Expert interactions that per-expert scores cannot represent

The complete-action cost in Equation (1) is essential even for four experts and top-2. Fix a state and assign the six unordered pairs costs

$$C_{12} = C_{34} = 0, \qquad C_{13} = C_{14} = C_{23} = C_{24} = K > 0. \tag{60}$$

Each expert has the same average cost across its three partners, $2K/3$. Expertwise conditional means therefore provide no information about which pairs are compatible.

Proposition 6 (A joint-interface interaction obstruction)**.** *No additive expert cost $d_i + d_j + c$ represents Equation (60). Its best uniform additive approximation has error at least $K/2$. Adding expert prices cannot remove this obstruction. Two stochastic top-2 policies can have identical expert loads while their expected compatibility costs differ by K.*

Proof. Every additive representation satisfies the three matching identities

$$(d_1 + d_2 + c) + (d_3 + d_4 + c) = (d_1 + d_3 + c) + (d_2 + d_4 + c) = (d_1 + d_4 + c) + (d_2 + d_3 + c).$$

The corresponding sums of true costs are $0, 2K, 2K$. If all six approximation errors were at most ϵ, the difference between the first two sums would be at most 4ϵ, giving $\epsilon \ge K/2$. This bound is attained by predicting $K/2$ for every pair. Expert prices add $\alpha_i + \alpha_j$ and preserve the matching identities, so they do not alter the interaction discrepancy.

For the last statement, let q choose $\{1, 2\}$ or $\{3, 4\}$ with probability $1/2$ each, while r chooses $\{1, 3\}$ or $\{2, 4\}$ with the same probabilities. Both dispatch each expert with probability $1/2$. Their expected costs are respectively 0 and K. Thus load and averaged per-expert scores cannot distinguish the policies, although complete-action compatibility does so exactly. □

This example has a direct predictive realization. Take deterministic target $Y = 0$ and four expert output vectors $v, -v, w, -w$ in $\mathbb{R}^2$, where v, w are orthogonal and have squared norm $2K$. The interface is the sum of the two chosen outputs and the prediction is $\mathcal{N}(\Phi_{ij}, I)$. After subtracting the shared Gaussian normalization constant, its negative log-likelihood is $\|\Phi_{ij}\|^2/2$. The complementary pairs cost zero; each mixed pair costs $2K$. Rescaling v, w by $1/\sqrt{2}$ gives exactly Equation (60). Thus an ordinary downstream predictive loss realizes the interaction obstruction.

I.5 Two levels of certification

These results separate two questions that can otherwise be conflated in a audit with M candidate actions. First, has the declared family of complete interfaces been evaluated sufficiently to distinguish complementary expert sets? Second, can the proposed per-state changes be assembled into a feasible assignment under the requested capacity? The first is a cost-model question and the second is a constrained-allocation question. More accurate scalar expert scores do not resolve missing interactions; more dual iterations do not repair an infeasible support.

A practical certificate records the action family, empirical loads, price vector, and reduced-cost slack. With exact forced costs these determine Equation (57) directly. With estimated costs, the quotient error bounds of Appendix G control the difference between the estimated and true certificates. Keeping those two error sources separate gives a reproducible account of both statistical approximation and finite-batch feasibility.

J Capacity Response to Changes in the State Distribution

Vision-language routing changes both the cost of an action and the mixture of states competing for expert capacity. A long image prefix and a short text answer need not place the same demand on the routed layers. To isolate this effect, we first hold conditional action costs fixed and vary only the state distribution. We then combine this change with the cost perturbation from Appendix G. In both cases the relevant population quantity is the induced change in expert demand.

J.1 A differentiable change of measure

Fix a reference probability measure μ and a bounded measurable function z satisfying $\mathbb{E}_\mu z = 0$. For sufficiently small t let

$$d\mu_t = (1 + tz)d\mu. \tag{61}$$

The action costs, reference kernel ν, temperature $\tau > 0$, and target load ρ initially remain fixed. Assume bounded costs, a reference uniformly bounded away from zero, a relative-interior feasible target, and the nondegeneracy conditions of Appendix G. Prices are represented on the load-difference span V. Define

$$\bar{g}(h) = \mathbb{E}_{q^\star(\cdot|h)} g(a), \quad K = \mathbb{E}_\mu \operatorname{Cov}_{q^\star}(g), \quad d = \mathbb{E}_\mu[z(H)\bar{g}(H)].$$

The vector d is the first-order load change obtained by applying the *unchanged* policy to the new state distribution. It lies in V because the component of $\bar{g}$ in $V^\perp$ is constant and z has mean zero. We take inverses on V throughout this section.

Theorem 13 (Distribution-induced capacity response)**.** *The optimal price and policy curves are differentiable at $t = 0$, with*

$$\dot{\alpha} = \tau K^{-1} d, \tag{62}$$

$$\dot{q}_a(h) = -q_a^\star(h)(g(a) - \bar{g}(h))^\top K^{-1} d. \tag{63}$$

Let $V(t)$ be the optimal regularized value under μ_t and the fixed load target. If Z_α is the normalizer in Equation (3), then

$$V'(0) = -\tau \mathbb{E}_\mu[z \log Z_\alpha], \tag{64}$$

$$V''(0) = \tau d^\top K^{-1} d \geq 0. \tag{65}$$

In particular, a state-distribution shift with $d = 0$ causes no first-order change in the optimal routing kernel and has zero second-order capacity cost along this linear measure path.

Proof. Write the capacity equation as

$$\mathcal{R}(\alpha, t) = \mathbb{E}_\mu[(1 + tz)\bar{g}_\alpha] - \rho = 0.$$

Differentiating the finite-action Gibbs kernel gives

$$\partial_\alpha \bar{g}_\alpha = -\tau^{-1} \operatorname{Cov}_{q^\alpha}(g).$$

At zero, $\partial_\alpha \mathcal{R} = -K/\tau$ is invertible on V and $\partial_t \mathcal{R} = d$. The implicit function theorem supplies the local price curve and $\dot{\alpha} = \tau K^{-1} d$. The policy derivative follows by differentiating its logarithm; its conditional mean is zero, as required by normalization. Boundedness allows differentiation under every expectation in these expressions.

The dual value can be written

$$V(t) = \sup_\alpha \left\{ -\tau \mathbb{E}_\mu[(1 + tz) \log Z_\alpha] - \alpha^\top \rho \right\}.$$

Stationarity cancels the derivative through α_t, giving Equation (64). Its derivative at zero uses $\partial_\alpha \log Z_\alpha = -\bar{g}_\alpha/\tau$ and is

$$\mathbb{E}_\mu[z\bar{g}^\top]\dot{\alpha} = \tau d^\top K^{-1} d.$$

This also exhibits convexity along a linear change of measure: the dual is a supremum of affine functions of t. Finally, $d = 0$ in the price and policy derivatives proves the last assertion. It does not imply that the linear value change $V'(0)$ vanishes, since the new states can have different prediction difficulty even when their expert demand is unchanged. □

Separating difficulty from competition. The first derivative in Equation (64) measures the change in reduced local free energy across the state population. The nonnegative second derivative measures the extra cost of redistributing capacity. States can become more difficult without competing differently for experts, or compete differently without a large change in their average native loss. These are distinct mechanisms and require different diagnostics in a multimodal or long-context audit.

J.2 Joint cost and distribution changes

Now let the costs also change as $C_{t,a} = C_a + te_a$, with bounded e. Retain μ_t from Equation (61), and define

$$\widetilde{e}_a = e_a - \mathbb{E}_{q^\star} e, \qquad b_{ge} = \mathbb{E}_\mu \operatorname{Cov}_{q^\star}(g, e), \qquad \widetilde{g}_a = g(a) - \bar{g}.$$

Both changes now enter the same load equation: the costs alter action preferences, while the new measure changes population demand. Solving that equation jointly gives the response below.

Theorem 14 (Cost–demand interaction law)**.** *Under the preceding boundedness and relative-interior assumptions,*

$$\dot{\alpha} = K^{-1}(\tau d - b_{ge}), \tag{66}$$

$$\frac{\dot{q}_a}{q_a^\star} = -\frac{1}{\tau}\left\{ \widetilde{e}_a + \widetilde{g}_a^\top K^{-1}(\tau d - b_{ge}) \right\}. \tag{67}$$

The optimal-value curvature along this joint linear path is

$$V''(0) = 2\mathbb{E}_\mu[z\mathbb{E}_{q^\star} e] - \frac{1}{\tau}\mathbb{E}_\mu \operatorname{Var}_{q^\star}(e) + \frac{1}{\tau}(\tau d - b_{ge})^\top K^{-1}(\tau d - b_{ge}). \tag{68}$$

The pure cost response is concave, the pure distribution response is convex, and the mixed response need have neither sign.

Proof. At fixed prices, differentiation through costs gives $\partial_t \bar{g} = -\operatorname{Cov}_{q^\star}(g, e)/\tau$. Differentiating the load equation therefore yields $0 = d - b_{ge}/\tau - K\dot{\alpha}/\tau$, proving Equation (66). Centering the derivative of $-(C_{t,a} + \alpha_t^\top g(a))/\tau$ gives Equation (67).

For the curvature, set

$$D(\alpha, t) = -\tau \mathbb{E}_\mu[(1 + tz) \log Z_{\alpha,t}] - \alpha^\top \rho.$$

At the reference point its partial derivatives are

$$D_{\alpha\alpha} = -K/\tau,$$
$$D_{t\alpha} = d - b_{ge}/\tau,$$
$$D_{tt} = 2\mathbb{E}_{\mu}[z\mathbb{E}_{q^\star} e] - \mathbb{E}_{\mu} \operatorname{Var}_{q^\star}(e)/\tau.$$

The envelope derivative along the stationary price curve is $V'' = D_{tt} + D_{t\alpha}\dot{\alpha}$. Substitution gives the displayed formula. If $d = z = 0$, it reduces to minus the projected residual energy from Appendix G, divided by τ. If $e = 0$, it reduces to Equation (65). The cross term can have either sign, proving the final distinction. □

The vector $\tau d - b_{ge}$ makes the mechanism concrete. A shift that increases demand for an expert can be offset by a cost change that makes that expert less attractive. Conversely, the two effects can reinforce one another. A stable observed load can therefore conceal changes in both compatibility and state composition. Neither an unchanged load histogram nor an unchanged average cost proves an unchanged routing mechanism.

J.3 A TWO-MODALITY SPECIALIZATION

Let $\mu_\vartheta = (1-\vartheta)\mu_T + \vartheta\mu_I$, for text and image states, and fix $0 < \vartheta_0 < 1$. Relative to μ_{ϑ_0}, the signed measure $\mu_I - \mu_T$ has a bounded density z: domination by the positive mixture bounds its positive and negative parts. Thus the previous theorem applies without requiring that the modalities have overlapping support.

In this case

$$d = \mathbb{E}_{\mu_I}\bar{g} - \mathbb{E}_{\mu_T}\bar{g}. \tag{69}$$

The price response is determined by the *difference in expert demand* between modalities, transformed by the inverse load susceptibility. Equation (65) becomes

$$\frac{d^2V}{d\vartheta^2}(\vartheta_0) = \tau\|\mathbb{E}_{\mu_I}\bar{g} - \mathbb{E}_{\mu_T}\bar{g}\|^2_{K^{-1}}.$$

This provides a quantity that a multimodal audit can estimate from a declared routing kernel without assigning linguistic categories to experts. It describes the frozen local transport problem; changing the number of image patches also changes the network's state distribution, so evaluating the complete model remains necessary for a model-level performance claim.

Binary example. For two experts write $q(h)$ for the probability of expert 1 and take the scalar load constraint $\mathbb{E}q = \rho_1$. Let $K = \mathbb{E}[q(1-q)]$, $d = \mathbb{E}_{\mu_I}q - \mathbb{E}_{\mu_T}q$, and $\alpha = \alpha_1 - \alpha_2$ be the scalar price difference. Then

$$\dot{\alpha} = \frac{\tau d}{K}, \qquad \dot{q}(h) = -q(h)(1-q(h))\frac{d}{K}, \qquad V'' = \frac{\tau d^2}{K}.$$

When both modalities demand experts equally, $d = 0$ even if their hidden representations are far apart. When their demands differ, the response is largest in uncertain states with large $q(1-q)$. The denominator shows that an almost deterministic router can require a large price adjustment to absorb a small change in modality proportions.

J.4 CAPACITY STIFFNESS AND THE HARD-ROUTING LIMIT

Under the smooth-interface assumptions of Appendix H, $K_\tau/\tau \to G_\partial$. Suppose also $d_\tau \to d_0$ on the gauge space. Then

$$\dot{\alpha}_\tau \longrightarrow G_\partial^{-1}d_0, \qquad V''_\tau \longrightarrow d_0^\top G_\partial^{-1}d_0. \tag{70}$$

Indeed $\tau K_\tau^{-1} = (K_\tau/\tau)^{-1} \to G_\partial^{-1}$ by continuity of matrix inversion on positive definite matrices. This identifies the same weighted interface matrix in two apparently different problems: fitting imperfect action costs and accommodating a different population of routed states. Cost errors enter through their boundary differences; population changes enter through their expert-demand vector.

The dependence on G_∂^{-1} also specifies a meaningful failure of smooth sensitivity. If the active boundary graph disconnects, some demand changes cannot be absorbed by an infinitesimal interface

displacement. Finite changes may require creating a new active interface. One should then solve the changed assignment problem instead of extrapolating the derivative outside the region where its nondegeneracy assumptions hold.

J.5 What transfers from calibration to a new distribution

Restoring population load does not by itself preserve a calibration guarantee. We therefore consider separately how the probability of harmful acceptance changes under a new sampling law. Let A be the event that a proposal is accepted and has negative true gain. If calibration under a distribution P gives $P(A) \le \delta$ and the test law satisfies $dQ/dP \le \kappa$, for a finite constant $\kappa \ge 1$, then

$$Q(A) = \int_A \frac{dQ}{dP}\, dP \le \kappa\delta. \tag{71}$$

Here P, Q describe the joint experiment, including the calibration sample. The same conclusion holds when only the new-group law changes and the calibration marginal remains fixed: the joint likelihood ratio is then the new-group likelihood ratio. No conditional coverage given a particular calibration sample is assumed. Alternatively $\mathrm{TV}(P, Q) \le \eta$ gives $Q(A) \le \delta + \eta$ by the definition of total variation. These elementary transfer bounds do not supply conditional safety among accepted proposals. Nor can a load response estimate by itself establish a bound on dQ/dP.

For a finite trajectory, suppose the conditional likelihood ratio at decision ℓ is bounded by κ_ℓ. The chain rule for path probabilities bounds the trajectory likelihood ratio by $\prod_\ell \kappa_\ell$. Applying Equation (71) with this product yields a valid but potentially loose coverage bound. This explains why a one-step acceptance guarantee must not be silently reused as a full-sequence guarantee. Group-level calibration of the entire declared sequence experiment, or a genuine distribution-shift bound, is needed to make that claim.

J.6 An audit that distinguishes the mechanisms

The results suggest three separate measurements at fixed model weights. First, compare modality-conditional expert demands and estimate d. Second, estimate the load susceptibility K from the same declared positive kernel, or the weighted boundary matrix G_∂ in a hard-routing analysis. Third, compare candidate predictive costs at a common state, which estimates e or a cost gap without conflating it with the change in population. Their combination predicts the price response through Equation (66).

These quantities serve different purposes. The norm of d measures competition for capacity; the spectrum of K measures how readily the router can redistribute that competition; the projected cost error measures decision-relevant misspecification. Reporting these quantities separately makes it possible to distinguish a change in demand from a change in cost estimation or in the ability to reallocate capacity.

K From Cost Estimation to Calibrated Execution

The preceding results specify what a correction must estimate and what it must keep fixed. We now put those requirements in execution order: choose the action family, fit its relative costs, estimate prices, and calibrate the resulting proposals. The protocol constructs a correction at frozen model weights; it does not identify the training objective of the native router.

K.1 A complete fitting and execution contract

1. Specify actions and split independent units. A complete action includes expert identities and the rule for combining their outputs. Let M be the maximum number of actions retained at a state, and k the number of experts executed by each action. The first controls the search budget; the second controls the sparse action executed. Candidate generation must use deployment-time information and include the native action. For top-k, evaluating one expert at a time is not equivalent to evaluating its complete joint action. Split documents, images, or independently generated states before collecting targets. Use separate fitting, pricing/development, calibration, and test units. A fitting set may

also supply development prices when no calibration or test unit enters either optimization. All hyperparameter selection ends before calibration. Candidate-family changes require a new proposal and therefore a fresh calibration, even when the old cost head is reused.

2. Estimate relative predictive costs. Collect forced continuation losses on fitting units only. For the image head, the inputs concatenate the pre-intervention hidden vector, all native router log probabilities, and the candidate dispatch vector. A one-hidden-layer GELU network of width 128 predicts the cost relative to the native action; normalization statistics are fitted on fitting inputs. No candidate expert output or target is an inference-time feature. Section E gives the optimizer and all iteration counts. Because common state offsets do not affect selection, centered cost error and candidate regret complement ordinary squared error. The finite-error theorem requires an error bound; minimizing an empirical regression loss alone does not establish that bound or a Lipschitz constant for the fitted function.

3. Estimate expert prices at the chosen target. For a fixed head, candidate rule, temperature, and load target, solve the convex development objective from Appendix M. A gradient step uses only soft candidate probabilities and dispatch vectors. Gauge-fix the prices on the load-difference span, and save the remaining load residual. For top-1 the problem is a transport matrix scaling problem. For joint actions, use the general moment dual: ordinary column scaling is not a substitute for arbitrary overlapping dispatch constraints. The Hessian is the conditional covariance of dispatch divided by temperature; a damped Newton solve can exploit it when the number of experts is small. This is the general linear-constraint extension of entropic transport, not an assertion that all such constraints reduce to balanced bipartite matching (Peyré & Cuturi, 2019; Liu et al., 2023b).

4. Calibrate, freeze, and execute. Compute residuals for the already fixed proposals on calibration groups. Use an image as one group for the image experiment; use a whole sequence when simultaneous sequence coverage is intended. A group maximum covers any accepted harmful proposal in that group. Larger groups increase the scope of the covered event but leave fewer independent calibration units at a fixed observation budget. At failure probability δ, at least $\lceil 1/\delta - 1 \rceil$ groups are needed for a finite order statistic. Even with enough groups for a finite threshold, the residual quantile determines how many proposals pass the calibrated gain threshold. Freeze the head, normalization, prices, candidate rule, and quantile. The batch filter can revert proposals but cannot replace them by a different uncalibrated action. The guarantee is marginal over the new group, and calibrated single-target losses have the observed-loss meaning specified in Appendix D.3.

K.2 Capacity accounting and computational cost

Let N_f, N_d, N_c be fitting, development, and calibration state counts; B the execution batch size; d_f the head input width; d_c its hidden width; and T_{cont} the cost of a forced continuation. Collecting supervision for M actions costs $O((N_f + N_c)MT_{\text{cont}})$ if calibration evaluates all actions. Evaluating only the fixed proposal and native action reduces calibration to two continuations per state. Fitting a one-hidden-layer head for I_f minibatches of size b_f costs $O(I_f b_f M d_f d_c)$; a shared multilayer head also includes its hidden-to-hidden matrix products. Development pricing takes $O(I_d N_d M k)$ arithmetic for I_d iterations when the k nonzero dispatch entries are used. These costs are paid offline; the inference-time action still executes its declared expert set.

At execution, head scoring costs $O(BMd_f d_c)$ for the image architecture, and reduced-cost selection costs $O(BMk)$. The literal threshold filter in the reference image implementation recomputes batch loads and has worst-case cost $O(B^2 k + B \log B + BE)$, where E is the number of experts. Its recorded timings include that implementation. An equivalent count-ledger implementation initializes all proposals, sorts their gains, and reverts equal-gain blocks in ascending order. Maintaining expert counts and the sum of absolute count deviations reduces filtering to $O(B \log B + Bk + E)$. Reverting one move changes at most $2k$ entries; update their absolute deviations before and after the move. Neither implementation assumes monotonicity of load TV in the threshold. The sensitivity code tests the ledger against literal recounting.

Normalized dispatch load is $\widehat{\rho}/k$. The image implementation stores g/k directly, so its prices are k times prices expressed against the unnormalized indicator g in the theorems. The reduced-cost scalar is identical after this rescaling. A TV tolerance allows small changes of aggregate load; it is not a

no-overflow guarantee. To enforce per-batch capacities b_i, check the integer counts against b_i at every retained threshold. Native dispatch must already satisfy those capacities for native fallback to certify termination. A batch allocation may be infeasible even if expected population loads satisfy the same average fractions.

K.3 Support and distribution changes

Top-k candidate counts can grow as $\binom{E}{k}$, before gate-weight variants are considered. We therefore report candidate-relative optima, not an exhaustive joint-action oracle when actions have been truncated. Candidate inclusion does not assert that omitted actions have high cost. Expanding a nested family improves the exact optimum at fixed feasible constraints, but can change estimation error and calibration, so learned-policy improvement is a separate measurement. The oracle references in the tables use measured test costs only for evaluation and do not constitute deployable policies.

Distribution shift introduces a second reason to revisit the fitted objects. To apply the response analysis, first distinguish a change in the conditional cost function from a change in the state measure. The price derivative in Appendix J fixes the conditional costs while perturbing that measure. With both changed, its joint perturbation formula includes both terms. A frozen price is not guaranteed to satisfy the new target. Updating it requires new development states and recalibration of the updated proposals. The bounded density-ratio results concern a specified shift class, not arbitrary autoregressive feedback. A correction at many layers is evaluated by a full forward under the resulting states, not by adding native-state local improvements.

K.4 Controlled sensitivity protocol

To test the choices in the execution protocol separately, we vary candidate count, predictor width, calibration group size, and distribution shift on the nonlinear eight-expert Gaussian prediction control. Each of 20 independent problems uses disjoint fitting, development, calibration, and test sets of 8,192, 2,048, 2,048, and 8,192 states. The latent state is $z \sim \mathcal{N}(0, I_4)$ and the hidden input is the 48-dimensional deterministic embedding specified in the exact-control protocol. Expert i predicts a unit-covariance Gaussian centered at c_i; the target mean is the same nonlinear map of z used in that protocol, with strength 0.65. Conditional risk is its exact squared-distance cost up to a shared constant. The native score is $z^\top c_i - \|c_i\|^2/2$, so candidate ordering uses an affine router, not measured test costs. All evaluated actions execute one expert.

The head has two SiLU hidden layers, width 32, 128, or 256, and eight outputs. It is trained for 600 steps with batch size 512 using centered Huber loss, AdamW, learning rate 0.0008, and weight decay 0.00001. The width-128 head is the reference. All widths use the same training examples and iteration budget. For candidate counts $M \in \{2, 4, 8\}$, retain nested native-score prefixes and reuse the head fitted to all eight action costs. This isolates inference-time support restriction, not a reduction in the offline cost-label budget. Development prices use temperature 0.2 and 400 damped scaling steps. The target is native development load. For each declared candidate family, freeze prices before calibrating the proposal-specific residual; no test cost determines a route.

Calibration groups contain $s \in \{1, 8, 32\}$ independently generated states, giving 2,048, 256, or 64 calibration groups at failure level 0.1. This changes simultaneous group coverage, not semantic strata. Test groups use the same construction. The one-factor grid has 13 settings, all reported in Table 11. The reported means describe the prespecified grid, not a test-selected tuning recommendation. The full JSON records calibration thresholds, parameter counts, residual loads, and pre- and post-filter harm rates.

Support, capacity, and grouping. The reference correction recovers 0.050 nats per state. Restricting to two candidates recovers 0.039; four candidates recover 0.052. The measured recovery is therefore not monotone in candidate count, even though the exact optimum over nested feasible families is monotone. Across head widths, observed group harm is 0.3%–1.5% at the nominal 10% level. Increasing group size from 1 to 32 retains a similar mean recovery while covering a larger simultaneous event and reducing the fraction of retained changes. These controls vary estimator capacity and coverage scope without changing the downstream prediction task.

Setting	Gain	Retained	Harm	Soft TV
Reference: $M = 8$, width 128, $s = 1$	0.050	4.46%	0.41%	0.022
$M = 2$	0.039	4.53%	0.43%	0.023
$M = 4$	0.052	4.69%	0.38%	0.022
Head width 32	0.055	5.94%	1.54%	0.022
Head width 256	0.057	5.73%	0.27%	0.022
Group size $s = 8$	0.050	4.34%	0.62%	0.022
Group size $s = 32$	0.050	4.09%	1.04%	0.022
Shift 0.5, frozen	0.040	2.10%	0.41%	0.141
Shift 1.0, frozen	0.035	1.74%	0.46%	0.265
Shift 0.5, refresh target	0.050	4.53%	0.41%	0.021
Shift 1.0, refresh target	0.054	5.29%	0.49%	0.020
Shift 0.5, retain target	0.022	1.21%	0.35%	0.023
Shift 1.0, retain target	0.016	1.06%	0.30%	0.023

Table 11: Controlled one-factor sensitivity over 20 independent problems. Gain is raw conditional-risk reduction after the hard load filter, in nats per state. Retained is the fraction of all test states with a changed action after both filters. Harm is the fraction of test groups with an accepted negative reduced-risk gain before the load filter; its nominal bound is 10% for exchangeable groups. Soft TV compares the unfiltered soft policy with its declared development target. All hard policies have TV at most 0.010 relative to current native dispatch.

Distribution and target load are separate choices. A shift adds 0.5 or 1.0 to the first latent Gaussian coordinate, holding expert predictors and the conditional target map fixed. The head is never retrained on shifted examples. Frozen deployment retains the original prices and quantile. Two refresh variants use fresh shifted development and calibration samples: one retains the original load target, and the other adopts shifted native load. With a unit shift, original-target soft-load TV is 0.265 under frozen prices and 0.023 after refitting prices at that same target. The reduction is consistent with the role of repricing in the distribution-response calculation. Adopting shifted native load defines a different capacity objective and is reported separately, not used as evidence of fixed-target feasibility. In every variant the hard filter remains relative to current native batch load, exactly as in Algorithm 1. The shifted frozen-quantile rows test transfer empirically; exchangeability does not provide their nominal coverage guarantee.

L Additional Datasets, Architectures, and Routing Controls

The main experiments separate proximity to a routing boundary from the predictive effect of crossing it. We now test whether this distinction persists across datasets, architectures, layers, and token positions. All comparisons intervene at frozen weights and measure observed-label loss changes after a specified expert exchange.

L.1 Image datasets and architectures

We first extend the original image model to CIFAR-100, keeping the architecture and training protocol fixed except for the classifier's 100 outputs. Across five independent checkpoints, the nearest rank-4/rank-5 exchange gives Spearman correlations of -0.040 for raw margin and 0.638 for Fisher displacement with absolute NLL change, compared with -0.023 and 0.707 on CIFAR-10. Small margin has harmful-exchange AUROC 0.502 on CIFAR-100 and 0.499 on CIFAR-10, where harm means an NLL increase above 0.1 nat.

To test structural dependence, we also train Conv-MoE and ViT-MoE on both datasets. Both models have four blocks, width 96, and eight experts with top-2 routing. Conv-MoE uses convolutional downsampling and spatial mixing; ViT-MoE uses 8×8 patches and four-head self-attention. Each condition uses seeds 9100–9104, 40,000 training images, 12 epochs, batch size 256, AdamW with learning rate 8×10^{-4} and weight decay 0.01, gradient clipping at 1, and load-balancing coefficient 0.01. Normalization uses fixed CIFAR-10 channel constants, with no augmentation or test-driven checkpoint selection. These are compact diagnostic models; mean test accuracies are $60.14\%/29.24\%$ for Conv-MoE and $54.90\%/28.16\%$ for ViT-MoE on CIFAR-10/100.

For each checkpoint, we audit all four layers at positions 0, 5, 10, and 15 on all 10,000 test images. Each intervention replaces the second-ranked expert with the third-ranked expert and recomputes the continuation. The resulting 64 architecture/dataset/layer/position settings contain 3.2 million audits across seeds. Statistical replication remains at the checkpoint level.

Model	Data	Raw margin	Prob. margin	Fisher	Margin AUROC
Conv	CIFAR-10	−0.028–0.009	0.032–0.077	0.773–0.786	0.498–0.507
Conv	CIFAR-100	−0.008–0.022	0.095–0.258	0.625–0.681	0.497–0.514
ViT	CIFAR-10	−0.009–0.016	0.081–0.138	0.742–0.785	0.497–0.514
ViT	CIFAR-100	−0.011–0.034	0.148–0.253	0.664–0.738	0.500–0.521

Table 12: Ranges of five-seed means over 16 layer/position settings per architecture and dataset; these ranges are not confidence intervals. Correlations are with absolute observed-label NLL change. The last column evaluates an independently calibrated raw-margin risk score.

Testing a nonlinear margin predictor. Low rank correlation alone can miss a nonmonotonic relationship. We therefore use 5,000 held-out training-set images to fit a fixed, Laplace-smoothed risk score in 20 margin-quantile bins, then evaluate it only on the official test set. Its AUROC is 0.497–0.521 across the 64 settings (Table 12). Within narrow margin bins, Fisher displacement retains correlations of 0.624–0.786 with loss magnitude. Restricting to initially correct predictions gives Fisher correlations of 0.832–0.953; in the lowest margin decile, with the cutoff set on calibration data, they are 0.625–0.790.

Raw margin still describes switching propensity: under independent Gaussian gate-logit noise of standard deviation 0.1, its negative correlates 0.829–0.918 with switching probability. This control uses 1,024 fixed test images and 64 noise draws per image and setting. Together, these results distinguish predicting a switch from predicting its loss consequence.

Alternative definitions. The probability margin contains additional information about mixture weights:

$$p_{(k)} - p_{(k+1)} = p_{(k)}(1 - e^{-m}),$$

where m is the raw logit margin and p is the unit-temperature softmax over all expert logits. Its correlations reach 0.258 in these experiments, so the evidence supports distinguishing margin definitions. Across harm thresholds $0.01, 0.05, 0.1, 0.2, 0.5$ nat, calibrated raw-margin AUROC ranges respectively over 0.495–0.507, 0.496–0.514, 0.497–0.521, 0.489–0.542, and 0.478–0.539. The last threshold retains 57 settings with defined AUROC in all five seeds; the others retain all 64.

L.2 Additional MoE-LLaVA language backbones

We next audit MoE-LLaVA-StableLM-1.6B-4e and MoE-LLaVA-Qwen-1.8B-4e (Lin et al., 2026). These extend the language backbone within the same model family. Each uses the same 256 AI2D questions and 256 POPE-random questions; the latter are based on COCO val2014 images (Li et al., 2023). The independent units are 209 AI2D image clusters and 213 POPE image clusters. Each model/dataset run audits three layers and two position types, producing 1,536 groups and 9,216 candidate forwards over all six expert pairs. The primary comparison keeps the highest-ranked expert and exchanges the second for the third.

Backbone	Data	Position	Margin correlation	Fisher correlation
StableLM	AI2D	Image	−0.048 [-0.119, 0.023]	0.908 [0.876, 0.934]
StableLM	AI2D	Answer	−0.041 [-0.107, 0.030]	0.859 [0.815, 0.895]
StableLM	POPE	Image	−0.048 [-0.118, 0.026]	0.974 [0.966, 0.981]
StableLM	POPE	Answer	0.068 [-0.005, 0.136]	0.966 [0.955, 0.974]
Qwen	AI2D	Image	0.010 [-0.059, 0.083]	0.933 [0.908, 0.953]
Qwen	AI2D	Answer	0.019 [-0.048, 0.083]	0.927 [0.901, 0.951]
Qwen	POPE	Image	−0.021 [-0.082, 0.044]	0.984 [0.978, 0.989]
Qwen	POPE	Answer	0.014 [-0.047, 0.085]	0.980 [0.973, 0.986]

Table 13: Nearest-expert exchanges in additional MoE-LLaVA backbones. Spearman correlations with absolute NLL change and descriptive 95% intervals from 1,000 image-cluster bootstrap replicates.

Fisher displacement correlates more strongly with loss magnitude in all eight comparisons (Table 13); the paired bootstrap intervals for Fisher correlation minus absolute margin correlation have positive lower bounds. These are descriptive intervals without multiplicity correction. All 6,144 native-action identity checks have zero loss error. POPE uses the leading-space answer tokens `Yes`/ `No`; an initial lowercase-token run is retained as a format check and is excluded from this table. Restricted-choice accuracies are $50.00\%/91.02\%$ for StableLM and $48.83\%/93.36\%$ for Qwen on AI2D/POPE.

L.3 An independent multimodal family and broader interventions

DeepSeek-VL2-Tiny (Wu et al., 2024) adds a separate model family with a SigLIP visual encoder, dynamic tiling, and a 12-layer sparse decoder. Each sparse block selects six of 64 routed experts and also uses two shared experts. We retain its softmax gate, unnormalized selected weights, and shared-expert contribution. The intervention exchanges rank 6 for rank 7.

For AI2D and POPE random, we audit layers 1, 3, 6, 9, and 11 at three global-image patch positions (49, 98, 147) and the answer-prediction position, with 256 questions per dataset. POPE popular and adversarial each add 256 questions at layers 3, 6, and 9 and two position types. These POPE variants change the absent-object question distribution on COCO images. Native prompts and answer tokens are fixed throughout.

Data	Position	Margin correlation	Fisher correlation
AI2D	Image	0.078 [0.041, 0.115]	0.978 [0.969, 0.984]
AI2D	Answer	−0.001 [-0.056, 0.050]	0.962 [0.945, 0.974]
POPE random	Image	0.009 [-0.026, 0.042]	0.972 [0.967, 0.977]
POPE random	Answer	0.061 [0.007, 0.117]	0.975 [0.968, 0.981]
POPE popular	Image	0.077 [-0.001, 0.148]	0.974 [0.967, 0.979]
POPE popular	Answer	0.079 [0.005, 0.149]	0.973 [0.964, 0.980]
POPE adversarial	Image	0.051 [-0.023, 0.125]	0.976 [0.968, 0.981]
POPE adversarial	Answer	0.109 [0.032, 0.183]	0.974 [0.966, 0.980]

Table 14: DeepSeek-VL2-Tiny with numerically isolated token interventions. Correlations with absolute NLL change and image-cluster bootstrap 95% intervals. Final-layer image interventions are structural zero controls and are excluded from correlation summaries.

Preserving intervention locality. Changing expert group sizes can alter BF16 arithmetic at non-target tokens. We therefore cache the native output of the intervened MoE, replace only the target token's output, and recompute all subsequent layers. This correction preserves the intended single-token intervention. All 13,312 native-action identity checks pass, and all 1,536 final-layer image controls have exactly zero loss change and Fisher displacement, as expected when no subsequent attention connects those tokens to the answer. Table 14 uses these corrected runs. In this broader audit, Fisher correlations remain 0.962–0.978, while raw-margin correlations range from -0.001 to 0.109.

These extensions support the distinction between boundary proximity and the predictive cost of an exchange across the tested architectures and datasets. Fisher displacement is computed from candidate continuations and serves here as an offline diagnostic of that distinction.

L.4 Computation and saved evidence

The exact-cost experiment in Appendix E proceeds from latent-state sampling to known expert costs, constrained assignment, router and cost-head fitting, and held-out evaluation. It saves policy metrics, tube masses, higher-order tie statistics, stochastic-path costs, and a per-problem summary. The planar plot uses the same problem generator and price solver, making the reference partition and router partition directly comparable.

The accompanying local experiment files separate generation from analysis. The toy reproduction directory contains the benchmark runner, aggregate analysis, planar plotting script, and explicit paper-scale arguments. The dataset/backbone extension retains per-state audits and image-cluster bootstrap summaries; the broader architecture audit retains per-seed, per-layer, and per-position statistics, including the corrected local-intervention records. Tables 12–14 are generated from these saved summaries.

M Correcting Costly Assignments

The analysis suggests correcting assignments by their predicted gains while accounting for expert load. Forced continuation evaluations supply training costs for a surrogate, and transport supplies the capacity prices; geometry helps choose informative regions to audit. Let $\mathcal{A}_M(h)$ contain at most M actions, including native action a_0, and let $\widehat{C}$ and $\lambda \in \mathbb{R}^E$ denote estimated costs and fitted prices. With positive native proposal probabilities p_a and prior strength $\beta \geq 0$, form

$$\widehat{q}_a(h) \propto p_a(h)^\beta \exp[-\{\widehat{C}_a(h) + \lambda^\top g(a)\}/\tau]. \tag{72}$$

The most probable action minimizes $\widehat{C}_a + \lambda^\top g(a) - \tau\beta \log p_a$, so the prior changes the objective unless its contribution is constant. For the guarantee below, we use the prior-free hard proposal and account for uncertainty in its estimated reduced-cost gain before accepting a correction.

Proposition 7 (Calibrated sparse correction)**.** *Write $D_a^\lambda = C_a + \lambda^\top g(a)$ and $\widehat{D}_a = \widehat{C}_a + \lambda^\top g(a)$ at the same fitted prices. If all relative errors to native a_0 on $\mathcal{A}_M$ are at most ε, the estimated minimum has candidate regret at most 2ε; predicted reduced-cost gain above ε is non-worsening on this event. Alternatively, freeze the head, prices, and proposal a_1 before calibrating gain residuals, with $G = D_{a_0}^\lambda - D_{a_1}^\lambda$ and $\widehat{G} = \widehat{D}_{a_0} - \widehat{D}_{a_1}$. Use group maxima $R_j = \max_{h \in j}(\widehat{G} - G)$ for n exchangeable groups and failure level $\delta \in (0,1)$. For $u_\delta = R_{(\lceil (n+1)(1-\delta) \rceil)}$, accepting only when $\widehat{G} > u_\delta$ gives*

$$\Pr\{\textit{accepted proposal has } G < 0\} \leq \delta. \tag{73}$$

A filter that only reverts proposals preserves the same marginal acceptance-and-harm guarantee.

The bound controls the event that a new exchangeable group contains an accepted harmful proposal; it is not a bound conditional on acceptance. Set $u_\delta = +\infty$ if the required rank exceeds n (Angelopoulos & Bates, 2023; Romano et al., 2019). Calibrating with observed losses ℓ_a covers realized loss instead of unobserved conditional risk; Appendix D.3 explains the distinction and a noisy-risk extension. Algorithm 1 then raises the gain threshold until the retained proposals satisfy the batch load tolerance. Since this filter only reverts proposals to native actions, it preserves the acceptance-and-harm guarantee; matching loads also cancels the aggregate price charge.

Fitting prices and choosing candidates. For a fixed head and candidate support, minimizing $\tau\mathbb{E} \log Z_\lambda + \lambda^\top \rho$ gives the update $\lambda \leftarrow \lambda + \eta(\widehat{\rho}_{\widehat{q}} - \rho)$ with step size $\eta > 0$. Its Hessian $\tau^{-1}\mathbb{E}\,\mathrm{Cov}_{\widehat{q}}(g)$ supplies a step-size bound and identifies unobservable price directions. Prices fitted on development states are frozen for calibration and testing; a new target distribution requires refitting and recalibration. Expanding nested candidate sets cannot worsen the exact feasible optimum, but need not improve an estimated hard policy. Appendix D bounds truncation, prior, estimation, and load effects.

Algorithm 1 Calibrated, load-aware compatibility correction

Require: Frozen head $\widehat{C}$, prices λ, candidate rule $\mathcal{A}_M$
Require: Quantile u_δ; load tolerance $\xi \geq 0$; native batch actions a_0
for each state h in the batch **do**
  $\widehat{D}_a \leftarrow \widehat{C}_a(h) + \lambda^\top g(a)$ for $a \in \mathcal{A}_M(h)$
  $a_1 \leftarrow \arg\min_a \widehat{D}_a$; $v(h) \leftarrow \widehat{D}_{a_0} - \widehat{D}_{a_1} - u_\delta$
end for
Sort the distinct positive values of v, and append $+\infty$
for thresholds t in ascending order, starting at 0 **do**
  $a_t(h) \leftarrow a_1(h)$ if $v(h) > t$, otherwise $a_0(h)$
  if $\mathrm{TV}(\widehat{\rho}(a_t), \widehat{\rho}(a_0)) \leq \xi$ **then**
    return a_t ▷ first feasible retained proposal set
  end if
end for

Execution cost and realized load. Offline supervision requires M continuation evaluations per audited state. Online, a width-d_c head with input width d_f costs $O(Md_fd_c)$, followed by the original k expert evaluations. It uses neither candidate expert outputs nor target labels, Fisher computations, or forced continuations. After hard selection in a batch of B states, the realized average load is $\widehat{\rho} = B^{-1}\sum_h g(a(h))$. Algorithm 1 checks total variation $\|\widehat{\rho} - \widehat{\rho}_0\|_1/(2k)$ relative to native dispatch; individual hard-capacity limits require a separate check of each expert's realized batch load.

Candidate count controls audit cost, not active expert count: the sequence test uses $M = 8$ single-expert actions; the image test uses $M = 5$ joint actions from eight experts. Both evaluate complete continuations offline. Appendix K details fitting, calibration groups, execution, and batch-filter cost; Appendix K.4 varies support, head width, and calibration with and without distribution shift.

N Interpreting existing routing methods

The preceding results compare complete actions at a fixed network interface. This appendix uses the same comparison to examine three components of existing routers: capacity coordination, continuous output construction, and gradient estimation. Each connection requires matching the implemented operation to the quantity being evaluated.

N.1 Capacity coordination and learned scores

BASE maximizes learned affinity under balanced training assignments (Lewis et al., 2021); sparse OT constrains an affinity-based transport plan (Liu et al., 2023b); Loss-Free Balancing adjusts expert-specific selection biases using recent loads (Wang et al., 2024). These methods learn task preferences and coordinate expert use. The COT comparison evaluates how their scores differ from the costs of the complete actions being executed.

For example, two tokens with expert costs $(0, 100)$ and $(0, 1)$ must split one slot at each of two experts. Both assignments are balanced, but their total costs are 1 and 100. The better assignment gives expert 1 to the token with the larger fallback cost. A uniform expert bias can regulate demand, while the scores still determine which token receives the slot.

Equation (7) identifies score errors that do not change this constrained decision. For arbitrary errors $e = \widehat{C} - C$, suppose $\widehat{q}$ is a δ-approximate minimizer of $\mathcal{J}_{\widehat{C}}$ over the same feasible set as the true-cost minimizer $q^\star$. Then

$$0 \leq \mathcal{J}_C(\widehat{q}) - \mathcal{J}_C(q^\star) \leq \delta + \inf_v \mathbb{E}_\mu \operatorname{osc}_a\{e_a - v^\top g(a)\}, \tag{74}$$

where $\operatorname{osc}_a r_a = \max_a r_a - \min_a r_a$. Indeed, subtracting the two objectives bounds the excess by $\delta + \mathbb{E}\sum_a (q^\star_a - \widehat{q}_a)e_a$. Equal loads remove $v^\top g(a)$, and the difference of two averages is at most the range. For unequal loads, assume separately that $\mathcal{J}_{\widehat{C}}(\widehat{q}) \leq \mathcal{J}_{\widehat{C}}(q^\star) + \delta$. The upper bound then retains $-v^\top\{\rho(\widehat{q}) - \rho(q^\star)\}$ before taking the infimum over v; the nonnegative lower bound need not

hold for an infeasible policy. At zero temperature the bound compares predictive risk directly; at positive temperature it compares regularized objectives.

The optimization premise is method-specific. BASE's training assignment and greedy inference have different constraints. Sparse OT uses support for dispatch and weights for output; equal transported mass need not imply equal call counts. Loss-Free Balancing certifies optimality at its realized load.

Proposition 8 (Biased selection at its realized load)**.** *Fix the state law μ, integrable router scores $s_i(h)$, and an expert-bias vector $\zeta \in \mathbb{R}^E$ shared across states. Let $\mathcal{A}$ contain all k-expert sets, with binary dispatch $g_i(a) = \mathbf{1}\{i \in S_a\}$. Let q_{LF} select a top-k set of $s_i(h) + \zeta_i$, with any measurable tie rule and no subsequent dropping or reassignment. For any fixed scale $\kappa > 0$, define*

$$\widehat{C}_a(h) = -\kappa \sum_{i \in S_a} s_i(h), \qquad \rho_{\mathrm{LF}} = \rho(q_{\mathrm{LF}}).$$

Then q_{LF} minimizes $\mathbb{E}_{\mu q}\widehat{C}$ over $\rho(q) = \rho_{\mathrm{LF}}$. For integrable complete-action costs C, let q_C minimize L at that same load and put $e = \widehat{C} - C$. Then

$$0 \le L(q_{\mathrm{LF}}) - L(q_C) \le \inf_v \mathbb{E}_\mu \operatorname{osc}_a\{e_a - v^\top g(a)\}.$$

Proof. Top-k selection minimizes $\widehat{C}_a - \kappa\zeta^\top g(a)$ at each state. Integrating its pointwise comparison with any q gives

$$\mathbb{E}_{\mu q_{\mathrm{LF}}}\widehat{C} - \mathbb{E}_{\mu q}\widehat{C} \le \kappa\zeta^\top\{\rho_{\mathrm{LF}} - \rho(q)\}.$$

Equal loads cancel the right side. Equation (74) at $\tau = \delta = 0$ then gives the risk bound. □

Before cancellation, for every comparator q and vector v,

$$L(q_{\mathrm{LF}}) - L(q) \le \mathbb{E}_\mu \operatorname{osc}_a\{e_a - v^\top g(a)\} + (\kappa\zeta - v)^\top\{\rho_{\mathrm{LF}} - \rho(q)\}.$$

Substitute $C = \widehat{C} - e$ and remove $v^\top g$ from the error. The load-price term remains; the comparison need not be nonnegative.

Native biases alter selection; unshifted scores weight outputs (Wang et al., 2024). Thus C_a must use those weights and the complete frozen continuation. For simultaneous token interventions, use complete group assignments as actions and sum scores and dispatch over tokens. Cancellation still holds without additive continuation costs, but tokenwise bounds need not control group errors.

The certificate requires no bias-update convergence. Reaching prescribed loads is a separate primal–dual question (Han & Zhong, 2026). This connection verifies surrogate optimality, but neither bounds the score–cost residual nor compares joint-training trajectories, along which scores, experts, and visited states evolve.

N.2 Continuous outputs and predictive loss

Write $\ell_h(z)$ for the conditional task loss after passing interface output z through the entire frozen continuation. For two complete outputs z_a, z_b, let $\phi(s) = \ell_h(z_a + s(z_b - z_a))$. The difference between sampling an endpoint and executing their average is

$$I_h(\theta) = (1-\theta)\phi(0) + \theta\phi(1) - \phi(\theta) = \int_0^1 \min(s,\theta)\{1 - \max(s,\theta)\}\phi''(s)\,ds. \tag{75}$$

For $\phi \in C^2$, integration by parts on $[0,\theta]$ and $[\theta,1]$ proves the identity. Its kernel is nonnegative. Convexity along the complete output path therefore favors mixing over sampling at the same probabilities. The distinction is classical (Jacobs et al., 1991); Equation (12) and the nonlinear pullback in Appendix C describe the relevant predictive curvature. Convexity in final logits alone does not establish convexity along an intermediate output path.

The effect across a narrow boundary. Let u be a router gap on a compact regular two-sided interface $\Gamma = \{u = 0\}$, with the hard rule choosing b for $u > 0$ and a otherwise. In $|u| < \epsilon$, use an affine mixture with weight $p_y(u/\epsilon)$, where $y \in \Gamma$ labels a normal chart, $p_y(-v) = 1 - p_y(v)$, and $p_y(-1) = 0$, $p_y(1) = 1$. Assume finite boundary measure $d\omega = f\|\nabla u\|^{-1} d\mathcal{H}^{r-1}$ and normally Lipschitz chart density and complete-path losses, uniformly over the mixture weights. For fixed profiles and endpoints, put

$$B = \int_\Gamma \int_{-1}^{1} I_y(p_y(v))\, dv\, d\omega(y).$$

If R_0 is hard-rule risk, R_{sel} uses random endpoint selection, and R_{mix} executes their affine mixture, then

$$R_{\mathrm{sel}} - R_0 = O(\epsilon^2), \qquad R_{\mathrm{mix}} - R_0 = -\epsilon B + O(\epsilon^2). \tag{76}$$

To prove the first relation, integrate $(p_y(u/\epsilon) - \mathbf{1}\{u > 0\})(C_b - C_a)$ in normal coordinates. Symmetry cancels its constant boundary trace, and normal variation contributes $O(\epsilon^2)$. Subtracting $I_h(p_y(u/\epsilon))$ gives the second relation after the substitution $u = \epsilon v$. This is a signed risk comparison; the absolute magnitude in Equation (9) need not have the same cancellation. The auxiliary endpoint policy also changes load by $O(\epsilon^2)$, while actual mixed execution may add expert calls throughout the band. For a native gate that follows another output path, its loss difference from the affine path must be included before applying the risk conclusion.

SmoothSMoE constructs a transition by adding experts below the selection threshold and renormalizing the retained weights (Huu et al., 2026). For an isolated top-1 exchange, its limiting boundary path is an affine mixture of the two hard outputs. For top-k, the same normalization also changes common-expert weights. Let Q be their total unnormalized weight, V their averaged output, and r the tied weight of each exchanging expert. With entering activation ψ, put $T = 1 + \psi$ and let $E_p = (e_{\mathrm{selected}} + \psi e_{\mathrm{entering}})/(1 + \psi)$ average the selected and entering expert outputs. The difference between the smoothed boundary output and the corresponding chord of hard outputs is

$$\frac{QV + rTE_p}{Q + rT} - \frac{QV + rE_p}{Q + r} = \frac{Qr(T - 1)}{(Q + rT)(Q + r)}(E_p - V). \tag{77}$$

This follows by taking a common denominator. Thus an explanation based on endpoint interpolation must also account for the common-expert term.

A convex-loss example makes the distinction explicit. Take $Q = 2$, $V = 0$, $r = 1$, and both exchanging expert outputs equal to 1. Both hard boundary outputs are $1/3$, with zero squared loss for target $1/3$. The smoothed limiting output is $(1 + \psi)/(3 + \psi)$ and reaches $1/2$ at $\psi = 1$. Its positive loss persists on an interval of scaled gaps. Under positive regular boundary density, this produces a positive order-ϵ risk change across a width-ϵ transition. All positive contributions are assumed to be executed; truncation or overflow changes this forward comparison. Coexecution also increases expert calls, which must be counted separately from the load of an auxiliary endpoint-sampling policy.

ReMoE uses unnormalized ReLU gates (Wang et al., 2025). A newly active expert contributes $[s]_+ e$ to the other experts' output z_-. The contribution vanishes at activation. If $\|e\| \le M$ and the complete loss is L-Lipschitz along this output path, removing contributions only for $0 < s < \epsilon$ changes their pointwise losses by at most LMs. Coarea on a compact regular interface with normally Lipschitz density therefore gives

$$|R_{\mathrm{native}} - R_{\mathrm{removed}}| \le \frac{LM\epsilon^2}{2} \int_{\{s=0\}} \frac{f}{\|\nabla s\|}\, d\mathcal{H}^{r-1} + O(\epsilon^3). \tag{78}$$

The saved-call count is $O(\epsilon)$, with a nonzero linear coefficient when the boundary measure has positive mass. Small output weight controls predictive impact under the stated sensitivity bound, but does not make an expert evaluation proportionally cheaper. At a fixed state, differentiability gives $\ell_h(z_- + se) - \ell_h(z_-) = s\langle \nabla \ell_h(z_-), e\rangle + o(s)$, so a negative directional derivative lowers conditional loss for sufficiently small positive s.

N.3 Sparse gradients and finite action differences

SparseMixer and GRIN approximate selection-gradient terms using a small number of derivative evaluations (Liu et al., 2023a; 2024a). The gradient target must be specified first. At a frozen

single-expert interface with locally fixed sampling support, restrict the router parameters to a region where every active branch loss is differentiable. Consider

$$\mathcal{F}(\theta) = \sum_i p_i(\theta)\ell_h(p_i(\theta)f_i), \qquad z_i = p_i f_i, \quad \phi_i(s) = \ell_h(sz_i), \quad u_i = \nabla_\theta p_i. \tag{79}$$

The probabilities sum to one and also weight the selected output. Experts and the continuation are fixed; their output amplitudes vary with the router parameters. The path from zero removes the selected expert contribution while retaining the residual context. Differentiating the finite sum gives

$$\nabla\mathcal{F} = A + B, \qquad A = \sum_i [\phi_i(1) - \phi_i(0)]u_i, \qquad B = \sum_i \phi_i'(1)u_i. \tag{80}$$

The common baseline cancels because $\sum_i u_i = 0$. The first term changes selection probabilities, and the second changes the selected output's amplitude.

On a smooth path, $\phi_i(1) - \phi_i(0) = \int_0^1 \phi_i'(s)\,ds$. Endpoint, midpoint and Heun approximations have errors at most $M_{2i}/2$, $M_{3i}/24$ and $M_{4i}/216$, respectively, where $M_{ji} = \sup_{[0,1]} |\phi_i^{(j)}|$ under the indicated differentiability. The Heun rule is $\phi_i'(1)/4 + 3\phi_i'(1/3)/4$. These are standard quadrature bounds: Taylor's integral remainder gives the endpoint and midpoint constants; integrating the absolute Peano kernel of the two-node rule gives $1/216$.

This calculation also explains why the ordinary branch gradient can already carry useful selection information. It gives $\|A - B\| \le R = \frac{1}{2}\sum_i M_{2i}\|u_i\|$. For $v = A + B$,

$$\langle v, A\rangle,\ \langle v, B\rangle \ge \tfrac{1}{2}\|v\|(\|v\| - R).$$

Thus $-A$ and $-B$ both descend the frozen objective when $\|v\| > R$. Affine path losses make the two terms identical. This is a local statement about the specified objective; it does not establish a training advantage from evaluating additional routes.

For a piecewise C^1 continuation with finitely many interior loss jumps $J_j = \phi(s_j^+) - \phi(s_j^-)$, finite one-sided traces, integrable derivatives and no endpoint jumps,

$$\phi(1) - \phi(0) = \int_0^1 \phi'(s)\,ds + \sum_j J_j. \tag{81}$$

Applying the fundamental theorem of calculus on each smooth interval proves this identity. The endpoint condition uses the actual common loss at zero contribution; choosing a different one-sided baseline on each ray would not justify the cancellation in Equation (80). Even a continuous path may have slope jumps, which also enter the quadrature error. For example, $\phi(s) = 2 + s - 2\mathbf{1}\{s > 3/5\}$ has slope one on both pieces and total change -1. Within-branch derivatives miss the value jump.

For the inspected single-expert implementations, SparseMixer's mean router-gradient estimate over expert and node sampling is B plus an approximation to A. GRIN's SparseMixer-v2 estimates the selection term of a frozen-amplitude surrogate; v2* includes both terms, using each variant's own probabilities. All three use the endpoint rule on the argmax branch, where the $M_2/2$ error bound remains; the Heun bound alone therefore does not bound the whole native estimator. Hence an accurate selection estimate alone need not descend Equation (79) when amplitudes are recomputed. Extending this comparison to joint top-k sampling requires its complete sampling law and action losses, since sequential choices can interact.

N.4 Existing diagnostics and their scope

The completed route-signal diagnostics use six checkpoints from two seeds of a six-layer, width-352, eight-expert, top-2 model. Each checkpoint and layer uses 1,024 training probes, eight independently reset update groups, and 16,384 held-out validation tokens. The primary router-parameter displacement is 0.001 times its original norm. Experts remain fixed, and validation recomputes the full network after each update. Parameter displacement is matched; complete-route evaluation requires additional computation.

For native-minus-swapped observed loss Δ and the smooth score-gradient component c along the rank-2/rank-3 exchange, informative probes satisfy $|\Delta| > 10^{-4}$ and $|c| > 10^{-8}$. Loss-weighted disagreement is $\sum |\Delta|\mathbf{1}\{c\Delta < 0\} / \sum |\Delta|$ over all selected probes.

At the final MoE layer, endpoint comparisons and isolated pair-exchange derivatives agree in sign on 96.38–97.56% of informative probes; loss-weighted disagreement is 0.37–1.18%. At layer 3 it is 3.87–9.90%, falling to 0.57–2.30% when downstream expert IDs are fixed on the same probes while states and mixture weights are recomputed. These exchange paths differ from SparseMixer's zero-to-output path. All six primary endpoint-versus-pairwise-smooth uncertainty intervals include zero at each layer. The intervals resample update groups and validation documents conditional on a checkpoint. The results support a contribution from downstream support changes, while leaving a consistent benefit from additional endpoint supervision unestablished.