WHICH COMMUNICATION ROUTES HELP A GNN? A GRAPHON THEORY OF TASK-CONDITIONED EDGE SENSITIVITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Which communication routes in a graph neural network actually help its task, and does the answer remain meaningful when the same population is represented by graphs of different sizes or by different random samples? On a single finite graph, this question can be answered locally by differentiating the task loss with respect to edge weights. Across graphs, however, those derivatives live in different coordinates and need not be directly comparable.

We use graphons to make this question tractable at the population level. A graphon provides common coordinates for communication in large dense graphs, allowing the local value of changing a route to be represented by a signed task-conditioned sensitivity field. For nonlinear residual message passing, we derive this field exactly: each route is valued through the interaction between the representation difference available to cross it and a backward signal describing what the task needs. We prove that the complete field is stable under graphon cut convergence. In particular, dense Bernoulli graphs may remain noisy edge by edge in raw L^2, while their sensitivities converge in expected L^1 at rate $O(n^{-1/2})$ to those of the corresponding mean communication kernel. Under stronger aligned L^2 assumptions, finite communication updates also track a common population trajectory over fixed horizons.

Experiments separate structural stability from operational usefulness. Synthetic and CiteSeer-calibrated studies show stable sensitivities despite persistent edge noise, and a fixed-forward diagnostic shows that changing the task alone can change communication preferences. A controlled routing problem shows that route-specific forward–backward alignment can matter for intervention, whereas on METR-LA much of the transferred intervention benefit is captured by a simpler signed magnitude score. Together, these results show when task-conditioned communication has a coherent population meaning while separating that guarantee from how much task-specific detail is useful in a particular model.

1 INTRODUCTION

A graph neural network uses its graph to decide where information can flow, but the graph alone does not tell us whether *more* communication along a particular route would help the task. The same exchange can be useful for one task and harmful for another: mixing two populations may suppress irrelevant variation, or it may erase a task-relevant contrast. We therefore ask a local question: *with the model's feature parameters fixed, where would a small increase or decrease in communication improve its task objective?*

On a finite graph, a loss derivative with respect to an edge weight answers this question. Comparing such derivatives across resolutions is harder. Different graphs have different node sets, so their derivatives live in different coordinates, and the scale represented by one finite edge changes with resolution. Dense random samples introduce a second difficulty: an adjacency matrix can remain noisy edge by edge even when the induced message-passing computation becomes regular. We therefore seek a population-level communication derivative that remains meaningful across these finite-graph differences.

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

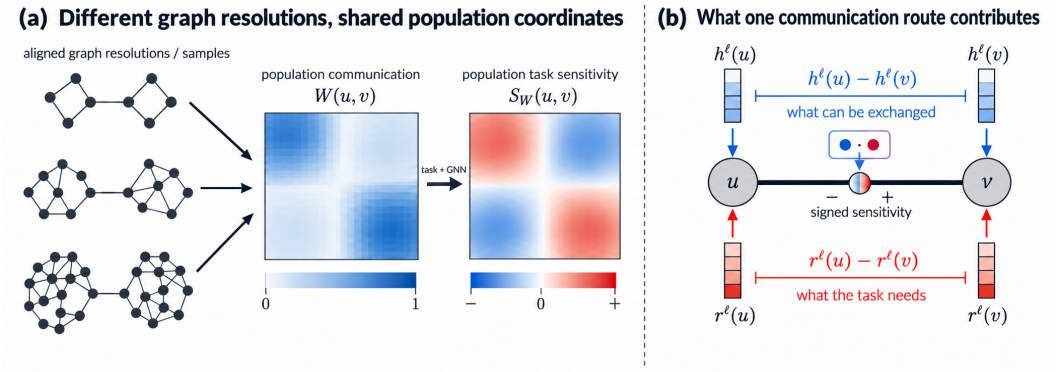


Figure 1: **From finite communication graphs to a population task sensitivity. (a)** Aligned graph resolutions or samples share population coordinates: $W(u, v)$ describes available communication and $S_W(u, v)$ how the task values changing it. **(b)** A route couples the forward difference $h^\ell(u) - h^\ell(v)$, describing what can be exchanged, with the backward difference $r^\ell(u) - r^\ell(v)$, describing what the task needs.

We use a graphon $W : [0, 1]^2 \to [0, 1]$ as a common population description of communication. A location $u \in [0, 1]$ indexes a population coordinate, while $h^\ell(u) \in \mathbb{R}^d$ stores finite-dimensional features at that location and layer. Thus the node coordinate becomes continuous while the feature dimension remains finite. In the same coordinates, a signed field $S_W(u, v)$ records how the task locally values strengthening or weakening communication between population locations u and v.

The field has a simple interpretation. At each layer, $h^\ell(u) - h^\ell(v)$ describes what information is available to cross a route. Backpropagation supplies a second endpoint difference describing what the downstream task needs. Their alignment determines whether strengthening that route lowers or raises the loss, and the exact sensitivity sums these forward–backward interactions across layers. Thus W describes *where communication is available*, whereas S_W describes *how a particular model and task value changing that communication*.

Our main theoretical result shows that this two-variable field has a stable population meaning. The graphon cut topology measures aggregate communication discrepancy rather than edgewise agreement. Under uniform boundedness, fixed depth and feature parameters, a smooth activation, and a ridge readout, cut convergence of aligned graphon–task instances implies L^1 convergence of their sensitivities. Stability propagates through the forward states, fitted readout, backward task signals, and finally the sensitivity field itself. For a Bernoulli sample A_n of a mean kernel W_n,

$$\mathbb{E}\|S_{A_n} - S_{W_n}\|_1 = O(n^{-1/2}), \tag{1}$$

even when $\mathbb{E}\|A_n - W_n\|_2^2$ stays bounded away from zero. A finite graph can therefore remain noisy edge by edge while inducing an increasingly stable population-level preference over communication routes.

This extends existing graph-limit analyses in a specific direction. Prior work studies forward stability, adjoint convergence, and gradients with respect to neural parameters. Here the differentiated object is the communication operator itself, so the derivative is a function of two population variables. Generic graphon optimization supplies the ambient geometry; our contribution identifies the message-passing-specific task derivative and proves its cut stability. A supporting result gives finite-horizon consistency of repeated communication updates under stronger aligned L^2 assumptions.

The experiments ask three questions. *Does the field stabilize?* Deterministic refinements, Bernoulli samples, and a CiteSeer-calibrated population test the theoretical predictions. *Does the task itself change the field?* A controlled diagnostic changes the targets while fixing the entire forward computation. *When does task-specific structure matter operationally?* Controlled routing gives a setting where route-specific alignment helps, while METR-LA shows improving cross-resolution transfer but also a regime where a simpler signed magnitude score captures most of the benefit.

2 Related Work

Population limits of neural computation. Graphons arise as limit objects for convergent dense graph sequences (Lovász & Szegedy, 2006; Lovász, 2012) and provide a natural population-level description of large networks. Graphon neural networks use such population limits to study size transfer in dense graphs (Ruiz et al., 2020), with extensions to random-graph generalization, graphon–signal stability, and parameter transference (Maskey et al., 2022; Levie, 2023; Cervino et al., 2023). Graphops provide an operator framework suited to sparse graph sequences (Le & Jegelka, 2023; Amran et al., 2026), while continuous-depth work studies graphon limits of neural differential equations (Yan et al., 2025). These works primarily ask whether a neural computation, prediction, or learned parameterization remains meaningful as graph size changes.

Most closely, Yan et al. (2026) establish forward, adjoint, and neural-parameter-gradient convergence on sparse random graphs. Their adjoint analysis provides relevant machinery, but the differentiated object is different. We differentiate the *communication operator itself*. A perturbation of W changes message exchange at every layer and produces a sensitivity field $S_W(u, v)$ with one population coordinate for each endpoint of a route. Convergence of a finite-dimensional neural-parameter gradient therefore does not by itself establish stability of this two-variable structural derivative.

Optimizing and interpreting communication. Graphon gradient flows provide continuum optimization geometry for generic functionals of a graphon (Oh et al., 2024). At finite size, graph-structure learning, diffusion, and rewiring modify communication during learning or preprocessing (Franceschi et al., 2019; Chen et al., 2020; Gasteiger et al., 2019; Topping et al., 2022). GNNExplainer, PGExplainer, and edge-gradient methods instead identify edges important for a prediction (Ying et al., 2019; Luo et al., 2020; He et al., 2025).

Our question is complementary to both directions. We do not propose a new finite-graph explainer or optimizer; we ask whether a local derivative with respect to communication has a coherent *population meaning* across graph resolutions and random edge realizations. Generic graphon optimization supplies the ambient geometry, while our contribution is the message-passing-specific forward–backward derivative and its stability as a task-conditioned field. This distinction has a significant consequence: stability of the communication derivative implies consistency of bounded first-order structural interventions, a conclusion that does not follow from convergence of forward predictions or neural-parameter gradients alone. Appendix B compares the differentiated objects and guarantees of the closest graph-limit results.

3 From a Finite Graph to a Population Model

Node coordinates and feature coordinates. We study mean-normalized difference aggregation. For a symmetric communication matrix $A \in [0, 1]^{n\times n}$ and node representations $H_i \in \mathbb{R}^d$,

$$(D_A H)_i = \frac{1}{n} \sum_{j=1}^{n} A_{ij}(H_j - H_i). \tag{2}$$

The graph acts along the node axis: it determines which d-dimensional feature vectors are exchanged while their dimension remains finite.

Partition $[0, 1]$ into equal cells I_i, set $W_A(u, v) = A_{ij}$ on $I_i \times I_j$, and set $h(u) = H_i$ on I_i. Then

$$(D_W h)(u) = \int_0^1 W(u, v)[h(v) - h(u)]\, dv \tag{3}$$

is exactly Equation (2) for these step functions. The population model therefore replaces discrete node indices by locations in $[0, 1]$ without changing the finite-dimensional feature computation. Write $\mathcal{K} = \{W \in L^2_{\mathrm{sym}}([0, 1]^2) : 0 \le W \le 1\}$.

Forward computation. Starting from $h^0 = x$, let $B_0, B_1 \in \mathbb{R}^{d\times d}$ be fixed feature maps and β a bias:

$$\begin{aligned} z^\ell(u) &= B_0 h^\ell(u) + B_1(D_W h^\ell)(u) + \beta, \\ h^{\ell+1}(u) &= (1 - \Delta\tau)h^\ell(u) + \Delta\tau\, \sigma(z^\ell(u)). \end{aligned} \tag{4}$$

Here B_0 transforms the local representation and B_1 the communicated difference. This separation lets us vary communication while holding ordinary feature parameters fixed.

Task and optimized readout. We allow the task head to adapt when the communication kernel changes while keeping the feature maps fixed. Let $y : [0,1] \to \mathbb{R}^c$ be the target and $m : [0,1] \to [0,1]$ a supervision weight, with $\bar{m} = \int_0^1 m(u)\,du > 0$. We fit the unique ridge readout

$$\Phi_W(R) = \frac{1}{2\bar{m}} \int_0^1 m(u)\|Rh_W^L(u) - y(u)\|^2 du + \frac{\mu}{2}\|R\|_F^2, \qquad R_W = \arg\min_R \Phi_W(R). \tag{5}$$

Its solution is $R_W = C_{yh}(C_h + \mu I)^{-1}$, with covariances weighted by $m/\bar{m}$. Positive ridge μ makes the optimum unique and uniformly controls the inverse.

When studying communication updates we may also penalize movement from a reference kernel:

$$\widehat{J}(W) = \Phi_W(R_W) + \frac{\lambda}{2}\|W - W_{\mathrm{ref}}\|_2^2, \qquad \lambda \geq 0. \tag{6}$$

We keep this fidelity term separate from task sensitivity: the latter asks which communication changes help the task, whereas the regularizer can alter the eventual optimization direction.

Standing assumptions. Signals x, y and masks are uniformly bounded, $\bar{m} \geq m_{\min} > 0$, $\sigma \in C_b^2$, $0 < \Delta\tau \leq 1$, and $\mu > 0$. Depth, dimensions, and feature parameters are fixed; kernels are dense, symmetric, and bounded in $[0,1]$. Cross-resolution statements assume the models are already expressed in aligned population coordinates.

4 The Task Derivative of Communication

To understand one route, imagine slightly increasing its weight while keeping feature parameters fixed. The forward endpoint difference describes what information is available to cross the route. Its usefulness depends on the downstream loss, whose backpropagation produces a corresponding backward message signal. The route sensitivity is determined by the alignment of these two endpoint differences.

Here p^ℓ is the backpropagated sensitivity to the layer-ℓ representation, q^ℓ applies the activation gate, and r^ℓ expresses the result in message coordinates.

Theorem 1 (Exact task-conditioned communication sensitivity)**.** *Set*

$$p^L(u) = \frac{m(u)}{\bar{m}} R_W^\top \big(R_W h^L(u) - y(u)\big), \tag{7}$$

and, backward for $\ell = L-1, \ldots, 0$, define

$$q^\ell = D\sigma(z^\ell)^\top p^{\ell+1}, \qquad r^\ell = B_1^\top q^\ell, \tag{8}$$

$$p^\ell = (1-\Delta\tau)p^{\ell+1} + \Delta\tau B_0^\top q^\ell + \Delta\tau D_W r^\ell. \tag{9}$$

Then, for every bounded symmetric perturbation V,

$$D\Phi_W(R_W)[V] = -\langle S_W, V\rangle_{L^2}, \tag{10}$$

where

$$\boxed{S_W(u,v) = \frac{\Delta\tau}{2} \sum_{\ell=0}^{L-1} \underbrace{[h^\ell(u) - h^\ell(v)]^\top}_{\textit{available exchange}} \underbrace{[r^\ell(u) - r^\ell(v)]}_{\textit{backward task signal}}.} \tag{11}$$

The full objective has gradient density $G(W) = \lambda(W - W_{\mathrm{ref}}) - S_W$.

The recursion is ordinary backpropagation in population coordinates. The activation Jacobian $D\sigma$ gates the local gradient, while $B_1^\top$ maps it into communicated-message coordinates. Equation (11) gives the central interpretation: a route is useful not merely because its endpoints carry different representations, but because that difference aligns with what the downstream task needs.

Why endpoint differences appear. Because a symmetric route affects both endpoints,

$$\langle g, D_V f\rangle = -\frac{1}{2}\int V(u,v)[g(u)-g(v)]^\top[f(u)-f(v)]\,du\,dv. \tag{12}$$

This gives the endpoint differences in Equation (11); the factor $1/2$ avoids double-counting an undirected route. Because R_W minimizes Φ_W, $\nabla_R\Phi_W(R_W)=0$, so the first-order derivative through the refitted readout vanishes by the envelope property. Appendix A.2 gives the derivation.

What the sign means. Positive $\langle S_W, V\rangle$ predicts a decrease in task loss to first order. If a route can be changed independently, positive sensitivity favors strengthening it and negative sensitivity favors weakening it. Under constraints such as row conservation, the relevant quantity is instead the sensitivity of the complete balanced perturbation. Appendix A.6 gives a two-block example in which within-block mixing suppresses nuisance while cross-block mixing erases task-relevant contrast. The derivative is local: it does not certify a large finite intervention or solve global rewiring.

5 Population Stability

The exact derivative tells us how a task values communication on one graph. For this quantity to be meaningful across resolutions or random realizations, the sensitivity fields must agree whenever the underlying systems represent nearly the same population.

Edgewise agreement is stronger than message passing requires. The operator D_W averages communication against node signals, so two kernels can differ on many individual edges while inducing nearly the same aggregate exchange. The cut topology captures this weaker notion.

5.1 Stability of the complete sensitivity field

For a bounded kernel U, define

$$\|U\|_{\square,*} = \sup_{\|a\|_\infty,\|b\|_\infty\le 1}\left|\int U(u,v)a(u)b(v)\,du\,dv\right|. \tag{13}$$

This is equivalent to the usual cut norm up to an absolute constant. For aligned instances $\mathcal{Z} = (W,x,y,m)$ and $\widetilde{\mathcal{Z}} = (\widetilde{W},\widetilde{x},\widetilde{y},\widetilde{m})$, define

$$d_{\square,1}(\mathcal{Z},\widetilde{\mathcal{Z}}) = \|W-\widetilde{W}\|_{\square,*} + \|x-\widetilde{x}\|_1 + \|y-\widetilde{y}\|_1 + \|m-\widetilde{m}\|_1. \tag{14}$$

Theorem 2 (Cut stability of task-conditioned sensitivity)**.** *Under the standing assumptions and common uniform bounds, there exists a constant C, independent of graph resolution, such that*

$$\boxed{\|S_{\mathcal{Z}} - S_{\widetilde{\mathcal{Z}}}\|_1 \le C\, d_{\square,1}(\mathcal{Z},\widetilde{\mathcal{Z}}).} \tag{15}$$

Forward states and adjoints satisfy analogous L^1 bounds, and the optimized readouts satisfy an analogous matrix-norm bound.

The constant is resolution-independent but is not claimed to be uniform in model depth or parameter scale; it may depend on the fixed feature-map norms, activation bounds, mask mass, and ridge conditioning, in particular through μ^{-1}.

The theorem controls more than the forward GNN: stability must propagate through the fitted readout, backward task signal, and two-variable sensitivity field. A cut-to-L^1 operator estimate provides the key step; at fixed depth it propagates through the forward and backward recursions, after which boundedness controls the products forming S_W. Appendix A.3 gives the complete argument.

Under a common measure-preserving relabeling ϕ, $S_{\mathcal{Z}^\phi}(u,v) = S_{\mathcal{Z}}(\phi(u),\phi(v))$, so the field is equivariant to population labels once communication and task signals are aligned.

5.2 Persistent edge noise, convergent sensitivity

Let W_n be an n-cell mean kernel and sample $A_{ij} \sim \mathrm{Bernoulli}(W_{ij})$ independently for $i<j$, reflecting symmetrically and fixing the diagonal. Then

$$\mathbb{E}\|A_n - W_n\|_2^2 = \frac{1}{n^2}\sum_{i\ne j} W_{ij}(1-W_{ij}), \tag{16}$$

which stays positive when a nonvanishing region has probabilities bounded away from zero and one. Individual sampled edges therefore remain noisy. In contrast, Hoeffding concentration and a union bound give $\mathbb{E}\|A_n - W_n\|_{\square,*} = O(n^{-1/2})$.

Corollary 3 (Fixed-grid Bernoulli consistency)**.** *For uniformly bounded step signals and fixed feature parameters, conditional on the latent grid, signals, and supervision mask, if the only remaining randomness is edge sampling, then*

$$\boxed{\mathbb{E}\|S_{A_n} - S_{W_n}\|_1 = O(n^{-1/2}).} \tag{17}$$

The comparison is to the *corresponding mean-kernel sensitivity* S_{W_n}. Convergence to one fixed population sensitivity additionally requires the deterministic approximation $S_{W_n} \to S_W$; independently resampling node signals across resolutions does not by itself guarantee this. Thus the adjacency can remain noisy edge by edge while its induced population preference over communication routes becomes stable.

What this guarantees for interventions. For any bounded symmetric perturbation V,

$$|D\Phi_n[V] - D\Phi[V]| \le \|V\|_\infty \|S_n - S\|_1. \tag{18}$$

Hence bounded candidate interventions receive increasingly consistent first-order values as the sensitivity fields converge. The guarantee therefore goes beyond visual agreement of the fields: every bounded candidate communication change receives an asymptotically consistent first-order value. For a finite candidate family separated by a uniformly positive mean-kernel decision margin, the probability that Bernoulli edge sampling changes the preferred candidate is also $O(n^{-1/2})$; Appendix A.4 gives the formal margin argument. This does not imply convergence of individual edge rankings or large finite interventions.

5.3 Repeated updates require stronger control

Cut convergence is sufficient for stability of the local sensitivity field, but repeated communication updates require stronger control of the kernel itself. We therefore use aligned L^2 convergence for optimization trajectories.

Theorem 4 (Finite-horizon update consistency)**.** *Suppose aligned equal-cell discretizations of $x, y, m, W_{\mathrm{ref}}, W^0$ converge in L^2, with common fixed feature parameters and step size η. Then, for every fixed horizon K,*

$$\max_{0\le k\le K} \|W_n^k - W^k\|_2 \le C_K \varepsilon_n \longrightarrow 0, \tag{19}$$

where ε_n is the total aligned L^2 approximation error. States, readouts, predictions, adjoints, gradient densities, and objectives converge over the same fixed horizon.

Thus the two results serve different purposes: cut convergence is enough to make the local task-conditioned sensitivity field stable, whereas aligned L^2 convergence additionally controls repeated communication updates. The required finite-to-population gradient scaling is n^2 for full matrix coordinates and $n^2/2$ for shared undirected-edge coordinates; the coordinate conversion and proof appear in Appendix A.5.

6 Experiments

We first test whether the population sensitivity behaves as predicted before asking when it is useful for intervention. The first studies match the theory's fixed-parameter setting; the task swap isolates supervision, and routing and METR-LA test operational meaning. Error bars are sample SDs unless stated otherwise; full protocols appear in Appendix C.

6.1 Derivative certification and deterministic consistency

Across three synthetic kernel families and five resolutions, the analytic sensitivity agrees with float64 autodiff to worst relative error 1.89×10^{-16}. A central finite-difference check that refits the ridge

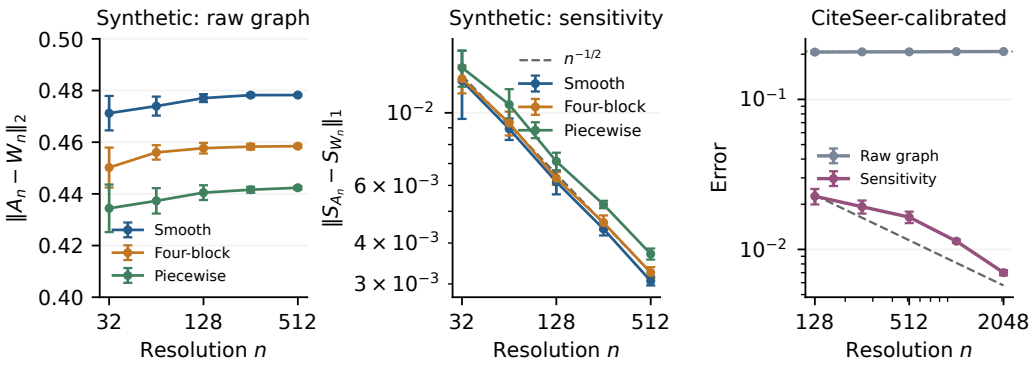


Figure 2: **Stable sensitivity despite persistent edge noise.** Means ± SD over 12 samples. Synthetic families and the CiteSeer-calibrated population show convergent sensitivity while raw graph error remains nonvanishing. Dashed lines indicate slope $-1/2$.

readout agrees to 4.8×10^{-10} for the task derivative and 1.3×10^{-9} for the full objective, certifying the sign, endpoint factor, activation gate, and envelope calculation.

For smooth, four-block, and piecewise-discontinuous kernels, successive refinements $n \to 2n$, $n = 32, 64, 128, 256$, give sensitivity RMS decay exponents 0.94–0.96. The minimum sensitivity cosine over six updates exceeds 0.9978 by $64 \to 128$ and 0.9995 at the two finest comparisons. At matched population-L^2 movement 0.008, coarse and fine sensitivities induce task-loss changes differing by at most 2.2×10^{-5} for $64 \to 128$.

The analytic derivative is numerically certified, and the population sensitivity becomes rapidly consistent under deterministic refinement.

6.2 A NOISY GRAPH CAN HAVE A STABLE SENSITIVITY

Synthetic populations. Across 12 Bernoulli samples per family and resolution, raw kernel L^2 error remains approximately 0.43–0.48 from $n = 32$ to 512, while mean sensitivity L^1 error falls from 1.25–1.37×10^{-2} to 3.06–3.71×10^{-3}. Fitted decay exponents are 0.508, 0.495, and 0.480; relative sensitivity error falls from 5.83–7.10% to 1.36–1.82%.

A dense population calibrated from CiteSeer. We estimate class proportions and class-to-class edge frequencies from Planetoid CiteSeer (Yang et al., 2016), construct a dense six-block population kernel, and bootstrap class-conditional 16-dimensional PCA features. Feature maps are trained on topic classification at reference resolution 1024 and then frozen. This is a calibrated dense population model; the theorem is not being applied to the raw sparse citation graph.

From $n = 128$ to 2048, raw error approaches the analytic Bernoulli plateau 0.20854, while sensitivity L^1 error falls from 0.02261 to 0.00700 and cosine rises from 0.99497 to 0.99983. The fitted slope is -0.415 over all five resolutions and -0.491 for $n \geq 256$, approaching the predicted $n^{-1/2}$ behavior. Features are resampled across resolutions, so this tests the edge-conditional comparison with S_{W_n} rather than convergence to one fixed population instance.

Across both synthetic and CiteSeer-calibrated populations, sensitivity converges at approximately the predicted $n^{-1/2}$ rate even while raw edge noise remains nonvanishing.

6.3 SAME POPULATION AND FORWARD MODEL, DIFFERENT TASKS

To isolate task dependence, we use a five-seed diagnostic at $n = 2048$. Two orthogonal teacher directions define balanced binary probes from within-class feature residuals; removing their span leaves a shared 14-dimensional input. The forward model is trained *only on topic classification* and then frozen. Between probes A and B, the kernel, inputs, masks, feature parameters, and every forward state are identical; only the targets, fitted readout, and backward chain change.

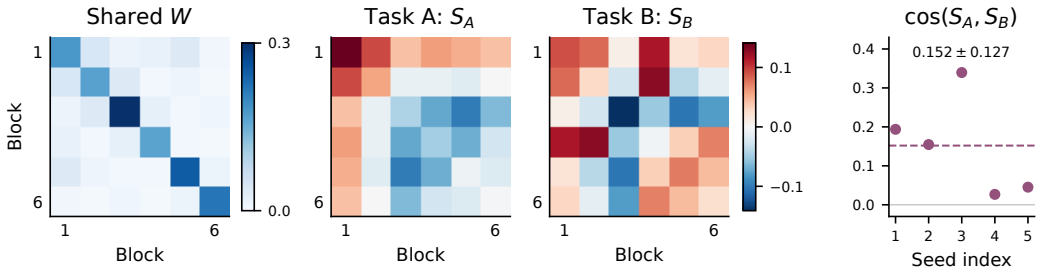


Figure 3: **Changing the task changes the field.** Common kernel and block sensitivities for probes A and B, followed by full-field cosine across five seeds. Only the targets and readout/backward chain differ within each comparison.

Across seeds, the full sensitivity cosine is 0.152 ± 0.127, and $28.0\% \pm 5.6\%$ of the 15 cross-block route types reverse sign. The corresponding backward-field cosines across the three layers average 0.111, 0.155, and 0.252, showing that the separation arises in the task-dependent backward chain while the forward representation remains shared.

Changing only the task substantially changes the communication sensitivity, even when the population, kernel, and entire forward representation are fixed.

These are mechanism probes: test accuracies are 54.46% for A and 52.91% for B, versus feature-only values 54.63% and 53.71%. They establish task dependence, not useful communication on these diagnostic tasks.

6.4 Route-specific alignment can matter

A four-population relational task makes useful and harmful source–target routes similar in forward statistics. We compare the exact sensitivity with forward-only, backward-only, and magnitude-product controls. Both signs are available; the best control and sign are selected on validation loss. All changes have equal population-L^2 movement and preserve each population's total communication. Let P denote projection onto the row-conserving feasible space.

The exact and positive forward-only feasible directions are nearly orthogonal: $\cos(PS, PF) = -0.0664 \pm 0.0206$. At movement 0.0108, full sensitivity changes held-out MSE by -0.00702 ± 0.00072, compared with -0.00052 ± 0.00013 for the selected simple control. Full sensitivity wins all 15 seed–budget comparisons and all 15 in a reduced-cue variant.

Route-specific forward–backward alignment can therefore carry useful intervention information that simpler forward-, backward-, or magnitude-based signals miss.

6.5 METR-LA: transfer toward the native sensitivity

We train traffic predictors independently on nested sets of 64, 128, and 207 METR-LA sensors (Li et al., 2018). With model parameters frozen, we lift a source sensitivity using geographic correspondence and compare it with the native 207-sensor derivative. This transfer experiment lies beyond the shared-parameter theorem.

Across different-seed source–target pairs, cosine with the native field rises from 0.7478 ± 0.0047 at 64 sensors to 0.9036 ± 0.0008 at 128. We match population-L^2 movement on each frozen target; the native gradient determines only the common budget, never the transferred direction.

Under row conservation, 128-sensor sensitivity recovers 69.70% of the native gain, improving MAE by 0.04149 mph (0.68% of the 6.078 mph baseline). A direct geographic control achieves 0.02528 mph, a lifted geographic control 0.02143 mph, and distance alone 0.00985 mph; shuffling the residual within distance strata reduces the gain to 0.00367 mph (Appendix C.6). Thus the transferred field contains actionable structure beyond these specified smooth geographic approximations.

Table 1: **METR-LA native versus transferred sensitivity.** Held-out MAE reduction (mph), mean $\pm$ SD over five targets, at matched population movement.

	Without row conservation		With row conservation	
Source	MAE reduction	Native gain	MAE reduction	Native gain
64	0.10494 ± 0.00156	83.61%	0.02554 ± 0.00017	42.91%
128	0.11349 ± 0.00265	90.41%	0.04149 ± 0.00049	69.70%
Native 207	0.12552 ± 0.00229	100%	0.05953 ± 0.00049	100%

The coupling also simplifies in this setting. A magnitude-product score retains forward and backward norms but removes route-specific angles. Its global sign is selected independently using only the source model's validation loss: all five 64-sensor and all five 128-sensor checkpoints select the negative orientation, consistently across source-side movement budgets 0.005, 0.010, and 0.015. At $128 \to 207$, this validation-signed magnitude score gains 0.04079 ± 0.00057 mph, or 98.3% of the transferred full-sensitivity gain. Thus, in this METR-LA configuration, route-specific angular information contributes little beyond forward/backward magnitudes and a coarse orientation that can be determined without access to the full sensitivity field.

Sensitivity becomes more transferable as source resolution increases, but in this METR-LA model nearly all of the transferred intervention gain is captured by a simpler signed-magnitude signal.

7 Discussion

Graphons provide common population coordinates for separating two questions that are conflated on a single finite graph: *where communication exists* and *how a particular task values changing that communication*. The exact derivative identifies the latter through forward–backward coupling, while the cut-stability theorem shows that this task-conditioned field can remain meaningful across dense graph samples even when the adjacency itself does not converge edgewise in raw L^2. The result is therefore not merely a transfer statement for predictions: it gives a population interpretation to the local structural preference induced by the task. This makes task-conditioned communication a population object on which questions of transfer, intervention, approximation, and compression can be posed independently of any particular finite graph realization.

The theory also clarifies what kind of stability is being claimed. Cut convergence is enough for the local sensitivity field and for bounded first-order intervention values, whereas repeated communication updates require stronger aligned L^2 control. The guarantees concern models represented in common population coordinates with fixed feature parameters; they do not provide an alignment procedure for unrelated graphs, compare independently retrained feature maps, or imply stability of individual edge rankings. In this sense, the graphon acts as a population coordinate system for the derivative rather than as an assumption that arbitrary finite graphs should be compared edge by edge.

The experiments deliberately separate existence of the population object from its usefulness as an intervention rule. Synthetic and calibrated populations verify that the field stabilizes despite persistent edge noise, and the fixed-forward diagnostic shows that changing only the task can substantially change it. Controlled routing then gives a setting where route-specific forward–backward alignment matters, whereas METR-LA shows a different regime: cross-resolution sensitivity transfers toward the native derivative, but a simpler signed-magnitude score captures nearly all of the resulting intervention gain. Thus stability of the full field does not imply that all of its structure is operationally necessary in every model.

The present guarantees assume dense symmetric kernels, aligned bounded signals, smooth activations, positive ridge, fixed depth, and fixed feature parameters. They do not establish sparse-graph transfer, joint feature–topology training, global rewiring optimality, or benefit when the original communication operator is already well adapted to the task. Extending the differentiated object to sparse operator limits (Le & Jegelka, 2023; Amran et al., 2026; Yan et al., 2026), learned alignments, joint feature–communication optimization, and settings with beneficial spatial mixing are natural next directions.

AI USE DISCLOSURE

Generative AI tools assisted with proof development and checking, experiment design and implementation, numerical analysis and interpretation of results, literature search and positioning, figure preparation, and manuscript drafting and editing. All AI-assisted outputs were reviewed and verified by the authors. The authors take responsibility for the mathematical arguments, empirical claims, citations, code, figures, and final submission.

Reproducibility statement. Complete proofs and extensive experimental materials are provided in the supplement. The package includes source code for the synthetic, controlled-routing, and CiteSeer experiments; archived seed-level results and numerical diagnostics; the CiteSeer calibration and task-intervention artifacts; METR-LA checkpoints, scores, lifted sensitivity fields, configurations, and evaluation records; and scripts that reconstruct the reported tables, summary statistics, and figures. The supplementary README maps the main experimental claims to the corresponding code and result files and provides commands for reaggregation and evaluation.

REFERENCES

Ofek Amran, Tom Gilat, and Ron Levie. A graphop analysis of graph neural networks on sparse graphs: Generalization and universal approximation. *arXiv preprint arXiv:2602.08785*, 2026.

Juan Cervino, Luana Ruiz, and Alejandro Ribeiro. Learning by transference: Training graph neural networks on growing graphs. *IEEE Transactions on Signal Processing*, 71:233–247, 2023. doi: 10.1109/TSP.2023.3242374.

Yu Chen, Lingfei Wu, and Mohammed J. Zaki. Iterative deep graph learning for graph neural networks: Better and robust node embeddings. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1938–1949, 2020.

Luca Franceschi, Mathias Niepert, Massimiliano Pontil, and Xiao He. Learning discrete structures for graph neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 1972–1982. PMLR, 2019.

Johannes Gasteiger, Stefan Weißenberger, and Stephan Günnemann. Diffusion improves graph learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jesse He, Akbar Rafiey, Gal Mishne, and Yusu Wang. Explaining gnn explanations with edge gradients. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 884–895, 2025. doi: 10.1145/3711896.3736947.

Thien Le and Stefanie Jegelka. Limits, approximation and size transferability for gnns on sparse graphs via graphops. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ron Levie. A graphon-signal analysis of graph neural networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 64482–64525, 2023.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018.

László Lovász. *Large Networks and Graph Limits*, volume 60 of *Colloquium Publications*. American Mathematical Society, Providence, RI, 2012. doi: 10.1090/coll/060.

László Lovász and Balázs Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933–957, 2006. doi: 10.1016/j.jctb.2006.05.002.

Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Sohir Maskey, Ron Levie, Yunseok Lee, and Gitta Kutyniok. Generalization analysis of message passing neural networks on large random graphs. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4805–4817, 2022.

Sewoong Oh, Soumik Pal, Raghav Somani, and Raghavendra Tripathi. Gradient flows on graphons: Existence, convergence, continuity equations. *Journal of Theoretical Probability*, 37(2):1469–1522, 2024. doi: 10.1007/s10959-023-01271-8.

Luana Ruiz, Luiz F. O. Chamon, and Alejandro Ribeiro. Graphon neural networks and the transferability of graph neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1702–1712, 2020.

Jake Topping, Francesco Di Giovanni, Benjamin P. Chamberlain, Xiaowen Dong, and Michael M. Bronstein. Understanding over-squashing and bottlenecks on graphs via curvature. In *International Conference on Learning Representations*, 2022.

Mingsong Yan, Charles Kulick, and Sui Tang. On the convergence and size transferability of continuous-depth graph neural networks. *arXiv preprint arXiv:2510.03923*, 2025.

Mingsong Yan, Zhida Wang, and Sui Tang. Zero-shot size transfer for neural odes on sparse random graphs: Graphon limits and adjoint convergence, 2026. URL https://arxiv.org/abs/2606.26662. arXiv:2606.26662.

Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pp. 40–48. PMLR, 2016. URL https://proceedings.mlr.press/v48/yanga16.html.

Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. GNNExplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1903.03894.

A Proofs and Technical Details

We use $\|\cdot\|_p$ for the $L^p([0,1])$ norm of vector-valued signals and $\|\cdot\|_{L^2}$ for kernels. Constants may depend on fixed depth, feature matrices, activation bounds, mask lower bound, ridge parameter, and common signal bounds, but not on discretization size.

A.1 Difference-operator identities

Lemma 5 (Pairwise identity and operator bounds)**.** *For symmetric $W \in \mathcal{K}$ and $f, g \in L^2([0,1]; \mathbb{R}^d)$,*

$$\langle g, D_W f\rangle_{L^2} = -\frac{1}{2}\int_{[0,1]^2} W(u,v)[g(u)-g(v)]^\top[f(u)-f(v)]\,du\,dv. \tag{20}$$

Moreover,

$$\|D_W f\|_2 \le 2\|f\|_2, \qquad \|D_W f\|_\infty \le 2\|f\|_\infty,$$

and, for $U \in L^2([0,1]^2)$ and bounded f,

$$\|D_U f\|_2 \le 2\|U\|_{L^2}\|f\|_\infty. \tag{21}$$

Proof. Expand the definition and interchange u and v. Symmetry and averaging give the pairwise identity. Write

$$D_W = T_W - M_{d_W}, \qquad (T_W f)(u) = \int_0^1 W(u,v)f(v)\,dv, \qquad d_W(u) = \int_0^1 W(u,v)\,dv.$$

Since $0 \le W \le 1$ on a unit-measure domain, $\|T_W\|_{2\to 2} \le \|W\|_2 \le 1$ and $\|d_W\|_\infty \le 1$. The L^∞ bound follows pointwise. For Equation (21), apply Cauchy–Schwarz first in v and then in u to the integral and degree terms. □

A.2 Exact derivative and adjoint

Differentiability is taken in the bounded-kernel topology, without assuming a nonlinear $L^2 \to L^2$ Fréchet derivative. For a bounded symmetric direction V, the tangent state $\xi^\ell = Dh^\ell_W[V]$ satisfies

$$\xi^0 = 0, \tag{22}$$

$$\xi^{\ell+1} = (1-\Delta\tau)\xi^\ell + \Delta\tau D\sigma(z^\ell)\left[B_0\xi^\ell + B_1 D_W \xi^\ell + B_1 D_V h^\ell_W\right]. \tag{23}$$

Pairing this recursion with $p^{\ell+1}$ and using self-adjointness of D_W produces the backward recursion in Equation (9). Summing over layers cancels the tangent-state terms and leaves

$$D\Phi_W(R_W)[V] = \Delta\tau\sum_{\ell=0}^{L-1}\langle r^\ell, D_V h^\ell\rangle. \tag{24}$$

Because R_W is the unique ridge minimizer,

$$\nabla_R \Phi_W(R_W) = 0.$$

Hence the first-order term passing through the dependence of R_W on W vanishes by the envelope property. Applying Equation (20) gives

$$\langle r^\ell, D_V h^\ell\rangle = -\frac{1}{2}\int V(u,v)[r^\ell(u)-r^\ell(v)]^\top[h^\ell(u)-h^\ell(v)]\,du\,dv.$$

The symmetric route contributes through both endpoint orderings; the factor $1/2$ removes this double counting. Substitution into Equation (24) proves Equation (11). Adding the fidelity derivative gives

$$G(W) = \lambda(W - W_{\mathrm{ref}}) - S_W.$$

Uniform bounds on states, adjoints, $D\sigma$, and B_1 imply boundedness and symmetry of the resulting density.

A.3 Proof of cut stability

The main estimate is a cut-to-L^1 stability property of difference aggregation.

Lemma 6 (Difference aggregation is cut-to-L^1 stable)**.** *Let $W, \widetilde{W} \in \mathcal{K}$ and let $f, \widetilde{f} : [0,1] \to \mathbb{R}^d$ satisfy $\|f\|_\infty, \|\widetilde{f}\|_\infty \le M$. Then*

$$\|D_W f - D_{\widetilde{W}} \widetilde{f}\|_1 \le 2\|f - \widetilde{f}\|_1 + (d+1)M\|W - \widetilde{W}\|_{\square,*}. \tag{25}$$

Proof. Set $Q = W - \widetilde{W}$ and write $D_W = T_W - M_{d_W}$. Since $0 \le W \le 1$,

$$\|D_W(f - \widetilde{f})\|_1 \le 2\|f - \widetilde{f}\|_1.$$

For each coordinate j, scalar L^1–L^∞ duality gives

$$\|(T_Q \widetilde{f})_j\|_1 = \sup_{\|a\|_\infty \le 1} \left| \int Q(u,v) a(u) \widetilde{f}_j(v)\, du\, dv \right| \le M\|Q\|_{\square,*}.$$

Summing coordinates gives $\|T_Q \widetilde{f}\|_1 \le dM\|Q\|_{\square,*}$. Likewise, $\|d_Q\|_1 \le \|Q\|_{\square,*}$, so $\|d_Q \widetilde{f}\|_1 \le M\|Q\|_{\square,*}$. Combining the terms proves the claim. □

Proof of Theorem 2. Because $0 < \Delta\tau \le 1$ and σ is bounded, all forward states are uniformly bounded. Let

$$e_h^\ell = \|h^\ell - \widetilde{h}^\ell\|_1.$$

Applying Lemma 6 to one forward layer and using Lipschitz continuity of σ gives

$$e_h^{\ell+1} \le c_h e_h^\ell + c_W \|W - \widetilde{W}\|_{\square,*},$$

where the constants depend only on fixed parameters and common bounds. Finite induction therefore yields

$$\max_{\ell \le L} \|h^\ell - \widetilde{h}^\ell\|_1 \le C_h \left(\|x - \widetilde{x}\|_1 + \|W - \widetilde{W}\|_{\square,*} \right). \tag{26}$$

The same argument controls $z^\ell - \widetilde{z}^\ell$.

Write

$$\rho = \frac{m}{\overline{m}}, \qquad \widetilde{\rho} = \frac{\widetilde{m}}{\overline{\widetilde{m}}}.$$

The common lower bound $m_{\min}$ gives

$$\|\rho - \widetilde{\rho}\|_1 \le \left(m_{\min}^{-1} + m_{\min}^{-2}\right) \|m - \widetilde{m}\|_1.$$

Uniform state bounds and Equation (26) control the two ridge covariance matrices in Frobenius norm. The resolvent identity together with

$$\|(C_h + \mu I)^{-1}\|_{\mathrm{op}} \le \mu^{-1}$$

then gives

$$\|R_W - R_{\widetilde{W}}\|_F \le C_R d_{\square,1}(\mathcal{Z}, \widetilde{\mathcal{Z}}). \tag{27}$$

The terminal adjoint is

$$p^L = \rho R_W^\top (R_W h^L - y),$$

so Equations (26) and (27) and the uniform data bounds control $\|p^L - \widetilde{p}^L\|_1$. For one backward layer,

$$q^\ell = D\sigma(z^\ell)^\top p^{\ell+1}, \qquad r^\ell = B_1^\top q^\ell.$$

Bounded σ' and σ'' control the activation-gate perturbation. Subtracting the adjoint recursions and applying Lemma 6 to $D_W r^\ell - D_{\widetilde{W}} \widetilde{r}^\ell$ gives a finite backward recurrence. Hence

$$\max_\ell \|p^\ell - \widetilde{p}^\ell\|_1 + \max_{\ell<L} \|r^\ell - \widetilde{r}^\ell\|_1 \le C_p d_{\square,1}(\mathcal{Z}, \widetilde{\mathcal{Z}}). \tag{28}$$

Finally, with $\Delta f(u,v) = f(u) - f(v)$,

$$S_{\mathcal{Z}} - S_{\widetilde{\mathcal{Z}}} = \frac{\Delta\tau}{2} \sum_\ell \left[\Delta(h^\ell - \widetilde{h}^\ell)^\top \Delta r^\ell + \Delta \widetilde{h}^{\ell\top} \Delta(r^\ell - \widetilde{r}^\ell) \right].$$

Using

$$\|\Delta f\|_{L^1([0,1]^2)} \le 2\|f\|_1$$

and the common uniform bounds, Equations (26) and (28) give the claimed L^1 sensitivity estimate. □

Relabeling-invariant form. For a measure-preserving bijection ϕ, the model is equivariant:

$$S_{\mathcal{Z}^\phi}(u,v) = S_{\mathcal{Z}}(\phi(u), \phi(v)).$$

For the full objective, relabel W_{ref} by the same ϕ; invariance of integration preserves both the optimized task loss and the fidelity term. Define

$$\delta_{\square,1}(\mathcal{Z}, \widetilde{\mathcal{Z}}) = \inf_\phi d_{\square,1}(\mathcal{Z}, \widetilde{\mathcal{Z}}^\phi), \qquad \delta_1(S, \widetilde{S}) = \inf_\phi \|S - \widetilde{S}^\phi\|_1.$$

Applying Theorem 2 along an approximating sequence of relabelings gives

$$\delta_1(S_{\mathcal{Z}}, S_{\widetilde{\mathcal{Z}}}) \le C \delta_{\square,1}(\mathcal{Z}, \widetilde{\mathcal{Z}}).$$

Bernoulli concentration and proof of Corollary 3. Condition on the mean kernel and latent-determined signals. Put

$$Q_{ij} = A_{ij}^{(n)} - W_{ij}^{(n)}, \qquad Q_{ii} = 0, \qquad Q_{ji} = Q_{ij}.$$

For an equal-cell step kernel, bounded test functions can be replaced by their cell averages. Maximizing the resulting bilinear form over two cubes attains its maximum at sign vectors, so

$$\|Q\|_{\square,*} = \frac{1}{n^2} \max_{a,b \in \{-1,1\}^n} \left| \sum_{i<j} (a_i b_j + a_j b_i) Q_{ij} \right|.$$

For fixed a, b, the summands are independent and centered, with

$$|a_i b_j + a_j b_i| \le 2.$$

Their squared range lengths sum to at most $2n(n-1) \le 2n^2$. Hoeffding's inequality and a union bound over 4^n sign pairs therefore give, for $t > 0$,

$$\Pr\{\|Q\|_{\square,*} > t\} \le \min\{1,\, 2\,4^n e^{-n^2 t^2}\}.$$

Integrating the tail and splitting at

$$t_n = \frac{\sqrt{(2n+1)\log 2}}{n}$$

gives

$$\mathbb{E}\|Q\|_{\square,*} \le t_n + (2n^2 t_n)^{-1} = O(n^{-1/2}).$$

Applying Theorem 2 with identical signals for A_n and W_n proves the corollary.

Because the diagonal is fixed, Equation (16) sums only over $i \neq j$; sampling the diagonal would add variance terms without changing the rate. For comparison with one limiting population task instance,

$$\mathbb{E}\|S_{A_n} - S_W\|_1 \le C n^{-1/2} + \|S_{W_n} - S_W\|_1. \tag{29}$$

Here S_{W_n} includes the discretized signals x_n, y_n, m_n. Their aligned convergence together with kernel cut convergence makes the second term vanish. Obtaining an overall $O(n^{-1/2})$ rate to S_W additionally requires that deterministic approximation term to decay at the same rate.

A.4 Bounded-intervention consequence of sensitivity stability

For any bounded symmetric perturbation V,

$$D\Phi[V] = -\langle S, V\rangle.$$

Therefore, for two aligned sensitivity fields,

$$\big|D\Phi_n[V] - D\Phi[V]\big| = \big|\langle S_n - S, V\rangle\big| \le \|V\|_\infty \|S_n - S\|_1,$$

which gives Equation (18).

More generally, if $V_n \to V$ in L^∞ and $\sup_n \|V_n\|_\infty \le B$, then

$$|\langle S_n, V_n\rangle - \langle S, V\rangle| \le B\|S_n - S\|_1 + \|S\|_1 \|V_n - V\|_\infty.$$

For a finite resolution-dependent candidate family $\{V_n^{(j)}\}_{j=1}^M$ with

$$\sup_{n,j} \|V_n^{(j)}\|_\infty \le B,$$

let $j_{\star,n}$ uniquely maximize $\langle S_{W_n}, V_n^{(j)}\rangle$ for the deterministic mean kernel W_n, and define the decision margin

$$\gamma_n = \min_{j \ne j_{\star,n}} \Big(\langle S_{W_n}, V_n^{(j_{\star,n})}\rangle - \langle S_{W_n}, V_n^{(j)}\rangle\Big).$$

Assume

$$\gamma_* := \inf_n \gamma_n > 0.$$

Each candidate score changes under Bernoulli sampling by at most $B\|S_{A_n} - S_{W_n}\|_1$, so the deterministic ordering is unchanged whenever

$$2B\|S_{A_n} - S_{W_n}\|_1 < \gamma_*.$$

Markov's inequality and Corollary 3 then give

$$\Pr(\widehat{j}_n \ne j_{\star,n}) \le \frac{2B}{\gamma_*}\mathbb{E}\|S_{A_n} - S_{W_n}\|_1 = O(n^{-1/2}).$$

If the benchmark decision is instead defined at a limiting population kernel W, one must additionally control $\|S_{W_n} - S_W\|_1$ and, when the candidate actions vary with resolution, $\|V_n^{(j)} - V^{(j)}\|_\infty$. Symmetry, fixed-diagonal conditions, row conservation, or other linear feasibility restrictions can be incorporated by restricting the candidate family accordingly. The argument does not assert stability of a resolution-dependent nonlinear projection when its active box constraints change.

A.5 L^2 stability and finite-horizon communication dynamics

Finite-to-population coordinate scaling. For cells of masses π_i, direct integration gives

$$\|W_A - W_B\|_2^2 = \sum_{ij} \pi_i \pi_j (A_{ij} - B_{ij})^2, \qquad \frac{\partial J}{\partial A_{ij}} = \pi_i \pi_j G_{ij}. \tag{30}$$

For equal cells, $\pi_i = 1/n$, so

$$\|W_A - W_B\|_2^2 = n^{-2}\|A - B\|_F^2, \qquad \nabla_A J_n = n^{-2} G^{(n)}.$$

Thus a full-matrix finite update multiplies its Euclidean gradient by n^2 to represent the population-scale update

$$W^{k+1} = \Pi_{\mathcal{K}}\big(W^k - \eta G(W^k)\big). \tag{31}$$

If one off-diagonal parameter is shared by the two symmetric matrix entries, the corresponding conversion factor is $n^2/2$; diagonal parameters retain their single-cell weight. These factors convert coordinate areas and do not define a different optimization algorithm.

Fixed-kernel L^2 stability. The residual recursion and bounded activation imply uniform L^∞ bounds on the forward states. For

$$e^\ell = h^\ell_W - h^\ell_{\widetilde{W}},$$

we have

$$z^\ell_W - z^\ell_{\widetilde{W}} = B_0 e^\ell + B_1 D_W e^\ell + B_1 D_{W-\widetilde{W}} h^\ell_{\widetilde{W}}.$$

Lemma 5 and Lipschitz continuity of σ give an L^2 finite-depth recurrence and therefore Lipschitz forward states. The covariance maps are Lipschitz on bounded sets, while the ridge inverse bound and resolvent identity give

$$\|R_W - R_{\widetilde{W}}\|_F \le C_R \|W - \widetilde{W}\|_2.$$

The terminal adjoint is uniformly bounded. Bounded σ'' controls changes in the backward gate, while the operator term splits as

$$D_W(r - \widetilde{r}) + D_{W-\widetilde{W}}\widetilde{r}.$$

Backward induction gives

$$\max_\ell \|p^\ell_W - p^\ell_{\widetilde{W}}\|_2 \le C_p \|W - \widetilde{W}\|_2.$$

The difference-product representation of S_W therefore yields an L^2 Lipschitz bound for the full gradient density. The same estimates accommodate perturbations in $x, y, m, W_{\mathrm{ref}}$ under common bounds and mask mass.

Proof of Theorem 4. Let

$$e_k = \|W^k_n - W^k\|_2.$$

Nonexpansiveness of the projection gives

$$e_{k+1} \le e_k + \eta \|G_n(W^k_n) - G(W^k)\|_2.$$

The fixed-kernel estimates imply

$$\|G_n(W^k_n) - G(W^k)\|_2 \le C(e_k + \varepsilon_n).$$

Hence

$$e_{k+1} \le (1 + \eta C) e_k + \eta C \varepsilon_n.$$

With $a = 1 + \eta C$ and $e_0 \le \varepsilon_n$, induction yields

$$e_k \le \left[a^k + \eta C \sum_{j=0}^{k-1} a^j \right] \varepsilon_n = (2a^k - 1)\varepsilon_n.$$

Therefore

$$C_K = 2(1 + \eta C)^K - 1$$

is valid for every fixed horizon K. The same bounds propagate to states, readouts, adjoints, gradient densities, predictions, and objectives. □

A.6 Solvable two-block example

This analytically solvable specialization makes the sign of the communication derivative concrete; it is not the discrete nonlinear architecture of Theorems 1–4. Split $[0, 1]$ into two equal populations, let $s(u) \in \{-1, +1\}$ be the centered block indicator with $\|s\|_2 = 1$, and set

$$W_{a,b}(u, v) = \begin{cases} a, & u, v \text{ in the same block,} \\ b, & u, v \text{ in different blocks.} \end{cases}$$

Consider the continuous-time linear flow

$$\partial_\tau h = D_{W_{a,b}} h$$

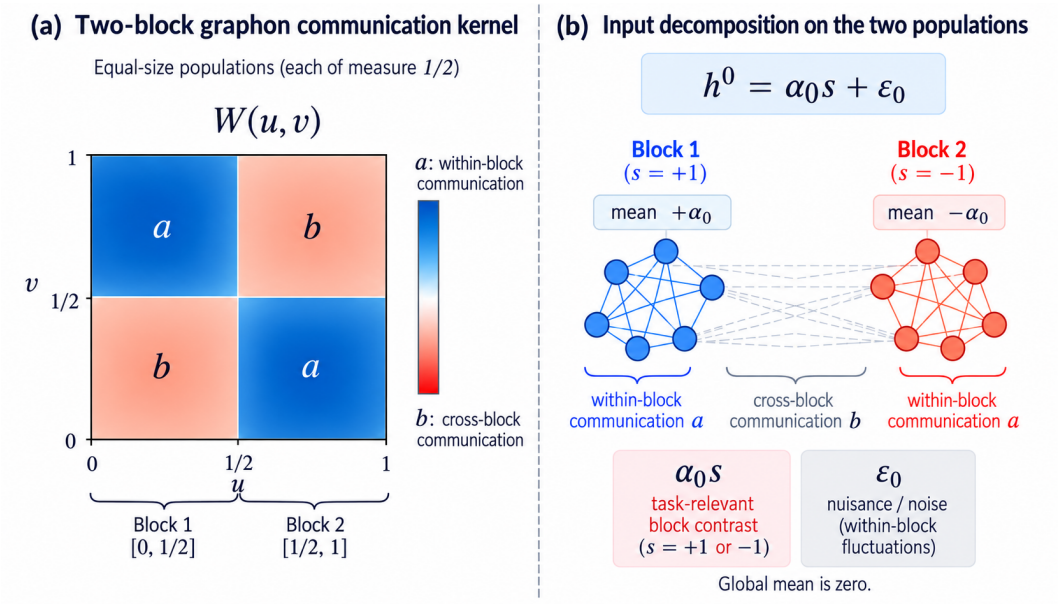


Figure 4: **Two-block graphon example. (a)** Two equal-measure populations with within-block communication a and cross-block communication b. **(b)** The centered input $h^0 = \alpha_0 s + \varepsilon_0$ separates into a task-relevant block contrast $\alpha_0 s$ and within-block nuisance ε_0. Within-block communication can average away nuisance without mixing the two population means, whereas cross-block communication also attenuates the task-relevant contrast.

with centered initial signal

$$h_0 = \alpha_0 s + \varepsilon_0,$$

where ε_0 has zero mean within each block and is therefore orthogonal to both constants and s. The constant, block-contrast, and within-block zero-mean subspaces are invariant, giving

$$h(T) = \alpha_0 e^{-bT} s + e^{-(a+b)T/2} \varepsilon_0. \tag{32}$$

Take the scalar target $y = s$ and fit a scalar ridge readout r using the same normalization convention as the main objective,

$$\Phi_T(r; a, b) = \frac{1}{2} \|r h(T) - s\|_2^2 + \frac{\mu}{2} r^2, \qquad \mu > 0.$$

Define

$$A_T = \alpha_0^2 e^{-2bT}, \qquad N_T = \|\varepsilon_0\|_2^2 e^{-(a+b)T}.$$

Since

$$\langle s, h(T) \rangle = \alpha_0 e^{-bT}, \qquad \|h(T)\|_2^2 = A_T + N_T,$$

the ridge optimum is

$$r_\star = \frac{\alpha_0 e^{-bT}}{A_T + N_T + \mu},$$

and the optimized objective is

$$\Phi_T^\star(a, b) = \frac{1}{2} \frac{N_T + \mu}{A_T + N_T + \mu}. \tag{33}$$

Direct differentiation yields

$$\frac{\partial \Phi_T^\star}{\partial a} = -\frac{T}{2} \frac{A_T N_T}{(A_T + N_T + \mu)^2} \leq 0, \tag{34}$$

$$\frac{\partial \Phi_T^\star}{\partial b} = \frac{T}{2} \frac{A_T (N_T + 2\mu)}{(A_T + N_T + \mu)^2} \geq 0. \tag{35}$$

The first inequality is strict when $T > 0$, $A_T > 0$, and $N_T > 0$; the second is strict when $T > 0$ and $A_T > 0$. Thus increasing within-block communication suppresses nuisance variation without directly attenuating the task-relevant block contrast, whereas increasing cross-block communication suppresses the contrast itself. Even this minimal block model can therefore favor one form of mixing and oppose another.

B Positioning Relative to Graph-Limit Literature

The graph-limit literature contains several notions of transfer and differentiation that are adjacent to, but distinct from, the object studied here. Table 2 summarizes the closest comparisons.

Table 2: **Closest graph-limit results.** A dash means that the corresponding object is not a claimed output of the cited work. The distinction from Yan et al. is the differentiated variable, not an advantage in graph regime or architecture.

Work	Population object	Forward	Adjoints	Param. grad.	Comm. deriv.
Ruiz et al. (Ruiz et al., 2020)	Graphon GNN	✓	–	–	–
Cerviño et al. (Cervino et al., 2023)	Growing-graph training	✓	–	✓	–
Levie (Levie, 2023)	Graphon-signal MPNN	✓	–	–	–
Le–Jegelka (Le & Jegelka, 2023)	Graphop GNN	✓	–	–	–
Yan–Kulick–Tang (Yan et al., 2025)	Graphon-NDE	✓	–	–	–
Yan et al. (Yan et al., 2026)	Sparse random Graphon-NDE	✓	✓	✓	–
Oh et al. (Oh et al., 2024)	Generic graphon functional	–	–	–	generic edge flow
Ours	GNN communication task sensitivity	✓	✓	–	✓

Graphon-GNN and graphon-signal analyses primarily establish that forward computation or prediction has a coherent population limit. Yan et al. (2026) additionally analyze adjoints and gradients with respect to learned neural parameters on sparse random graphs. Their adjoint machinery is closely related to ours, but a neural-parameter gradient and a communication derivative are different maps: perturbing W changes message exchange at every layer and produces a two-variable field indexed by both route endpoints.

Conversely, generic graphon gradient flows (Oh et al., 2024) already provide an optimization geometry for functionals whose variable is a graphon. Our claim is therefore not that graphons themselves provide a new notion of functional differentiation. The contribution is the message-passing-specific forward–backward derivative induced by a task, together with a stability result showing that this structural field remains meaningful under the weak topology appropriate for dense graph sampling.

C Experimental Protocols and Extended Results

C.1 Deterministic resolution study

All theorem-matched synthetic experiments use midpoint coordinates

$$u_i = (i + 1/2)/n, \qquad n \in \{32, 64, 128, 256, 512\},$$

float64 arithmetic, six feature coordinates, two outputs, six residual layers, $\Delta\tau = 0.22$, $\tanh$ activation, ridge parameter $\mu = 0.08$, fidelity coefficient $\lambda = 0.04$, and random seed 7. The supervised term is normalized exactly as in Equation (5), by the supervised weight sum only. The

topology step size is $\eta = 0.035$, trajectories use six projected updates ($k = 0, \dots, 6$), and matched one-step interventions use population-L^2 movement 0.008.

The three communication families are

$$W_{\mathrm{smooth}}(u,v) = 0.46 + 0.18\cos 2\pi(u-v) + 0.07\cos 2\pi(u+v) + 0.05\sin 4\pi u \sin 4\pi v, \quad (36)$$

$$W_{\mathrm{piece}}(u,v) = 0.20 + 0.38\mathbf{1}\{|u-v| \le 0.17\} + 0.16\mathbf{1}\{u+v \ge 1.10\} \quad (37)$$

$$+ 0.10\mathbf{1}\{u, v \in [0.30, 0.70]\}. \quad (38)$$

For the four-block stochastic-block graphon, let $b(u) = \lfloor 4u \rfloor$. The kernel equals 0.74 for the same block, 0.43 for cyclically adjacent blocks, and 0.20 for opposite blocks. Each finite kernel is symmetrized and clipped to $[0.05, 0.95]$. The literal matrix diagonal is set to 0.45; it is inactive for difference aggregation because it multiplies $H_i - H_i = 0$.

The six input coordinates are

$$\big(\sin 2\pi u,\ \cos 2\pi u,\ \sin 4\pi u,\ \cos 4\pi u,\ 2u-1,\ 0.55\sin 6\pi u + 0.25\cos 2\pi u\big),$$

and the targets are

$$y_1(u) = 0.80\sin 2\pi u - 0.35\cos 4\pi u + 0.25\sin 6\pi u + 0.15\sin 2\pi u\cos 4\pi u, \quad (39)$$

$$y_2(u) = 0.60\cos 2\pi u + 0.30\sin 4\pi u - 0.20\cos 6\pi u + 0.10(2u-1)^2. \quad (40)$$

Supervision uses

$$m(u) = 1 \quad \text{on} \quad [0.04, 0.39] \cup [0.50, 0.74] \cup [0.82, 0.97],$$

and zero elsewhere. Let P_+ and P_- denote circular feature-coordinate shifts. The fixed transforms are

$$B_0 = 0.62I + 0.055P_+ - 0.025P_-, \qquad B_1 = 0.82I + 0.060P_- + 0.020P_+,$$

with bias coordinates linearly spaced from -0.035 to 0.035. The reference kernel is

$$W_{\mathrm{ref}} = \mathrm{clip}(0.92W + 0.08\bar{W}, 0, 1)$$

off diagonal, with diagonal reset to 0.45. Here $\bar{W} = \int_{[0,1]^2} W(u,v)\,du\,dv$ denotes the mean kernel value.

Successive $n \to 2n$ comparisons lift coarse fields as step functions to the finer grid. Task-sensitivity comparisons use S_W alone, excluding fidelity, whereas topology trajectories use

$$-G = S_W - \lambda(W - W_{\mathrm{ref}}).$$

For the trajectory kernel metric, lifted coarse diagonal cells are masked because communication there is unobservable for the constant coarse step signal. Smooth and four-block topology trajectories have fitted exponents 0.964 and 0.964; the discontinuous piecewise family gives 0.534.

The closed-form sensitivity and automatic differentiation agree to worst relative error 1.89×10^{-16} across all three families and five resolutions. On a smooth $n = 32$ instance, a central finite-difference directional derivative that refits the ridge readout after each kernel perturbation agrees with the optimized-task derivative to relative error 4.8×10^{-10} and with the full-objective derivative to 1.3×10^{-9}.

Table 3: Successive-resolution task-sensitivity cosine for S_W (excluding fidelity), minimum over the six-update trajectory.

Resolution	Smooth	4-block SBM	Piecewise
$32 \to 64$	0.9803	0.9794	0.9767
$64 \to 128$	0.9981	0.9979	0.9978
$128 \to 256$	0.99968	0.99959	0.99954
$256 \to 512$	0.99959	0.99961	0.99953

C.2 BERNOULLI SAMPLING PROTOCOL

The Bernoulli experiment uses the kernels, signals, masks, feature transforms, ridge readout, and resolutions above. Conditional on the fixed-grid W_n, off-diagonal undirected edges are sampled independently as

$$A_{ij} \sim \mathrm{Bernoulli}((W_n)_{ij})$$

and reflected symmetrically. No additional feature, label, or mask randomness is introduced. The diagonal remains fixed at $A_{ii} = (W_n)_{ii} = 0.45$. Each family and resolution uses 12 independent graph samples.

Sensitivity error is the step-kernel L^1 distance between S_{A_n} and S_{W_n}, excluding the fidelity term. Raw graph error is the step-kernel L^2 distance to W_n.

Table 4: Bernoulli sensitivity L^1 error, mean $\pm$ SD over 12 graph samples.

n	Smooth	4-block SBM	Piecewise
32	0.01254 ± 0.00296	0.01271 ± 0.00124	0.01375 ± 0.00173
64	0.00894 ± 0.00067	0.00931 ± 0.00078	0.01060 ± 0.00121
128	0.00615 ± 0.00052	0.00631 ± 0.00028	0.00711 ± 0.00045
256	0.00442 ± 0.00020	0.00462 ± 0.00024	0.00525 ± 0.00014
512	0.00306 ± 0.00010	0.00325 ± 0.00013	0.00371 ± 0.00014

Relative sensitivity error. Relative error is the sampled L^1 sensitivity error normalized by the deterministic mean-kernel sensitivity norm $\|S_{W_n}\|_1$. At $n = 32$, mean relative errors are 5.829%, 6.110%, and 7.100% for the smooth, four-block, and piecewise families; at $n = 512$ they are 1.360%, 1.492%, and 1.818%. An independent central finite-difference validation with the ridge readout refitted after each perturbation gives relative errors below 2×10^{-10} at $n = 32$.

C.3 CITESEER-CALIBRATED DENSE POPULATION

This experiment uses the class structure and feature distribution of Planetoid CiteSeer to define a dense population problem. The raw graph has 3327 nodes, six classes, and 3703-dimensional bag-of-words features. Class sizes are

$$(264, 590, 668, 701, 596, 508),$$

whose relative sizes define interval masses π.

Let B_{emp} be the empirical symmetric class-to-class edge-frequency matrix. We set

$$B = 0.3 \frac{B_{\mathrm{emp}}}{\max B_{\mathrm{emp}}},$$

which preserves relative block-interaction strengths while giving

$$\pi^\top B \pi = 0.05286954.$$

This deliberately changes the density regime: the experiment is neither a standard sparse CiteSeer benchmark nor an application of the dense theorem directly to the original citation graph.

The 3703-dimensional bag-of-words features are reduced to 16 dimensions by PCA, and population features are sampled with replacement conditional on CiteSeer class. Features, labels, and supervision masks are held fixed across the 12 Bernoulli edge samples at each resolution. The three-layer model uses $\Delta\tau = 0.5$ and $\mu = 0.01$. Feature parameters are trained on the mean-kernel problem at $n = 1024$ and frozen for cross-resolution evaluation. An independent autodiff check at $n = 64$ gives relative error 2.38×10^{-6}.

Across all five resolutions, the fitted log–log exponent for mean sensitivity error is -0.4147 ($R^2 = 0.938$). Restricting to $n \geq 256$ gives -0.4913, consistent with approach to the asymptotic $-1/2$ scaling at larger resolutions. The corresponding limiting raw-noise RMS is

$$\left(\sum_{a,b=1}^{6} \pi_a \pi_b B_{ab}(1 - B_{ab}) \right)^{1/2} = 0.2085444.$$

Table 5: CiteSeer-calibrated Bernoulli results, mean ± SD over 12 samples. Sensitivities are compared with their mean-kernel counterparts at the same resolution.

n	Raw kernel L^2	Sensitivity L^1	Sensitivity cosine
128	0.206656 ± 0.004802	0.022612 ± 0.002672	0.994974 ± 0.001443
256	0.207381 ± 0.001778	0.019259 ± 0.001923	0.997770 ± 0.000726
512	0.207454 ± 0.000564	0.016410 ± 0.001440	0.998558 ± 0.000442
1024	0.208154 ± 0.000278	0.011344 ± 0.000268	0.999333 ± 0.000044
2048	0.208475 ± 0.000225	0.007000 ± 0.000281	0.999833 ± 0.000033

Finite-grid diagonal handling contributes only a vanishing correction.

Connectivity versus topic sensitivity. At the reference resolution, all six diagonal block sensitivities are positive, while 10 of the 15 cross-block entries are negative. Connectivity and topic sensitivity have Pearson correlation 0.8203 and cosine 0.7971 over the 21 upper-triangular entries including the diagonal; the correlation over only the 15 cross-block entries is 0.4598. Thus the strong all-block association partly reflects within-class communication, which is related to both the calibration and classification target. Connectivity is not thereby identified with task utility; the fixed-forward task swap below tests this distinction directly.

C.4 Fixed-forward task-conditioning diagnostic

The diagnostic uses $n = 2048$ and seeds 1103, 2207, 3313, 4421, and 5531. Two orthonormal directions $a, b \in \mathbb{R}^{16}$ define balanced binary teacher targets from within-class feature residuals, thresholded by within-class medians. An orthonormal complement $Q \in \mathbb{R}^{16\times 14}$ removes their span, and both tasks receive the same input XQ. The saved teacher/projection arrays verify negligible projection leakage. Orthogonal teacher directions do not imply statistically independent labels.

Feature maps are trained only on the six-way topic task and selected using its validation objective. They are then frozen before fitting either binary ridge readout. Thus Tasks A and B never influence the learned forward representation. Within each seed, W, X, masks, feature maps, and all forward states are identical. Only the target, optimized readout, terminal adjoint, backward fields, and resulting sensitivity change.

Training uses a maximum of 400 epochs with validation-based model selection. Four seeds select epoch 400; seed 3313 selects epoch 117.

Table 6: Task-conditioning diagnostic. Block sign changes count the 15 distinct cross-population pairs. The cosine is computed on the full sensitivity fields, not on the displayed block averages.

Seed	$\cos(S_A, S_B)$	Block sign changes	Topic acc.	A acc.	B acc.
1103	0.19375	5/15	70.52%	52.99%	51.02%
2207	0.15463	4/15	69.62%	53.07%	54.05%
3313	0.33952	5/15	62.82%	51.43%	54.46%
4421	0.02672	4/15	68.22%	54.55%	50.53%
5531	0.04522	3/15	67.98%	60.28%	54.46%

The mean full-field sensitivity cosine is

$$0.151967 \pm 0.126504.$$

Backward-field cosines are

$$0.110677 \pm 0.074652, \qquad 0.155049 \pm 0.140123, \qquad 0.251679 \pm 0.221798$$

at layers 0, 1, and 2. Block averaging discards within-block variation, so the cosine of block means should not be substituted for the full-field statistic.

Topic accuracy is $67.83\% \pm 2.99\%$ versus $61.41\% \pm 1.72\%$ for the feature-only predictor. The diagnostic binary tasks achieve $54.46\% \pm 3.43\%$ versus $54.63\% \pm 2.66\%$ for A, and $52.91\% \pm 1.96\%$ versus $53.71\% \pm 1.91\%$ for B. These values limit the interpretation to task dependence: they do not establish that the frozen representation or communication is useful for these binary probes.

C.5 CONTROLLED FORWARD/BACKWARD MECHANISM EXPERIMENT

The relational mechanism experiment uses four equal blocks, $n = 512$, feature dimension six, eight residual layers, $\Delta\tau = 0.25$, $\tanh$ activation, $\mu = 0.10$, and five seeds. The initial block communication matrix is

$$Q_0 = \begin{pmatrix} 0.40 & 0.18 & 0.34 & 0.26 \\ 0.18 & 0.40 & 0.26 & 0.34 \\ 0.34 & 0.26 & 0.36 & 0.24 \\ 0.26 & 0.34 & 0.24 & 0.36 \end{pmatrix}.$$

The first feature coordinate has block means $(2, -2, 0, 0)$ with Gaussian noise of scale 0.12; the remaining five coordinates are Gaussian noise of scale 0.30. Labels have block means $(-1, +1, +1, -1)$, with supervision only on blocks 2 and 3. Within each supervised block, examples are split $60\%/20\%/20\%$ into train/validation/test sets. The fixed feature maps are

$$B_0 = 0.55I, \qquad B_1 = 0.95I, \qquad \beta = 0.$$

For each layer and route, let Δh^ℓ and Δr^ℓ denote the forward and backward endpoint differences. Up to a common positive scale, the controls are

$$F_{uv} = \sum_\ell \Delta\tau \|\Delta h^\ell_{uv}\|_2^2, \tag{41}$$

$$B_{uv} = \sum_\ell \Delta\tau \|\Delta r^\ell_{uv}\|_2^2, \tag{42}$$

$$M_{uv} = \sum_\ell \Delta\tau \|\Delta h^\ell_{uv}\|_2 \|\Delta r^\ell_{uv}\|_2. \tag{43}$$

Scores are averaged into the symmetric 4×4 block-action coordinates before intervention. Both global signs of each simple control are considered; the control and sign are selected on validation loss only and then evaluated on the held-out test set. Population-L^2 movement budgets are

$$0.002165, \qquad 0.010825, \qquad 0.021651.$$

For the row-conserving feasible set, write the off-diagonal symmetric action as a vector z and let B be the unsigned incidence matrix mapping edge changes to row-sum changes. The orthogonal tangent projection is

$$Pz = z - B^\top (BB^\top)^{-1} Bz,$$

with the diagonal fixed. Tentative changes also obey

$$-W_{ij} \le \delta_{ij} \le 1 - W_{ij}.$$

We use Dykstra projection onto the intersection of the affine row-sum constraint and box, followed by scalar bisection to match the requested population-L^2 movement. The box constraint is inactive at all reported budgets; maximum row-sum residuals are of order 10^{-14}. The analytic sensitivity and autodiff agree to relative error approximately 1.5×10^{-16}.

Table 7: Relational task under the row-conserving feasible set. Negative ΔMSE indicates improvement; mean $\pm$ SD across five seeds.

Movement	Full sensitivity	Best validation-selected simple control
0.00217	-0.001231 ± 0.000145	-0.000103 ± 0.000025
0.01083	-0.007015 ± 0.000718	-0.000519 ± 0.000132
0.02165	-0.016129 ± 0.001414	-0.001050 ± 0.000275

The reduced-cue variant changes the useful/harmful source–target initialization from $0.34/0.26$ to $0.305/0.295$ and preserves the same qualitative separation. This experiment is a controlled existence test rather than an estimate of how often route-specific angular coupling matters in natural datasets.

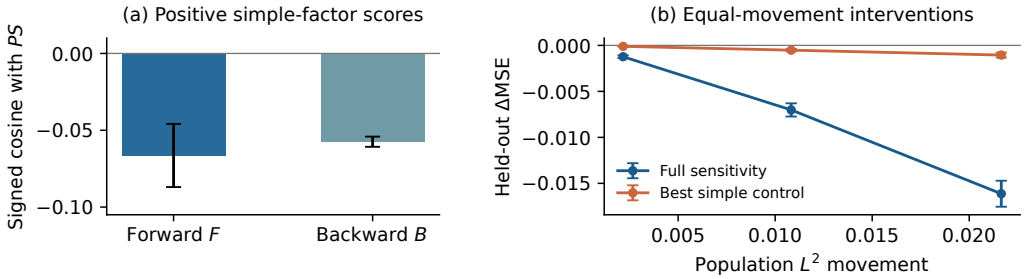


Figure 5: **Route-specific alignment in the controlled relational task.** Left: signed cosine of positive forward and backward controls with the full feasible sensitivity. Right: held-out loss changes at equal movement for the full sensitivity and validation-selected best simple control. Error bars are SD across five seeds.

C.6 METR-LA TRANSFER DETAILS

Data, nesting, and predictor. METR-LA contains 207 sensors. We construct a deterministic farthest-point ordering from the geographic-coordinate centroid and use the first 64, 128, and 207 sensors as nested resolutions. Pairwise coordinates are converted to local kilometers. The physical communication kernel is

$$W^0_{ij} = \exp\left(-\frac{d^2_{ij}}{2\sigma^2}\right),$$

where σ is the 20th percentile of nonzero pairwise distances and the diagonal is one.

Each example uses 12 historical steps and predicts the next 12. The chronological split is $70\%/10\%/20\%$ train/validation/test. Zero or missing measurements are masked; observed values are standardized using training-set statistics. Missing inputs use causal forward-fill, with leading gaps replaced by the sensor's training mean or, if unavailable, the global training mean.

The predictor uses a temporal encoder

$$\mathrm{Linear}(12, 32)\text{–GELU–}\mathrm{Linear}(32, 32)\text{–GELU},$$

followed by a four-unit-time spatial flow implemented with 16 Euler difference-diffusion steps ($\Delta\tau = 0.25$) and a linear $32 \to 12$ forecast head. Models are trained for 10 epochs with AdamW using learning rate 10^{-3}, weight decay 10^{-4}, batch size 64, and a 10^{-4} head ridge penalty. The validation-selected checkpoint is used.

Source and target models are trained independently, and all ordinary model parameters are frozen during communication interventions. This experiment therefore lies deliberately outside Theorem 4.

Sensitivity, lifting, and feasible movement. The source score is the negative, symmetrized training-MSE derivative with respect to W, converted to population coordinates by the n^2 factor and with diagonal zero. Let P be a row-normalized Gaussian interpolation matrix from target sensors to source sensors using the same physical distance scale σ, with identity rows on source sensors. We define

$$S_{\mathrm{lift}} = PSP^\top,$$

then symmetrize and zero the diagonal.

Before intervention, score entries are clipped to $[-5, 5]$. Let z be the off-diagonal symmetric coordinate vector and B its unsigned incidence matrix. The zero-row-sum tangent projection is

$$P_{\mathrm{row}} z = z - B^\top (BB^\top)^{-1} Bz.$$

After scaling, Dykstra projection jointly enforces this affine constraint,

$$0 \le W + \Delta W \le 1,$$

and a fixed diagonal; scalar bisection matches the target population RMS movement $n^{-1}\|\Delta W\|_F$. Unlike the controlled-routing experiment, box constraints may become active. All compared directions use the same feasibility operator and target-specific movement. For $128 \to 207$, movements

range from 0.012021 to 0.012106, with mean 0.012066. The native target gradient determines only this common movement budget, never the transferred direction.

Five independently trained source models and five independently trained target models are used. Primary summaries exclude equal-seed source–target pairs, yielding 20 source–target comparisons. Because these pairs reuse the same five frozen targets, uncertainty is summarized by descriptive SD across targets rather than by treating all 20 pairs as independent. Baseline test MAE is

$$6.07818 \pm 0.02219 \text{ mph}.$$

The $128 \to 207$ transferred direction improves MAE by

$$0.04149 \pm 0.00049 \text{ mph},$$

approximately 0.68% of baseline.

Native sensitivity and intervention comparison. The native direction is computed from the frozen 207-sensor target's own training objective. Different-seed source–target pairs give score cosines

$$0.747839 \pm 0.004661 \qquad (64 \to 207),$$

and

$$0.903555 \pm 0.000787 \qquad (128 \to 207).$$

Corresponding correlations are 0.599889 and 0.856506, while relative L^2 discrepancies are 0.665537 and 0.428565. These quantities measure score agreement before intervention constraints.

For both native and transferred interventions, the target-specific budget is the native one-step population RMS movement. "Without row conservation" still enforces symmetry, box bounds, and a fixed diagonal. For each source resolution, equal-seed pairs are excluded; the four remaining sources are averaged within each target, followed by mean and sample SD across the five targets. Fractions of native gain are computed per target before aggregation.

The unconserved fractions of native gain are

$$83.61\% \pm 0.46\% \quad \text{and} \quad 90.41\% \pm 0.49\%$$

for source sizes 64 and 128. Under row conservation, they are

$$42.91\% \pm 0.16\% \quad \text{and} \quad 69.70\% \pm 0.30\%.$$

Actionable geographic controls. Geographic controls are fit to the already clipped and row-projected source score, so they approximate the action that can actually be taken rather than an unprojected score that feasibility later changes. The distance-only family uses 24 radial RBFs with knots spanning observed source distances. The richer spatial family augments these with symmetrized products of node RBFs centered at farthest-point geographic centers.

For $n = 128$, sensor-held-out three-fold cross-validation selects among 8, 16, or 32 centers; bandwidth multipliers

$$\{0.35, 0.7, 1.4, 2.8, 5.6\};$$

and ridge penalties

$$\{10^{-5}, 10^{-3}, 10^{-1}, 10\}.$$

Held-out-node incident edges are excluded from the corresponding fitted projector.

At $128 \to 207$, MAE reductions are

$$0.02528 \pm 0.00031 \text{ mph}$$

for the direct geographic control,

$$0.02143 \pm 0.00032 \text{ mph}$$

for the lifted geographic control, and

$$0.00985 \pm 0.00027 \text{ mph}$$

for the distance-only lifted control. Twenty within-distance-stratum shuffles of the residual around the actionable geographic fit reduce the gain to

$$0.00367 \pm 0.00012 \text{ mph}.$$

This establishes structure beyond the specified smooth geographic family; it does not imply that the residual is intrinsically non-geographic.

Coupling ablations. The layerwise controls use the same endpoint differences as the exact sensitivity: forward-only accumulates $\|\Delta h\|^2$, backward-only accumulates $\|\Delta r\|^2$, and the magnitude product uses $\|\Delta h\|\,\|\Delta r\|$. Source scores are RMS-normalized before lifting.

For the magnitude-product control, the global sign is selected independently using only each frozen source model's validation MAE. At the predeclared source-side population-RMS movement 0.01, all five $n = 64$ and all five $n = 128$ checkpoints select the negative orientation; the choice is unchanged at movements 0.005 and 0.015. The mean validation-MAE separation between the selected and rejected signs is 0.07338 mph for $n = 64$ and 0.07764 mph for $n = 128$. Applying these independently selected signs after transfer gives MAE reductions of 0.02408 ± 0.00024 mph for $64 \to 207$ and 0.04079 ± 0.00057 mph for $128 \to 207$. The latter is 98.3% of the full transferred-sensitivity gain.

For reference, orienting the magnitude score by its alignment with the full source sensitivity selects the same negative sign on all ten source checkpoints. Thus the strong magnitude-control result is not an artifact of borrowing the orientation of the full sensitivity.

C.7 Reproducibility and supplementary materials

The theoretical assumptions and proofs are given in Appendix A, and the experimental protocols are specified in the preceding subsections. The supplementary ZIP contains experiment code, saved numerical records, model checkpoints where available, and scripts for reconstructing the reported summaries. Its top-level README maps each experiment to its code and results, lists dependencies and data requirements, and provides replay commands.

Synthetic and controlled-routing experiments. The package includes drivers for derivative certification, deterministic cross-resolution comparisons, finite-horizon communication updates, Bernoulli sampling, and controlled routing. Saved records include all 180 synthetic Bernoulli samples, derivative diagnostics, and seed-level routing interventions. The aggregation scripts check validation-based selection of the simple controls and reconstruct the held-out comparisons. A separate NumPy calculation reconstructs the reference sensitivity norms and checks directional derivatives by finite differences with the ridge readout refitted after each perturbation.

CiteSeer-calibrated experiments. The original notebook provides the class-aligned calibration, dense population construction, feature processing, training, and task-conditioning generators. The package retains all 60 Bernoulli measurements and the final five-seed task-conditioning records, including block sensitivities, teacher/projection arrays, training histories, and backward-field summaries. The historical full-resolution sensitivity arrays, trained checkpoints, and exact sampled feature/mask arrays for the final task-conditioning runs were not saved. Consequently, block-sign statistics and reported aggregate measurements can be checked from the archive, but the historical full-field cosines cannot be independently reconstructed from retained arrays alone. The notebook supports rerunning the protocol with resampling and retraining; bit-for-bit reproduction of those completed runs is not claimed.

METR-LA experiments. The package includes the code and configurations, 15 independently trained checkpoints across the three sensor resolutions, native scores, interpolation matrices, lifted fields, intervention kernels, model histories, and source–target evaluation records. Geographic, coupling, magnitude, and temporal diagnostics are also provided. The original fitted geographic-control binaries are not retained; their evaluation records and fitting code are included.

Independent magnitude-sign selection. The validation-only sign-selection script is included together with all 30 source-checkpoint/budget records, covering ten checkpoints at movements 0.005, 0.01, and 0.015. At the primary movement 0.01, the sign is selected from each frozen source model's validation MAE, without consulting the full sensitivity or target data. The package retains the 40 selected different-seed target evaluations and reconstructs the summaries by averaging sources within each target before reporting the mean and sample standard deviation over five targets. The earlier full-sensitivity-based sign diagnostic is retained separately and is not the sign-selection protocol used for the primary magnitude comparison.

Verification, data, and replay. File hashes and nested archives are checked by `scripts/verify_package.py`. The aggregation script `scripts/reaggregate.py` reconstructs the principal numerical summaries, including the validation-selected magnitude comparison. Figure-generation code reconstructs the experimental plots from saved records. These checks require no raw datasets or model retraining; they reaggregate completed evaluations rather than rerunning prediction inference. Full dataset-based replay requires public CiteSeer or METR-LA data, as appropriate. Raw datasets are not redistributed. The METR-LA validation-sign replay checks the HDF5 checksum and uses the archived sensor design, so it does not require the sensor-location CSV. Other METR-LA replay commands and their data requirements are documented in the README. Ten regenerable layerwise component arrays are omitted to meet the upload limit; aggregated source scores, lifted directions, diagnostics, and regeneration code are retained. Software dependencies and the recorded METR-LA training environment are documented; exact numerical agreement across other software versions and devices is not guaranteed.