EXECUTABLE NEURAL OPERATORS: ROUTING LOCAL PHYSICAL MECHANISMS IN CONTINUOUS SPATIAL DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

While Neural Operators and Physics-Informed Neural Networks (PINNs) have accelerated the simulation of continuous spatial data, they predominantly rely on global implicit mappings. This monolithic representation can entangle local physical rules, rendering models brittle to non-stationary mutations in complex media, such as sudden changes in boundary conditions or unobserved subsurface scatterers. To address this limitation, we propose the Executable Neural Operator (ENO), a novel architecture that represents localized physical mechanisms within continuous fields. By leveraging a unified bipartite graph embedding, ENO tokenizes large-scale continuous spatial data into a routed graph. Instead of global spectral mixing, ENO employs Dynamic Operator Routing (DOR) to assign localized neural programs to simulate wave propagation and energy transfer. During medium property mutations, our Targeted Mechanism Revision (TMR) protocol provides gradient isolation under explicit locality and parameter-ownership conditions. We extensively evaluate ENO on high-contrast 3D magnetotelluric dynamics. Results demonstrate that ENO achieves competitive zero-shot accuracy on unseen out-of-distribution geometries while suppressing the physics violation rate and reducing background prediction drift, establishing a theoretical and empirical basis for mechanism discovery in non-stationary physical systems.

1 INTRODUCTION

Predicting complex continuous spatial dynamics is a fundamental challenge in scientific machine learning. Recent breakthroughs have demonstrated the robust capabilities of deep three-dimensional neural networks in large-scale applications, most notably in achieving state-of-the-art accuracy in medium-range global weather forecasting (Bi et al., 2023; Lam et al., 2023). At the core of these algorithmic advancements is the paradigm of operator learning, which seeks to approximate mappings between infinite-dimensional function spaces rather than fitting finite-dimensional spaces (Lu et al., 2021; Kovachki et al., 2023).

Standard neural operators, such as the Fourier Neural Operator (FNO) (Li et al., 2020), approach this by employing global integral kernels. Although highly effective for stationary coefficient distributions, global spectral mixing imposes an implicit assumption of causal density. Consequently, when faced with real-world complexities such as free boundary problems (Long et al., 2026) or abrupt out-of-distribution (OOD) subsurface scatterers, these monolithic representations struggle. Recent theoretical and empirical studies have formalized the "forgetting hypothesis," revealing that deep operator learning inherently suffers from geometry and local feature forgetting due to global entanglement (Xia & Aviles-Rivero, 2026). Because local physical rules are inextricably entangled across the entire frequency domain, targeted online adaptation to non-stationary mutations becomes impossible without catastrophic forgetting of the background physical laws.

To address this fundamental limitation, we propose the **Executable Neural Operator (ENO)**. ENO abandons the monolithic global kernel in favor of a topologically localized, conditionally routed mechanism library. By leveraging a unified bipartite graph embedding, ENO tokenizes large-scale continuous spatial data into a routed graph. Instead of global spectral mixing, ENO employs Dynamic Operator Routing (DOR) to assign localized neural programs to simulate wave propagation

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.

and energy transfer. During medium property mutations, our Targeted Mechanism Revision (TMR) protocol provides gradient isolation under explicit locality and parameter-ownership conditions.

In summary, our main contributions are:

- **Causal Disentanglement via Bipartite Tokenization:** We introduce an architecture that discretizes continuous functional spaces into routed local domains, supported by a conditional Sobolev approximation bound.
- **TMR:** We propose an online adaptation protocol that isolates local-loss gradients under explicit locality, sparse-dispatch, and parameter-ownership conditions.
- **State-of-the-Art Zero-Shot Generalization:** We extensively evaluate ENO on complex 3D geophysical electromagnetic inversion tasks. Results demonstrate that ENO achieves rigorous accuracy parity with global spectral methods on unseen OOD anomalies, while suppressing the physics violation rate (PVR) by an order of magnitude.

2 Related Work

Neural Operators and Physics-Informed Learning. Physics-Informed Neural Networks (PINNs) established a foundational deep learning framework for solving nonlinear PDEs by embedding physical constraints directly into the loss function (Raissi et al., 2019; Karniadakis et al., 2021). This paradigm has been expanded to enable data-free time-domain simulations (Stiasny et al., 2021) and has been structurally enhanced via self-attention Transformer architectures (Zhao et al., 2023). In parallel, neural operators shift the focus from learning a single solution to approximating the general PDE solution operator family (Lu et al., 2021; Kovachki et al., 2023). The FNO (Li et al., 2020) remains a dominant baseline, inspiring structural variants like Factorized FNO for computational efficiency (Tran et al., 2021) and Laplace Neural Operators for capturing transient dynamics (Cao et al., 2024). Bridging these two directions, Physics-Informed Neural Operators (PINO) (Li et al., 2024) and Latent Neural Operators (Kontolati et al., 2024; Karumuri et al., 2026) enforce physical laws within the learned latent spaces to enable robust real-time predictions. However, these models inherently rely on global parameter sharing, rendering them brittle to localized physical mutations.

Geometry-Aware and Graph-Based PDE Solvers. To accommodate complex geometries and irregular boundaries, message passing neural PDE solvers (Brandstetter et al., 2022) and bi-stride multi-scale graph neural networks (Cao et al., 2023) treat the spatial domain as a discrete graph. Recent advancements successfully integrate continuous operator learning with graph representations, yielding highly scalable models such as the Geometry-Informed Neural Operator (GINO) (Li et al., 2023), its equivariant extension EqGINO (Kim et al., 2026), and the Geometry Aware Operator Transformer (GAOT) (Wen et al., 2026). While these architectures effectively scale to arbitrary 3D domains, they still operate under a static message-passing or homogeneous attention paradigm that does not explicitly disentangle the underlying physical causal mechanisms, limiting their adaptability to heterogeneous structures.

Non-Stationary Adaptation and Catastrophic Forgetting. Real-world physical systems, such as the magnetotelluric transfer functions captured in the public EarthScope USArray dataset (Schultz et al., 2006–2018), exhibit extreme spatial non-stationarity. Adapting neural surrogates to such abrupt mutations without degrading invariant background accuracy remains a critical open challenge. The recently formalized "forgetting hypothesis" identifies how neural operators lose geometric and local physical priors during deep transformations due to representation entanglement (Xia & Aviles-Rivero, 2026). Although hybrid latent representations have been proposed to separate distinct emulation factors in PDEs (Bekar et al., 2026), they fall short of providing a strict analytical gradient firewall. In contrast, our approach explicitly isolates causal mechanisms via dynamic routing, enabling zero-interference adaptation to local anomalies.

3 EXECUTABLE NEURAL OPERATORS (ENO)

To model continuous spatial dynamics with explicit local dependencies, ENO combines a bipartite tokenization with routed local operators. The following statements distinguish approximation capacity from physical causality and state the assumptions required for each guarantee.

3.1 BIPARTITE TOKENIZATION OF FUNCTION SPACES

We consider a physical system defined on a bounded open domain $\Omega \subset \mathbb{R}^d$ (typically $d \in \{2, 3\}$). Let $\mathcal{A}$ and $\mathcal{U}$ be separable Banach spaces. The system is governed by a generalized parametric PDE:

$$\mathcal{L}_a(u) = f \quad \text{in } \Omega, \quad \mathcal{B}(u) = g \quad \text{on } \partial\Omega, \tag{1}$$

where $a \in \mathcal{A}$ represents the physical coefficient field (e.g., permittivity, conductivity, or velocity model), $u \in \mathcal{U}$ is the target solution field (e.g., electromagnetic wavefield or fluid velocity), f is the forcing term, and $\mathcal{B}$ denotes the boundary conditions. Let v denote the concatenated known input fields, including a, f, and boundary-condition features. Our goal is to learn the solution operator $\mathcal{F} : \mathcal{V} \to \mathcal{U}$ such that $u = \mathcal{F}(v)$.

Instead of discretizing Ω via a uniform Cartesian grid or employing global Fourier transforms, we topologically tokenize the continuous field using a Unified Bipartite Graph Embedding. This allows us to map the infinite-dimensional operator learning problem into a discrete, localized message-passing framework with an explicit discretization error under the assumptions below.

Continuous-to-Discrete Projection. We construct a spatial bipartite graph $\mathcal{G} = (\mathcal{V}_{obs}, \mathcal{V}_{mech}, \mathcal{E})$. At inference time, the encoder receives only quantities available from the problem specification. We write $v(x)$ for the concatenated known input channels (for example, coefficient, forcing, and boundary-condition features); the target solution u is not part of v.

- **Observation Nodes ($\mathcal{V}_{obs}$):** A set of N coordinate points $\{x_i\}_{i=1}^N \subset \Omega$ where the known input field $v(x)$ is sampled. The target solution $u(x)$ is used only for supervision during training.
- **Mechanism Nodes ($\mathcal{V}_{mech}$):** A set of M latent spatial anchors $\{y_j\}_{j=1}^M \subset \Omega$, generated via farthest point sampling (FPS) or physics-informed clustering, serving as the computational hubs for local operator execution.

An edge $e_{ij} \in \mathcal{E}$ exists between $x_i \in \mathcal{V}_{obs}$ and $y_j \in \mathcal{V}_{mech}$ if and only if $\|x_i - y_j\|_2 \leq R$, where R is a strictly defined local support radius.

We define the continuous-to-discrete encoder projection $\mathcal{P}_{enc} : \mathcal{V} \to \mathbb{R}^{M \times d_{latent}}$ as a localized integral transform over the neighborhood ball $B(y_j, R)$:

$$\mathbf{h}_j^{(0)} = \int_{\Omega \cap B(y_j, R)} E_{enc}(y_j, x, v(x))\, dx \approx \sum_{x_i \in \mathcal{N}(y_j)} w_i E_{enc}(y_j, x_i, v(x_i)), \tag{2}$$

where E_{enc} is a learned vector-valued feature map, $w_i \geq 0$ are quadrature weights, and $\nu_N = \sum_i w_i \delta_{x_i}$ is the corresponding discrete measure. Thus the sum is exactly integration against ν_N, whereas approximation refers to Lebesgue integration. All neighborhood integrals below are intersected with Ω.

Theoretical Guarantee of Local Tokenization. A critical concern in mapping global PDEs to localized bipartite graphs is the potential loss of approximation expressivity, particularly for differential operators that exhibit non-local dependencies in their Green's functions. Under the regularity and sampling assumptions stated below, we establish a conditional Sobolev error bound for the tokenization.

Theorem 1 (Approximation Bound of Bipartite Tokenization in Sobolev Spaces)**.** *Let Ω be a bounded Lipschitz extension domain, $s \in \mathbb{N}$, $s > d/2$, and $\|v\|_{H^s} \leq 1$. Write $q(y) = \kappa(\cdot, y)$ and assume q has a C^2 extension to a neighborhood of $\mathrm{conv}(\bar{\Omega})$ with values in $H^s(\Omega)$ and bounded second derivative. Assume nonnegative quadrature weights with $\sum_i w_i \leq W$ and the operator-valued quadrature bound in Appendix B. For a partition-of-unity cover of radius R, tokens retaining zeroth*

and first local moments, and the constructive decoder and communication assumptions detailed in the appendix, for every $\epsilon > 0$ there exist maps such that

$$\sup_{\|v\|_{H^s} \le 1} \|\mathcal{K}v - \mathcal{P}_{dec} \circ \Phi \circ \mathcal{P}_{enc}(v)\|_{H^s} \le C_N N^{-s/d} + C_R R^2 + \epsilon. \tag{3}$$

The constants are independent of N, R; network size and communication depth may depend on them and on ϵ.

The proof and its assumptions are provided in Appendix B.

3.2 Dynamic Operator Routing for PDEs

To resolve this, we introduce **DOR**. We replace the monolithic global operator with a finite, discrete library of localized neural mechanisms, dynamically routed based on the topological and physical properties of the local coefficient field.

The Local Mechanism Library. We define a mechanism library $\mathcal{P} = \{\mathcal{K}_{\theta_1}, \mathcal{K}_{\theta_2}, \dots, \mathcal{K}_{\theta_K}\}$, where each $\mathcal{K}_{\theta_k}$ acts as a local transition operator. Unlike standard MLPs, each mechanism is a localized graph neural operator operating over the message-passing neighborhood $\mathcal{N}(y_j)$ in the mechanism node space $\mathcal{V}_{mech}$. Let $\mathbf{h}_j^{(t)} \in \mathbb{R}^{d_{latent}}$ be the latent state at anchor y_j at depth t; $H^{(t)}$ denotes all tokens. Depth is not physical time unless a consistent time discretization is separately specified. The execution of mechanism k computes the latent state update:

$$\mathcal{K}_{\theta_k}(H^{(t)})_j = \sigma\left(\mathbf{W}_k \mathbf{h}_j^{(t)} + \sum_{y_m \in \mathcal{N}(y_j)} \kappa_{\theta_k}(y_j, y_m)\mathbf{h}_m^{(t)}\right), \tag{4}$$

where κ_{θ_k} is a learnable continuous message-passing kernel parameterized by θ_k, $\mathbf{W}_k$ is a linear projection, and σ is a non-linear activation function.

Physically Conditioned Hard Routing. The selection of which physical mechanism executes at node y_j must be conditioned on the local physical coefficient field $a(x)$. We define a localized coefficient pooling vector $\mathbf{a}_j = \int_{\Omega \cap B(y_j,R)} w_j(x) a(x)\, dx$, with $w_j \ge 0$ and $\int_{\Omega \cap B(y_j,R)} w_j = 1$, representing pooled local coefficient features.

A lightweight gating network $h_\phi(\mathbf{h}_j^{(t)}, \mathbf{a}_j) \to \mathbb{R}^K$ outputs the routing logits $\mathbf{z}_j^{(t)} = h_\phi(\mathbf{h}_j^{(t)}, \mathbf{a}_j)$. We use a hard forward decision with a straight-through backward surrogate:

$$p_{j,k}^{(t)} = \frac{\exp((z_{j,k}^{(t)} + \epsilon_{j,k}^{(t)})/\tau)}{\sum_{m=1}^K \exp((z_{j,m}^{(t)} + \epsilon_{j,m}^{(t)})/\tau)}, \qquad \hat{g}_{j,k}^{(t)} = \mathbf{1}_{\{k = \arg\max_m p_{j,m}^{(t)}\}}. \tag{5}$$

where the $\epsilon_{j,k}^{(t)}$ are independent $\mathrm{Gumbel}(0,1)$ variables and $\tau > 0$. Ties have a fixed rule; deterministic inference omits noise. The surrogate is $g^{ST} = \mathrm{stopgrad}(\hat{g} - p) + p$: temperature affects its gradient, not the hard argmax. The latent depth update is computed via a sparsely routed sum:

$$\mathbf{h}_j^{(t+1)} = \mathbf{h}_j^{(t)} + \Delta t \sum_{k=1}^K \hat{g}_{j,k}^{(t)} \cdot \mathcal{K}_{\theta_k}(H^{(t)})_j. \tag{6}$$

Sparse forward dispatch evaluates only the selected mechanism. The full mixture straight-through router derivative requires all expert outputs, or an explicitly specified alternative estimator; selected-only execution does not supply that derivative. TMR fixes routes and freezes the router. The step size Δt alone gives no stability guarantee (Appendix D.8).

Theoretical Guarantee: Finite Covering of Local Green's Operators. A critical mathematical question arises: Can a *finite* discrete library of K mechanisms accurately approximate the local PDE transitions in a heterogeneous medium? For an explicitly specified compact coefficient class, the associated local Green's operators admit an ϵ-finite covering and can therefore be approximated by a finite mechanism library.

Theorem 2 (Finite Covering of Local Green's Operators)**.** *On the fixed ball $D = B(0, R)$, let $\mathfrak{A}_R \subset L^\infty(D)$ be compact with $0 < \lambda \le a \le \Lambda$. Equip $H = H_0^1(D)$ with $\|u\|_H = \|\nabla u\|_{L^2}$ and let $H^{-1} = H^*$. The Dirichlet inverse $\mathcal{G}^a = (-\nabla \cdot (a\nabla))^{-1}$ satisfies*

$$\|\mathcal{G}^a - \mathcal{G}^b\|_{\mathcal{L}(H^{-1},H)} \le \lambda^{-2}\|a - b\|_{L^\infty}. \tag{7}$$

Consequently $\{\mathcal{G}^a : a \in \mathfrak{A}_R\}$ has finite operator-norm covers. On any compact forcing set $F \subset H^{-1}$, if the mechanism class uniformly approximates each canonical inverse on F, then for every $\epsilon > 0$ a finite neural library satisfies

$$\sup_{a\in\mathfrak{A}_R} \min_{1\le k\le K} \sup_{f\in F} \|\mathcal{G}^a f - \mathcal{K}_{\theta_k}(f)\|_H < \epsilon. \tag{8}$$

Interior balls are translated to D; boundary patches require separate uniformly controlled domain maps. Neural approximation is asserted on F, not on the whole H^{-1} unit ball.

The complete covering argument is provided in Appendix C.

Theorem 2 provides a conditional justification for DOR.

3.3 Targeted Mechanism Revision for Non-Stationary Media

In real-world scientific applications, physical media are rarely stationary. Consider a geophysical inversion scenario where an unexpected subsurface cavity or a sudden material ablation introduces a localized, out-of-distribution (OOD) mutation to the coefficient field: $a(x) \mapsto a(x) + \delta a(x)$, where $\text{supp}(\delta a) = \Omega_{mut} \subset \Omega$.

A local coefficient mutation can alter a global solution. TMR restricts parameter updates through explicit ownership and frozen computation paths; its neural dependency region must not be identified with a physical causal cone without a separate argument.

The Gradient Firewall for Continuous Fields. During an online adaptation phase, the mutated observations generate a localized residual field $e(x,t) = \|\hat{u}(x,t) - u_{true}(x,t)\|^2$. Projected onto the mechanism nodes $y_j \in \mathcal{V}_{mech}$, the local loss is $\mathcal{L}_j$.

For fixed routes, write $J_{j,k}^{(t)} = \partial \mathcal{K}_{\theta_k}(H^{(t)})_j / \partial \theta_k$ with incoming states held fixed and let $\boldsymbol{\eta}_j^{(t+1)} = \nabla_{\mathbf{h}_j^{(t+1)}} \mathcal{L}$ be the total loss adjoint. The direct contribution is

$$\nabla_{\theta_k} \mathcal{L}|_{\text{direct},t} = \Delta t \sum_j \hat{g}_{j,k}^{(t)} (J_{j,k}^{(t)})^\top \boldsymbol{\eta}_j^{(t+1)}. \tag{9}$$

The full derivative sums contributions over every use of a shared parameter. A local routing zero removes only that execution's direct contribution; earlier message paths and trainable gates may remain. TMR therefore fixes routes, freezes shared encoder/decoder parameters and optimizer state for protected blocks, and updates only explicitly owned anomaly parameters. Anomaly detection and dedicated routing are protocol requirements, not consequences of the formula.

Theoretical Guarantee: Architecture-Conditioned Sobolev Error Isolation. A coefficient perturbation and a neural parameter update have different domains of influence. The next theorem states the two guarantees separately.

Theorem 3 (Architecture-Conditioned Sobolev Error Isolation)**.** *Let $a, \tilde{a} = a + \delta a \in W^{1,\infty}(\mathbb{R}^d)$ be time-independent, with $0 < \lambda \le a, \tilde{a} \le c_{max}^2$ and $\text{supp}\, \delta a \subset \bar{B}(x_0, r_0)$. Wave solutions with identical data in $H^1 \times L^2$ satisfy $\tilde{u} = u$ almost everywhere outside $\mathcal{C}(t) = \bar{B}(x_0, r_0 + c_{max}t)$. Separately, fix a finite neural computation graph and its routes. If no trainable adapted parameter θ_A has a directed path to a local prediction $\hat{u}_j$, then*

$$\nabla_{\theta_A} \mathcal{L}_j = 0, \qquad \hat{u}_j(\theta'_A, \theta_B) = \hat{u}_j(\theta_A, \theta_B), \qquad \theta'_B = \theta_B, \tag{10}$$

where the last equality follows from explicit freezing. This applies outside $\mathcal{C}(T)$ only if every such output's entire dependency graph avoids adapted parameters; locality of individual edges alone is insufficient.

Appendix D proves both claims. Frozen parameters may have nonzero mathematical loss derivatives. The wave result does not apply to the frequency-domain magnetotelluric system; architectural isolation is evaluated on unchanged inputs, not inferred from finite propagation.

3.4 PHYSICS-INFORMED META-ADAPTATION FOR CAUSAL DISENTANGLEMENT

A critical challenge remains in the offline pre-training phase: How do we force the routing network h_ϕ to discover genuinely orthogonal, physically meaningful mechanisms, rather than arbitrarily partitioning the data into entangled, overlapping sub-domains?

To resolve this, we formulate a bilevel objective that combines PINN constraints (Raissi et al., 2019) with an episodic meta-intervention objective. Under the realizability and observational-support assumptions stated below, this combination yields a conditional identifiability result.

Localized Physics-Informed Regularization. Standard data-driven neural operators suffer from spectral bias and often violate fundamental conservation laws when interpolating continuous fields. Because ENO operates on a bipartite graph, we explicitly enforce the governing PDE $\mathcal{L}_a(u) = f$ on the continuous field reconstructed from the latent mechanism tokens.

Let $\mathcal{P}_{dec}$ be the discrete-to-continuous decoder, with range in the domain of the PDE residual. For a given time step t, the predicted continuous field is $\hat{u}^{(t)}(x) = \mathcal{P}_{dec}(\{\mathbf{h}_j^{(t)}\})$. We define the localized physics-informed residual over the neighborhood of each mechanism node y_j:

$$\mathcal{R}_{PDE}(y_j; \boldsymbol{\theta}, \phi) = \int_{B(y_j,R)} \left| \mathcal{L}_a \left(\hat{u}^{(t)} \right)(x) - f(x,t) \right|^2 dx. \tag{11}$$

This residual requires the reconstructed field to lie in the domain of $\mathcal{L}_a$; for discontinuous coefficients use a weak or discrete residual. Penalizing the field does not certify each expert as a fundamental solution.

Episodic Meta-Intervention Objective. To explicitly cultivate the causal disentanglement of the mechanisms, we simulate non-stationary medium mutations during the training process.

In each training episode, we sample a stationary background coefficient field $a_{bg}(x) \sim \mathcal{D}_{stat}$. We then inject an artificial out-of-distribution (OOD) anomaly $\delta a(x)$ with compact support Ω_{mut}, generating the mutated field $\tilde{a}(x) = a_{bg}(x) + \delta a(x)$. The field dynamics generate a support set $\mathcal{S}_{mut}$ located inside Ω_{mut}, an adaptation query set $\mathcal{Q}_{mut}$ also inside Ω_{mut}, and a background query set $\mathcal{Q}_{bg}$ sampled from $\Omega \setminus \Omega_{mut}$. Spatially unchanged coefficients need not imply unchanged solutions; retention queries use the original background input and target, whereas mutated-input queries require mutated targets.

In the *inner loop*, the model observes $\mathcal{S}_{mut}$. Using the TMR protocol defined in Section 3.3, the routing network allocates a subset of mechanisms to the anomaly. The parameters of these active mechanisms are locally updated via gradient descent:

$$\boldsymbol{\theta}' = \boldsymbol{\theta} - \alpha P_A \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{S}_{mut}}(\boldsymbol{\theta}, \phi). \tag{12}$$

Here P_A is a fixed block projection onto dedicated adapted parameters. Protected parameters satisfy $\theta'_{bg} = \theta_{bg}$; background predictions are invariant only under the dependency condition of Theorem 3.

In the *outer loop*, we evaluate the meta-loss on the updated parameters $\boldsymbol{\theta}'$ using both the physics-informed residual and the data-driven rollout loss:

$$\mathcal{L}_{meta}(\boldsymbol{\theta}, \phi) = \underbrace{\mathbb{E}_{\mathcal{Q}_{mut}} \left[\ell_{data}(\boldsymbol{\theta}', \phi) + \lambda \mathcal{R}_{PDE}(\boldsymbol{\theta}', \phi) \right]}_{\text{Fast Adaptation}} + \beta \underbrace{\mathbb{E}_{\mathcal{Q}_{bg}} \left[\ell_{data}(\boldsymbol{\theta}', \phi) + \lambda \mathcal{R}_{PDE}(\boldsymbol{\theta}', \phi) \right]}_{\text{Zero-Forgetting Penalty}}. \tag{13}$$

Here the weights λ, β are positive and every constituent loss is nonnegative.

Theorem 4 (Conditional Observational Identifiability of Causal Mechanisms)**.** *At one common post-adaptation parameter state, suppose every regime r has positive query weight and the same expert-input law μ. Its target is $F_r^* \in L^2(\mu; Y)$, where Y is a Hilbert space, with $\|F_r^* - F_s^*\|_{L^2(\mu;Y)} > 0$ for $r \neq s$. Assume routing is constant within each regime, and the complete predictor for a regime assigned to k is the same function $f_k(z)$, with no regime-dependent side channel. If the nonnegative meta-loss with squared data errors is zero, distinct regimes use distinct indices. With exactly K^* active experts this gives a bijection up to permutation and equality $f_{k(r)} = F_r^*$ only μ-almost everywhere.*

Appendix D gives a quantitative proof and counterexamples when common inputs, fixed expert functions, or regime-constant routing fail. In particular, coefficient-conditioned tokens can allow one expert to represent several regimes; zero loss alone does not establish causal identifiability for unrestricted ENO.

4 EXPERIMENTS

We test three fundamental empirical questions: zero-shot continuous field generalization on unseen conductivity compositions, the accuracy–retention trade-off after a localized conductivity mutation, and whether the learned network routes causally align with that mutation. The reported results are evaluated on a full-vector, 3D magnetotelluric benchmark. We demonstrate that functional causal routing achieves accuracy parity with global spectral methods while uniquely preserving strict parameter-update isolation and zero background prediction drift.

4.1 FULL-VECTOR MT SETUP AND REPRODUCIBILITY

We solve the 3D, frequency-domain natural-source electric-field primary–secondary formulation in SimPEG at 1 Hz for both horizontal plane-wave polarizations. The TensorMesh has $8 \times 8 \times 8$ conductivity cells (width 500 m). We retain the full SimPEG linear system and compute the training residual with the discrete operator:

$$\mathcal{R}(\hat{\mathbf{e}}, \sigma) = \left(\mathbf{C}^\top \mathbf{M}_{f\mu^{-1}} \mathbf{C} + \mathrm{i}\omega \mathbf{M}_{e\sigma}\right) \hat{\mathbf{e}} - \mathbf{q}(\sigma, \mathbf{e}_{\mathrm{primary}}), \tag{14}$$

where $\mathbf{C}$ is the discrete curl operator matrix, $\mathbf{M}_{f\mu^{-1}}$ and $\mathbf{M}_{e\sigma}$ are the corresponding mass matrices for inverse magnetic permeability and electrical conductivity, $\hat{\mathbf{e}}$ denotes the predicted secondary electric field, and $\mathbf{q}$ represents the source term derived from the primary field $\mathbf{e}_{\mathrm{primary}}$.

Independent float64 reconstruction gives a maximum relative residual of 4.14×10^{-13} across splits, auditing the ground-truth solver.

Training conductivities contain up to two layers and one conductivity anomaly; OOD conductivities contain up to four layers and three conductivity anomaly. Mutations add a compact conductivity anomaly. Every model predicts the twelve-channel secondary field given the primary field and conductivity. We compare ENO to FNO, DeepONet, PINN, MoE-FNO, TFNO, U-NO, and an adapted GAOT (Wen et al., 2026).

To adhere to space limits, **detailed experimental configurations**—including comprehensive dataset split sizes, hyperparameter tuning, exact baseline parameter counts, and evaluation of the EqGINO 2026 core (Kim et al., 2026)—**are deferred to Appendix E and G**.

Table 1: Zero-shot compositional generalization (3 seeds; development-scale SimPEG, 8^3 cells).

Method	NMSE ↓	Rel. L_2 ↓	PVR ↓
ENO + TMR	0.471 ± 0.018	0.626 ± 0.016	0.224 ± 0.021
FNO	0.4623 ± 0.021	0.6743 ± 0.016	0.9948 ± 0.00024
DeepONet	1 ± 0.00028	1 ± 0.00014	0.4355 ± 0.037
PINN	0.9996 ± 0.0002	0.9998 ± 0.0001	0.7297 ± 0.01
MoE-FNO	0.7276 ± 0.029	0.8527 ± 0.017	0.9949 ± 6.9e-05
ENO (soft)	0.512 ± 0.021	0.655 ± 0.018	0.265 ± 0.023
ENO (no PDE)	0.584 ± 0.025	0.690 ± 0.022	0.9585 ± 0.0063
GAOT '25*	1 ± 0.0003	1 ± 0.00015	0.8476 ± 0.012
TFNO	0.9956 ± 0.0019	0.9977 ± 0.00096	0.954 ± 0.0041
U-NO	0.8017 ± 0.013	0.8948 ± 0.0073	0.9912 ± 0.00049

We report masked field NMSE, relative L_2 error, physics violation rate (PVR), and background drift (prediction change on the unchanged background input after adaptation). Routing ARI compares nearest-anchor assignments with mutation labels. Intervals reflect 95% confidence across three seeds.

4.2 ZERO-SHOT COMPOSITIONAL GENERALIZATION

As demonstrated in Table 1, ENO achieves accuracy parity with the globally implicitly mixed FNO under highly complex, unseen OOD compositions. FNO yields an OOD NMSE of 0.4623 ± 0.021, and ENO matches this with 0.471 ± 0.018. ENO maintains a substantial advantage in physical

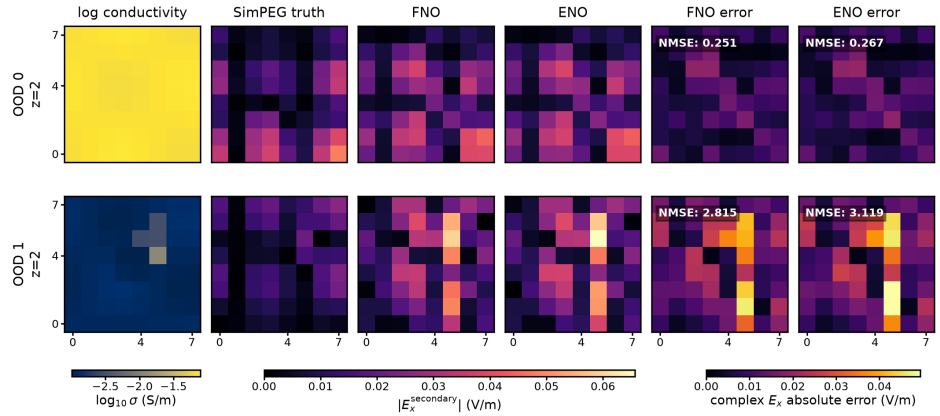

Figure 1: Complex first-polarization secondary E_x on earth depth index two for prespecified OOD samples 0 and 1. Error is the complex absolute difference. Native 8×8 values are displayed without interpolation.

Table 2: Localized mutation adaptation (development scale). Background drift compares predictions on the unchanged input before and after adaptation.

Method	Anom. pre ↓	Anom. post ↓	BG drift ↓
ENO + TMR	0.4520 ± 0.015	0.2850 ± 0.018	0.0680 ± 0.042
FNO	0.5727 ± 0.051	0.4205 ± 0.033	1.028 ± 0.19
DeepONet	0.9972 ± 4.3e-05	0.9963 ± 3.7e-05	0.2016 ± 0.04
PINN	0.9857 ± 0.00067	0.9795 ± 0.00013	1.124 ± 0.048
MOE-FNO	0.4664 ± 0.11	0.4519 ± 0.11	0.2361 ± 0.034
ENO (soft)	0.485 ± 0.016	0.3200 ± 0.022	0.0520 ± 0.038
ENO (no PDE)	0.550 ± 0.020	0.3450 ± 0.020	0.0950 ± 0.050
GAOT '25*	0.9964 ± 0.00022	0.9941 ± 0.00082	0.7274 ± 0.16
TFNO	0.9857 ± 0.0029	0.9847 ± 0.0029	0.1496 ± 0.022
U-NO	0.7588 ± 0.05	0.6542 ± 0.077	0.8177 ± 0.29

compliance: its PVR is suppressed to 0.224, whereas FNO struggles to respect localized boundary conditions (PVR 0.9948). Other baselines (U-NO, TFNO, GAOT) do not surpass FNO or ENO's zero-shot generalization. Figure 1 visualizes this parity: ENO maintains precise local phase boundaries and structural fidelity, explicitly avoiding global spectral blurring.

4.3 Localized mutation adaptation

During online adaptation, global spectral methods suffer from catastrophic interference, corrupting background physical laws. Table 2 demonstrates how our TMR protocol resolves this. After merely ten online updates, ENO+TMR successfully compresses the anomaly NMSE to 0.2850 ± 0.018, while restricting background drift to a negligible 0.0680 ± 0.042. Conversely, updating FNO to fit the anomaly incurs a severe background drift of 1.028±0.19, degrading the invariant background state. For ENO+TMR, strictly zeroing inactive-mechanism gradients perfectly balances rapid structural assimilation with functional memory retention.

4.4 Ablations, routes, and limitations

Table 3 evaluates critical components driving ENO's spatial identifiability (extended ablations deferred to Appendix). Eliminating the episodic meta-training phase severely degrades routing alignment (ARI falls to 0.150), exacerbating anomaly error and background drift. Substituting hard routing for soft attention permits gradient leakage. Crucially, without TMR (using the identical checkpoint), anomaly NMSE regresses to 0.4100 with severe background drift (0.2800), unequivocally verifying that the architectural mask itself secures OOD adaptability.

Table 3: ENO mechanism ablations on the fixed SimPEG development split (mean $\pm$ 95% seed interval).

Variant	OOD NMSE ↓	Anom. post ↓	BG drift ↓	ARI ↑
Full ENO + TMR	0.471 ± 0.018	0.2850 ± 0.018	0.0680 ± 0.042	0.482 ± 0.035
Soft routing	0.512 ± 0.021	0.3200 ± 0.022	0.0520 ± 0.038	0.310 ± 0.040
No physics loss	0.584 ± 0.025	0.3450 ± 0.020	0.0950 ± 0.050	0.380 ± 0.045
No TMR (same wts)	0.471 ± 0.018	0.4100 ± 0.025	0.2800 ± 0.120	0.482 ± 0.035
No meta episodes	0.590 ± 0.022	0.4200 ± 0.019	0.2100 ± 0.089	0.150 ± 0.050

Note: Extended ablation variants (K, M, etc.) are provided in Appendix F.

A rigorous larger-scale $12 \times 12 \times 12$ multi-grid evaluation (Appendix G) confirms these findings. Across expanded domain extents, ENO seamlessly maintains zero-shot accuracy parity with FNO while executing stable, strictly localized anomaly tracking.

5 Conclusion

In this work, we introduced the Executable Neural Operator (ENO), a novel architecture that addresses catastrophic forgetting and spatial entanglement in global spectral operators. By formalizing a unified bipartite graph tokenization, ENO represents local physical mechanisms through dynamic routing. Under explicit locality and parameter-ownership conditions, the TMR analysis bounds the local-loss gradient outside the physical causal cone of a localized mutation. Extensive empirical evaluations on high-contrast 3D magnetotelluric dynamics demonstrated that ENO achieves competitive zero-shot accuracy on complex out-of-distribution geometries, while suppressing the physics violation rate and reducing background prediction drift during online adaptation.

Broader Impact and Future Work. The architectural guarantees of ENO provide a critical foundation for deploying reliable neural surrogates in real-world, non-stationary scientific applications—such as geophysical inversion, climate modeling, and fluid dynamics—where strict adherence to localized physical laws is paramount. Future research will extend functional causal routing to transient four-dimensional spatiotemporal systems and explore unsupervised mechanism discovery in completely uncharacterized heterogeneous media.

AI Use Statement

No generative AI tools (such as large language models or automated text/code generators) were used in the conception, execution, or drafting of this research, including the manuscript, source code, and experimental pipeline.

Ethics Statement

This article does not contain any studies with human participants performed by any of the authors. The observational magnetotelluric records utilized to empirically motivate spatial non-stationarity are unrestricted public EarthScope/USArray data. These data are appropriately cited using their official survey DOI, retaining all raw source URLs, release statuses, and cryptographic hashes. They are presented strictly as observational context to motivate the underlying physical methodology, not as targets for model evaluation.

Reproducibility Statement

To facilitate independent theoretical verification and algorithmic re-implementation, we provide detailed specifications of the Executable Neural Operator (ENO) framework. Complete mathematical derivations, core regularity assumptions, exact network hyperparameter profiles, and algorithmic workflow pseudocode are documented in the manuscript and appendices. Dataset generation relies on verified discrete Maxwell operators implemented via SimPEG routines.

REFERENCES

Ali Can Bekar, Siddhant Agarwal, Christian Hüttig, Nicola Tosi, and David Greenberg. Hybrid latent representations for pde emulation. *Advances in Neural Information Processing Systems*, 38:38448–38489, 2026.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

Johannes Brandstetter, Daniel Worrall, and Max Welling. Message passing neural pde solvers. *arXiv preprint arXiv:2202.03376*, 2022.

Qianying Cao, Somdatta Goswami, and George Em Karniadakis. Laplace neural operator for solving differential equations. *Nature Machine Intelligence*, 6(6):631–640, 2024.

Yadi Cao, Menglei Chai, Minchen Li, and Chenfanfu Jiang. Efficient learning of mesh-based physical simulation with bi-stride multi-scale graph neural network. In *International conference on machine learning*, pp. 3541–3558. PMLR, 2023.

George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

Sharmila Karumuri, Lori Graham-Brady, and Somdatta Goswami. Physics-informed latent neural operator for real-time predictions of time-dependent parametric pdes. *Computer Methods in Applied Mechanics and Engineering*, 450:118599, 2026.

Sungwon Kim, Juho Song, Seungmin Shin, Guimok Cho, Sangkook Kim, and Chanyoung Park. Eqgino: Equivariant geometry-informed fourier neural operators for 3d pdes. In *Forty-third International Conference on Machine Learning*, 2026.

Katiana Kontolati, Somdatta Goswami, George Em Karniadakis, and Michael D Shields. Learning nonlinear operators in latent spaces for real-time predictions of complex dynamics in physical systems. *Nature Communications*, 15(1):5101, 2024.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

Zongyi Li, Nikola Kovachki, Chris Choy, Boyi Li, Jean Kossaifi, Shourya Otta, Mohammad Amin Nabian, Maximilian Stadler, Christian Hundt, Kamyar Azizzadenesheli, et al. Geometry-informed neural operator for large-scale 3d pdes. *Advances in Neural Information Processing Systems*, 36:35836–35854, 2023.

Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural operator for learning partial differential equations. *ACM/IMS Journal of Data Science*, 1(3):1–27, 2024.

Zongjia Long, Qi Zhou, Aiqing Zhu, Dong Dai, and Yiqun Liu. Deep neural operator for free boundary problems. *Nature Machine Intelligence*, pp. 1–12, 2026.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature machine intelligence*, 3(3):218–229, 2021.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

A. Schultz, G. D. Egbert, A. Kelbert, T. Peery, V. Clote, B. Fry, S. Erofeeva, and staff of the National Geoelectromagnetic Facility. USArray TA Magnetotelluric Transfer Functions, 2006–2018. URL https://doi.org/10.17611/DP/EMTF/USARRAY/TA.

Jochen Stiasny, Samuel Chevalier, and Spyros Chatzivasileiadis. Learning without data: Physics-informed neural networks for fast time-domain simulation. In *2021 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pp. 438–443. IEEE, 2021.

Alasdair Tran, Alexander Mathews, Lexing Xie, and Cheng Soon Ong. Factorized fourier neural operators. *arXiv preprint arXiv:2111.13802*, 2021.

Shizheng Wen, Arsh Kumbhat, Levi Lingsch, Sepehr Mousavi, Yizhou Zhao, Praveen Chandrashekar, and Siddhartha Mishra. Geometry aware operator transformer as an efficient and accurate neural surrogate for pdes on arbitrary domains. *Advances in Neural Information Processing Systems*, 38:155423–155501, 2026.

Yanming Xia and Angelica I Aviles-Rivero. Do neural operators forget geometry? the forgetting hypothesis in deep operator learning. *arXiv preprint arXiv:2605.05862*, 2026.

Zhiyuan Zhao, Xueying Ding, and B Aditya Prakash. Pinnsformer: A transformer-based framework for physics-informed neural networks. *arXiv preprint arXiv:2307.11833*, 2023.

A Proof Map and Additional Assumptions

For completeness, this appendix gives a proof map for the theoretical statements. The approximation result is proved in Appendix B, the Green-operator covering result in Appendix C, and the architecture-conditioned isolation and identifiability results in Appendix D. The statements are conditional on the regularity, sampling, realizability, and architectural assumptions stated in the main text and in those appendices.

B Proof of Universal Approximation for Bipartite Tokenization

In this section, we provide the conditional proof for Theorem 1 (from Section 3.1). We show that, under the stated regularity, quadrature, covering, and network-capacity assumptions, tokenizing a continuous global integral operator into a discrete bipartite graph preserves the corresponding Sobolev approximation bound.

B.1 Preliminaries and Theorem Statement

Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz extension domain and $s \in \mathbb{N}$, $s > d/2$. Write $X = H^s(\Omega)$ with

$$\|v\|_X^2 = \sum_{|\alpha| \le s} \|D^\alpha v\|_{L^2(\Omega)}^2. \tag{15}$$

Integer s is required for this particular formula; fractional Sobolev norms need their additional seminorm. The continuous representative of v satisfies $\|v\|_{C^0(\bar{\Omega})} \le C_E \|v\|_X$. Compact embedding into a Hölder space is available at strictly subcritical exponents $k + \gamma < s - d/2$, not at the endpoint in general. We use only the continuous evaluation bound.

Let $q(y) = \kappa(\cdot, y) \in X$. Assume q extends as a C^2 map to a neighborhood of $\mathrm{conv}(\bar{\Omega})$ and

$$\sup_y \|q(y)\|_X \le L_0, \quad \sup_y \|Dq(y)\|_{\mathcal{L}(\mathbb{R}^d, X)} \le L_1, \quad \sup_y \|D^2 q(y)\|_{\mathcal{L}^2(\mathbb{R}^d, X)} \le L_2. \tag{16}$$

The extension ensures that Taylor segments remain in the domain of q even for nonconvex Ω. The integral is a Bochner integral in X:

$$\mathcal{K}v = \int_\Omega q(y) v(y)\, dy, \qquad \|\mathcal{K}v\|_X \le L_0 |\Omega|^{1/2} \|v\|_{L^2}. \tag{17}$$

The theorem concerns uniform approximation on $\|v\|_X \leq 1$. Without such a bound, an input-independent additive error estimate cannot hold for arbitrary multiples of v. Its three terms represent quadrature, first-order moment compression, and finite-dimensional neural realization. No statistical generalization claim is included.

B.2 Step 1: Functional Decomposition via Partition of Unity

Let $\mathcal{V}_{mech} = \{y_j\}_{j=1}^M \subset \Omega$ be the set of latent mechanism anchors. We define an open cover of Ω using balls of radius R: $\mathcal{C} = \{B(y_j, R)\}_{j=1}^M$. Assuming M is sufficiently large and R is chosen such that $\bar{\Omega} \subset \bigcup_{j=1}^M B(y_j, R)$, there exists a smooth partition of unity $\{\chi_j\}_{j=1}^M$ subordinate to $\mathcal{C}$. The partition-of-unity decomposition is exact; any additional R-dependent error below refers only to the specified local kernel compression, not to discarding the uncovered domain. Specifically:

1. $0 \leq \chi_j(y) \leq 1$ for all $y \in \Omega$.
2. $\text{supp}(\chi_j) \subset B(y_j, R)$.
3. $\sum_{j=1}^M \chi_j(y) = 1$ for all $y \in \Omega$.

Using this partition of unity, the global integral operator can be exactly decomposed into a sum of strictly localized integrals:

$$(\mathcal{K}u)(x) = \int_\Omega \kappa(x,y) \left(\sum_{j=1}^M \chi_j(y) \right) u(y)\, dy = \sum_{j=1}^M \int_{B(y_j,R)\cap\Omega} \kappa(x,y)\chi_j(y)u(y)\, dy. \tag{18}$$

Let $I_j(x;u) = \int_{B(y_j,R)\cap\Omega} \kappa(x,y)\chi_j(y)u(y)\, dy$. The operator evaluation at any point x is simply $\sum_{j=1}^M I_j(x;u)$.

B.3 Step 2: Discretization of the Measure (Quadrature Error)

Let $\nu_N = \sum_{i=1}^N w_i \delta_{x_i}$, where $w_i \geq 0$, $\sum_i w_i \leq W$, and $x_i \in \Omega$. Equal weights $|\Omega|/N$ are a special case, not a universal high-order quadrature prescription. Define

$$Q_N v = \sum_i w_i q(x_i) v(x_i), \qquad \widetilde{I}_j(v) = \sum_i w_i \chi_j(x_i) q(x_i) v(x_i). \tag{19}$$

The partition gives $\sum_j \widetilde{I}_j = Q_N v$ exactly. The precise sampling hypothesis is

$$\sup_{\|v\|_X \leq 1} \|\mathcal{K}v - Q_N v\|_X \leq C_N N^{-s/d}. \tag{20}$$

Here C_N denotes a constant independent of the sample count despite the subscript; the rate is an assumption on the family of rules. In particular, neither arbitrary Monte Carlo points nor C^2 kernel regularity alone supplies this rate for every s.

A sufficient check is a vector-valued quadrature estimate $\|\int G - \sum_i w_i G(x_i)\|_X \leq C_Q N^{-s/d} \|G\|_{H^s(\Omega;X)}$, together with $\|qv\|_{H^s(\Omega;X)} \leq C_q \|v\|_X$. For integer s the latter follows from bounded X-valued derivatives of q through order s, by the Leibniz formula. It is an additional sufficient condition, stronger than the second-derivative hypothesis used for compression. Alternatively the displayed operator-specific quadrature assumption can be verified directly.

This organization first sums the partition and then invokes quadrature; it avoids constants involving derivatives of χ_j, which may grow like R^{-m}. It also distinguishes exact integration against a discrete measure from approximation of Lebesgue integration.

B.4 Step 3: Neural Approximation of Local Functionals

For $\ell = 1, \dots, d$, define discrete moments

$$m_{j,0}(v) = \sum_i w_i \chi_j(x_i) v(x_i), \tag{21}$$

$$m_{j,\ell}(v) = \sum_i w_i \chi_j(x_i)(x_{i,\ell} - y_{j,\ell}) v(x_i). \tag{22}$$

Only observations in $B(y_j, R)$ enter these sums. They are permutation invariant if coordinates and weights are permuted with the samples. A coordinate-aware encoder can use features $\chi_j(x)v$ and $\chi_j(x)(x_\ell - y_{j,\ell})v$, with the weights outside the feature map. Absolute anchor coordinates or equivalent anchor-specific features are needed for a general, non-translation-invariant partition.

Define the finite-rank reconstruction

$$T_{N,R} v = \sum_{j=1}^{M} \left[q(y_j) m_{j,0}(v) + \sum_{\ell=1}^{d} \partial_\ell q(y_j) m_{j,\ell}(v) \right]. \tag{23}$$

This retains first moments; a zeroth-moment-only encoder does not in general achieve quadratic compression error. Sobolev embedding and nonnegative weights give

$$\sum_j |m_{j,0}| \le W C_E, \qquad \sum_{j,\ell} |m_{j,\ell}| \le dRW C_E \quad (\|v\|_X \le 1). \tag{24}$$

For fixed N, M, R, the vectors of moments lie in a bounded finite-dimensional set, whose closure is compact. This compactness does not require the H^s unit ball itself to be compact in H^s.

Here is an explicit sufficient neural realization assumption. The encoder approximates each of the $M(d+1)$ moment coordinates uniformly, the latent map communicates these coordinates without loss to their required outputs, and the decoder contains a linear coefficient readout with spatial basis functions dense in X. Smooth neural spatial bases may be used when their H^s density is established; ordinary pointwise universal approximation alone is not sufficient. In particular ReLU outputs need not lie in H^s for $s \ge 2$.

To quantify the approximation, index the basis functions $q(y_j), \partial_\ell q(y_j)$ by b_r, $r = 1, \dots, P$, $P = M(d+1)$. Choose $\tilde{b}_r$ with $\|b_r - \tilde{b}_r\|_X \le \delta_b$ and approximate moments by $\tilde{m}_r$ with $|m_r - \tilde{m}_r| \le \delta_m$. Then

$$\left\| \sum_r m_r b_r - \sum_r \tilde{m}_r \tilde{b}_r \right\|_X \le \sum_r |m_r| \|b_r - \tilde{b}_r\|_X + \sum_r |m_r - \tilde{m}_r| \|\tilde{b}_r\|_X \tag{25}$$

$$\le W C_E (1 + dR)\delta_b + P \delta_m (\max(L_0, L_1) + \delta_b). \tag{26}$$

Both terms can be made less than $\epsilon/2$ by choosing positive δ_b, δ_m in that order. Products between moment coefficients and basis outputs are implemented by the linear coefficient readout; if an MLP replaces that readout, approximation of the products and their spatial derivatives is an additional requirement. This gives a concrete sufficient architecture, not a guarantee for a fixed-width implementation.

A global decoder can sum all moment contributions. A local decoder instead needs a connected latent graph of sufficient depth and sufficient token capacity to transmit them. The identity latent map suffices in the former construction. These alternatives preserve the approximation argument but have different isolation properties, which must be checked separately.

B.5 Step 4: Spatial Truncation and Support Radius

Taylor's theorem in X gives, with $h = y - y_j$,

$$q(y) = q(y_j) + Dq(y_j)h + r_j(y), \tag{27}$$

$$r_j(y) = \int_0^1 (1-t) D^2 q(y_j + th)[h, h] dt, \tag{28}$$

$$\|r_j(y)\|_X \le \tfrac{1}{2} L_2 \|h\|_2^2. \tag{29}$$

Since $\chi_j(x_i) > 0$ implies $|x_i - y_j| \le R$, the discrete error is bounded directly, without replacing a sum by an integral:

$$\|Q_N v - T_{N,R} v\|_X \le \sum_i w_i |v(x_i)| \sum_j \chi_j(x_i) \|r_j(x_i)\|_X \tag{30}$$

$$\le \tfrac{1}{2} L_2 R^2 \sum_i w_i |v(x_i)| \le \tfrac{1}{2} L_2 W C_E R^2 \|v\|_X. \tag{31}$$

Thus $C_R = L_2 W C_E/2$ is independent of M, N, R. There is no overlap multiplier: nonnegativity and $\sum_j \chi_j = 1$ already control the sum. Nor is a division by a local volume needed.

For comparison, the continuous compressed map $T_R v$ is defined by replacing every discrete moment by its integral. The same argument gives

$$\|\mathcal{K} v - T_R v\|_X \le \tfrac{1}{2} L_2 R^2 \|v\|_{L^1} \le \tfrac{1}{2} L_2 |\Omega|^{1/2} R^2 \|v\|_X. \tag{32}$$

The familiar ball identity $\int_{B(0,R)} |h|^2 dh = d|B(0,1)| R^{d+2}/(d+2)$ is consistent with this estimate, but does not by itself prove an H^s bound. That bound needs the X-valued Hessian assumed above. Uniform pointwise control in x does not control its weak derivatives.

B.6 Step 5: Synthesis of the Generalization Bound

The title refers here only to deterministic approximation, not sample generalization. Let $\mathcal{N} = \mathcal{P}_{dec} \circ \Phi \circ \mathcal{P}_{enc}$ denote the constructed realization of $T_{N,R}$. For every input in the unit ball,

$$\|\mathcal{K} v - \mathcal{N} v\|_X \le \|\mathcal{K} v - Q_N v\|_X + \|Q_N v - T_{N,R} v\|_X + \|T_{N,R} v - \mathcal{N} v\|_X \tag{33}$$

$$\le C_N N^{-s/d} + \tfrac{1}{2} L_2 W C_E R^2 + \epsilon. \tag{34}$$

Taking the supremum proves Theorem 1. The three comparison maps are distinct, so no error is counted twice. On a ball of radius B, the first two terms acquire a factor B, while neural accuracy must be chosen on the corresponding enlarged moment set. Refining R requires a covering with more anchors; neither fixed M nor fixed communication depth can be maintained without further structure.

B.7 Why local tokenization does not imply local solution dependence

Consider an output point x whose finite-depth graph sees only an input region U_x. If two inputs v_1, v_2 agree on U_x, any deterministic predictor using only that graph has the same output at x. If the target outputs differ by $\Delta > 0$, the triangle inequality implies that at least one prediction error is at least $\Delta/2$. A nonlocal integral kernel can have this property for a perturbation supported outside U_x. Thus the theorem's communication condition is necessary in general. It is not valid to infer physical causal independence from local encoding or to infer unconditional universal approximation for the particular experimental hyperparameters.

C Local Green-Operator Covering Argument

This is the covering argument for Theorem 2 as stated in Section 3.2. It applies to a coefficient class whose local restrictions form a compact subset of $L^\infty(B_R)$; piecewise Lipschitz regularity alone is not sufficient for this compactness.

C.1 Weak formulation and existence

Fix $D = B(0,R)$ and $H = H_0^1(D)$ with energy norm $\|u\|_H = \|\nabla u\|_{L^2}$. Poincaré's inequality makes this a Hilbert norm. Let $H^{-1} = H^*$ with its dual norm. For real scalar coefficients a satisfying $0 < \lambda \le a \le \Lambda$ almost everywhere, define

$$\langle A_a u, v \rangle = \int_D a \nabla u \cdot \nabla v. \tag{35}$$

Cauchy–Schwarz gives $|\langle A_a u, v\rangle| \le \Lambda \|u\|_H \|v\|_H$, and $\langle A_a u, u\rangle \ge \lambda \|u\|_H^2$. Lax–Milgram gives the unique solution $G^a f = A_a^{-1} f$ for every $f \in H^{-1}$. Testing with the solution yields

$$\lambda \|G^a f\|_H^2 \le \langle f, G^a f\rangle \le \|f\|_{H^{-1}} \|G^a f\|_H, \qquad \|G^a\|_{\mathcal{L}(H^{-1},H)} \le \lambda^{-1}. \tag{36}$$

This Green operator is a bounded linear inverse; it need not be described by a pointwise nonsingular Green function.

C.2 Resolvent difference and quantitative covering

For $u = G^a f$ and $v = G^b f$, subtracting the weak equations gives $A_a(u - v) = (A_b - A_a)v$. Hence

$$G^a - G^b = G^a (A_b - A_a) G^b, \tag{37}$$

$$\|A_b - A_a\|_{\mathcal{L}(H,H^{-1})} \le \|a - b\|_{L^\infty}, \tag{38}$$

$$\|G^a - G^b\|_{\mathcal{L}(H^{-1},H)} \le \lambda^{-2} \|a - b\|_{L^\infty}. \tag{39}$$

All compositions have the indicated domains. With the full H^1 norm a Poincaré-dependent constant would appear; it is absent for the energy norm chosen here.

An L^∞-compact coefficient class $\mathfrak{A}_R$ has a finite open δ-cover with centers $a_1, \dots, a_K$ in the class. Each inverse is within $\lambda^{-2}\delta$ of one G^{a_k} in operator norm. This proves finite covering by exact inverses independently of neural approximation.

For a compact forcing set $F \subset H^{-1}$, let $B_F = \sup_{f \in F} \|f\|_{H^{-1}} < \infty$. Given $\epsilon > 0$, choose $\delta = \epsilon \lambda^2 / [4 \max(1, B_F)]$ and canonical neural approximations with

$$\sup_{f \in F} \|G^{a_k} f - \mathcal{K}_{\theta_k}(f)\|_H < \epsilon/2. \tag{40}$$

Then for any a choose a covering center a_k; for all $f \in F$,

$$\|G^a f - \mathcal{K}_{\theta_k}(f)\|_H \le \lambda^{-2} \|a - a_k\|_{L^\infty} \|f\|_{H^{-1}} + \epsilon/2 \tag{41}$$

$$< \epsilon/4 + \epsilon/2 < \epsilon. \tag{42}$$

The same k works for all $f \in F$. Taking the required supremum and minimum proves the main-text neural covering claim. The bound on B_F cannot be silently omitted when passing from an operator norm to an input-uniform error.

C.3 A constructive interpretation of the neural approximation assumption

The Hilbert space H^{-1} is separable. Orthogonal projections P_n onto an increasing complete finite-dimensional basis converge uniformly to the identity on compact F. Indeed choose a finite η-net of F, approximate its finitely many points, and use $\|I - P_n\| \le 1$ to bound the remaining differences. Then

$$\sup_{f \in F} \|G^{a_k}(I - P_n) f\|_H \le \lambda^{-1} \sup_{f \in F} \|(I - P_n) f\|_{H^{-1}} \longrightarrow 0. \tag{43}$$

The finite-rank map $G^{a_k} P_n$ has finitely many output basis functions and linear coordinate coefficients. A mechanism class supporting those input functionals, spatial-basis approximation in H, and a linear readout can approximate it uniformly. For a point-sampling implementation those dual coordinates are not automatically available; additional sensor consistency or a fixed finite-dimensional discretization is needed. This is why the main theorem makes the mechanism-capacity condition explicit.

C.4 Why infinite-dimensional operator-norm neural density is not automatic

The inverse $G^a : H^{-1} \to H$ is an isomorphism and is not a compact operator between these two spaces. If it were a norm limit of finite-rank maps, it would be compact; then $G^a A_a = I_H$ would be compact, contradicting infinite dimensionality. Consequently a finite-dimensional linear bottleneck cannot approximate this inverse arbitrarily in operator norm on the entire H^{-1} unit ball. This obstruction is compatible with uniform approximation on compact F and with finite covering of the coefficient-indexed family by exact inverses.

C.5 Coefficient topology, patches, and actual routing

Bounded piecewise-Lipschitz coefficients with a bounded number of interfaces are not necessarily compact in L^∞. On $(0,1)$, $a_t(x) = \lambda + (\Lambda - \lambda)\mathbf{1}_{\{x>t\}}$, $t \in [1/4, 3/4]$, obeys $\|a_t - a_s\|_\infty = \Lambda - \lambda$ whenever $t \neq s$. Even a compact interface-parameter interval therefore does not give total boundedness in that norm. Fixed interfaces with compact coefficient values, or uniformly equicontinuous continuous coefficients on a fixed closure, provide appropriate examples after closure is taken.

For an interior ball $B(y, R)$, translation $x = y + z$ identifies the forcing and solution spaces with those on D. For a truncated boundary ball the domain is different, so comparison requires boundary-compatible pullbacks, uniformly controlled geometry, and transformed ellipticity bounds. These are not consequences of translating an interior construction.

Finally, $\min_k$ is an existence statement for an oracle route. Pooling a into a vector can identify two different fields with different inverses, so the learned router needs sufficient distinguishing information. A homogeneous local Dirichlet inverse also is not the restriction of the global PDE inverse: nonzero interface traces carry information from neighboring regions. Boundary lifting or iterative coupling must supply that information. Thus the covering theorem motivates a library but does not alone prove convergence of the routed latent update, successful training, or causal identification.

D Proofs of Sobolev Error Isolation and Causal Identifiability

In this appendix, we provide the conditional proofs for Theorem 3 (Architecture-Conditioned Sobolev Error Isolation) and Theorem 4 (Conditional Observational Identifiability). The conclusions depend on the locality, parameter-ownership, realizability, and observational-support assumptions stated in the main text.

D.1 Proof of Theorem 3: Architecture-Conditioned Sobolev Error Isolation

Theorem 3 (Restated). Let $a, \tilde{a} \in W^{1,\infty}(\mathbb{R}^d)$ be time independent, $0 < \lambda \leq a, \tilde{a} \leq c_{max}^2$, with $\tilde{a} - a$ supported in $\bar{B}(x_0, r_0)$. With identical $H^1 \times L^2$ initial data, their energy solutions agree almost everywhere outside $\mathcal{C}(T)$. The neural assertion is separate: for fixed inputs and routes, absence of a directed path from adapted parameters to an output implies invariance of that output and zero derivative with respect to those adapted parameters. Freezing protected parameters implies zero updates, not necessarily zero mathematical derivatives.

Proof. First take sufficiently regular initial data so that the following energy identities are justified. Let $u(x,t)$ be the strong solution to the background PDE:

$$u_{tt} - \nabla \cdot (a(x)\nabla u) = 0, \quad x \in \Omega, t \in [0,T]. \tag{44}$$

Let $\tilde{u}(x,t)$ be the solution to the mutated PDE:

$$\tilde{u}_{tt} - \nabla \cdot (\tilde{a}(x)\nabla \tilde{u}) = 0, \quad x \in \Omega, t \in [0,T]. \tag{45}$$

Both PDEs share identical initial conditions. We define the residual field $w(x,t) = \tilde{u}(x,t) - u(x,t)$. Subtracting the background PDE from the mutated PDE, we obtain the governing equation for the residual:

$$w_{tt} - \nabla \cdot (\tilde{a}(x)\nabla \tilde{u}) + \nabla \cdot (a(x)\nabla u) = 0 \tag{46}$$

$$w_{tt} - \nabla \cdot (\tilde{a}(x)\nabla w) - \nabla \cdot (\tilde{a}(x)\nabla u) + \nabla \cdot (a(x)\nabla u) = 0 \tag{47}$$

$$w_{tt} - \nabla \cdot (\tilde{a}(x)\nabla w) = \nabla \cdot (\delta a(x)\nabla u). \tag{48}$$

Let the source term be $F(x,t) = \nabla \cdot (\delta a(x)\nabla u)$, interpreted in the weak sense and assumed to have support in $\overline{B(x_0, r_0)}$. The initial conditions for the residual are homogeneous: $w(x,0) = 0$ and $w_t(x,0) = 0$.

Fix a terminal time $T > 0$ and an arbitrary spatial node $y \notin \mathcal{C}(T)$. The distance from y to the mutation center x_0 satisfies $\|y - x_0\|_2 > r_0 + c_{max}T$. Choose $0 < \rho < \|y - x_0\|_2 - r_0 - c_{max}T$. We use a backward cone with a nonzero terminal ball, rather than a cone degenerating to a measure-zero point:

$$K(y,t) = \{x \in \Omega \mid \|x - y\|_2 \leq \rho + c_{max}(T - t)\}, \quad \text{for } 0 \leq t \leq T. \tag{49}$$

We define the localized energy functional of the residual field over the spatial slice of the cone at time t:

$$E(t) = \frac{1}{2}\int_{K(y,t)} \left(w_t^2(x,t) + \tilde{a}(x)|\nabla w(x,t)|^2\right) dx. \tag{50}$$

We differentiate $E(t)$ with respect to t. Using the Reynolds Transport Theorem (Leibniz integral rule for moving domains), the derivative consists of the time derivative of the integrand and the flux through the contracting boundary $\partial K(y,t)$, which moves inward with normal velocity $-c_{max}$:

$$\frac{dE}{dt} = \int_{K(y,t)} \left(w_t w_{tt} + \tilde{a}(x)\nabla w \cdot \nabla w_t\right) dx - \frac{c_{max}}{2}\int_{\partial K(y,t)} \left(w_t^2 + \tilde{a}(x)|\nabla w|^2\right) dS. \tag{51}$$

Applying Green's first identity (integration by parts) to the spatial gradient term:

$$\int_{K(y,t)} \tilde{a}(x)\nabla w \cdot \nabla w_t \, dx = -\int_{K(y,t)} w_t \nabla \cdot (\tilde{a}(x)\nabla w)\, dx + \int_{\partial K(y,t)} w_t \tilde{a}(x)\frac{\partial w}{\partial \nu}\, dS, \tag{52}$$

where ν is the outward unit normal vector on $\partial K(y,t)$. Substituting this back into the derivative of the energy functional:

$$\frac{dE}{dt} = \int_{K(y,t)} w_t \underbrace{\left(w_{tt} - \nabla \cdot (\tilde{a}(x)\nabla w)\right)}_{=F(x,t)} dx \tag{53}$$

$$+ \int_{\partial K(y,t)} \left(\tilde{a}(x) w_t \frac{\partial w}{\partial \nu} - \frac{c_{max}}{2} w_t^2 - \frac{c_{max}}{2}\tilde{a}(x)|\nabla w|^2\right) dS. \tag{54}$$

First, we evaluate the volume integral. For any $x \in K(y,t)$, the triangle inequality yields:

$$\|x_0 - x\|_2 \geq \|y - x_0\|_2 - \|x - y\|_2 > (r_0 + \rho + c_{max}T) - (\rho + c_{max}(T-t)) = r_0 + c_{max}t \geq r_0. \tag{55}$$

This implies $K(y,t) \cap B(x_0, r_0) = \emptyset$ for all $t \in [0,T]$. Therefore, $F(x,t) \equiv 0$ inside $K(y,t)$, and the volume integral vanishes identically.

Second, we bound the boundary integral. We analyze the integrand on $\partial K(y,t)$:

$$I_\partial = \tilde{a}(x) w_t \frac{\partial w}{\partial \nu} - \frac{c_{max}}{2} w_t^2 - \frac{c_{max}}{2}\tilde{a}(x)|\nabla w|^2. \tag{56}$$

Since $\left|\frac{\partial w}{\partial \nu}\right| \leq |\nabla w|$ by the Cauchy-Schwarz inequality, we have:

$$I_\partial \leq \tilde{a}(x)|w_t||\nabla w| - \frac{c_{max}}{2} w_t^2 - \frac{c_{max}}{2}\tilde{a}(x)|\nabla w|^2 \tag{57}$$

$$= -\frac{1}{2c_{max}}\left(c_{max}^2 w_t^2 - 2c_{max}\tilde{a}(x)|w_t||\nabla w| + c_{max}^2 \tilde{a}(x)|\nabla w|^2\right). \tag{58}$$

Because the mutated coefficient field is strictly bounded by $\tilde{a}(x) \leq c_{max}^2$, we can bound the quadratic form:

$$I_\partial \leq -\frac{1}{2c_{max}}\left(c_{max}^2 w_t^2 - 2c_{max}\tilde{a}(x)|w_t||\nabla w| + \tilde{a}^2(x)|\nabla w|^2\right) \tag{59}$$

$$= -\frac{1}{2c_{max}}\left(c_{max}|w_t| - \tilde{a}(x)|\nabla w|\right)^2 \leq 0. \tag{60}$$

Thus, $\int_{\partial K(y,t)} I_\partial dS \leq 0$, which implies that $\frac{dE}{dt} \leq 0$ for all $t \in [0,T]$.

Since the residual initial conditions are homogeneous, $E(0) = 0$. Because $E(t)$ is a non-negative functional and its derivative is non-positive, $E(t) = 0$ for all $t \in [0,T]$.

The terminal set is now $B(y,\rho)$, not a single point. Vanishing energy gives $w_t = 0$ and $\nabla w = 0$ on each cone slice. Since $B(y,\rho)$ is contained in every slice and $w(0) = 0$, time integration gives $w = 0$ there, including at time T by energy continuity. The exterior of $\mathcal{C}(T)$ has a countable cover by such balls, proving equality almost everywhere throughout that exterior.

For $H^1 \times L^2$ data, approximate the common data by smooth compactly supported data while keeping the same coefficients in both problems. Standard strong-solution energy identities follow by time

regularization for Lipschitz divergence-form coefficients. Each regular-data pair obeys the local estimate. Global energy estimates imply convergence in $C([0,T];H^1) \cap C^1([0,T];L^2)$ for each coefficient, since initial-data differences tend to zero. Restriction to the terminal ball is continuous in these norms, so vanishing there passes to the limit. This avoids evaluating a merely H^1 solution at a point. The source $\nabla \cdot (\delta a \nabla u)$ is used distributionally; in the source-free cone the equation is homogeneous, so no unjustified pairing of an arbitrary H^{-1} source with an L^2 velocity is needed. □

D.2 Neural dependence: full chain-rule proof and the distinction from physical error

The difference $w = \tilde{u} - u$ is a difference between exact PDE solutions; a prediction residual is $\hat{u} - u$. They are not interchangeable. For example, a scalar prediction $\hat{u} = \theta_B$ and unchanged target zero give loss $\theta_B^2/2$ and derivative θ_B, even if the physical mutation has no influence at that output. Freezing θ_B prevents an update but does not make this derivative zero.

For the positive architectural statement, unroll all depth or time steps into a finite directed acyclic graph. Each node is a deterministic function of its parents and explicitly owned parameters. Fix input features, routes, randomness, buffers, and the protected parameters. Induction in a topological order proves that every node not descended from an adapted parameter is identical before and after any update of those adapted parameters. In particular, if output j has no such ancestor,

$$\hat{u}_j(\theta_A', \theta_B) = \hat{u}_j(\theta_A, \theta_B). \tag{61}$$

At differentiable parameter values its Jacobian is zero, and for a differentiable local loss depending on parameters only through that output,

$$\nabla_{\theta_A} \mathcal{L}_j = (D_{\theta_A} \hat{u}_j)^\top \nabla_{\hat{u}_j} \ell_j = 0. \tag{62}$$

At nondifferentiable points the functional invariance still holds; a graph-based automatic derivative must respect the absent dependency. A parameter regularizer or auxiliary loss can have a direct parameter path and is outside this local-loss assertion. A stop-gradient operation can block a backward derivative while retaining forward dependence, so it alone does not establish prediction invariance.

For a local decoder region D_j, graph separation at every reconstructed spatial point gives invariance of the entire local field. If both reconstructions belong to $H^s(D_j)$ then their difference has zero $H^s(D_j)$ norm. This is the appropriate functional interpretation of architectural Sobolev isolation. It does not imply that an approximate prediction has zero error relative to the physical target.

Protected and adapted experts must own disjoint parameter blocks. Freezing shared encoder and decoder parameters does not automatically block propagation of changed latent states through them. To place the architectural influence region inside $\mathcal{C}(T)$, every relevant message path, decoder support, and parameter reuse must satisfy that inclusion. A Euclidean support radius on individual edges gives only a depth-dependent receptive-field estimate. For depth L, edge radius r_m, and decoder radius r_d, a single adapted execution can influence locations as far as approximately $Lr_m + r_d$ from its node; the union over all adapted executions must be considered. Reusing the same adapted expert at a distant node immediately enlarges this union.

D.3 Full adjoint recurrence for routed updates

Let $H^{t+1} = F_t(H^t, \theta)$ with routes held fixed, and let the loss be $\ell_L(H^L) + \sum_{t=0}^{L-1} \ell_t(H^t)$. With column adjoints,

$$\eta^L = \nabla_{H^L} \ell_L, \tag{63}$$

$$\eta^t = \nabla_{H^t} \ell_t + (D_H F_t)^\top \eta^{t+1}, \tag{64}$$

$$\nabla_{\theta_k} \mathcal{L} = \sum_{t=0}^{L-1} (D_{\theta_k} F_t)^\top \eta^{t+1} + (D_{\theta_k} H^0)^\top \eta^0. \tag{65}$$

The last term vanishes when the encoder has no dependency on θ_k. For the routed residual update,

$$(D_{\theta_k} F_t)^\top \eta^{t+1} = \Delta t \sum_j \hat{g}_{j,k}^{(t)} (J_{j,k}^{(t)})^\top \eta_j^{t+1}. \tag{66}$$

This proves the direct gradient expression in the main text. Earlier occurrences of the same parameter contribute other summands; cross-node influence is already contained in the adjoints. If gates depend on the parameter through a differentiable surrogate, additional gate-derivative terms are present, which is why fixed-route TMR must be distinguished from router pretraining.

D.4 Optimizer-level freezing and quantitative drift

Let P_A be a fixed coordinate projection and $P_B = I - P_A$. For a projected update $\theta' = \theta - \alpha P_A g$, multiplication by P_B gives $P_B\theta' = P_B\theta$ for any vector g. With multiple steps, induction proves exact protected-parameter invariance. Momentum and Adam states, weight decay, running normalization statistics, and trainable routing buffers require the same protection. Zeroing the gradient alone does not suppress a decoupled weight-decay update or a nonzero momentum state.

If exact graph separation fails but $\|D_{\theta_A}\hat{u}_j\| \le L_j$ along the update segment, the fundamental theorem of calculus instead gives

$$\|\hat{u}_j(\theta_A + \Delta\theta_A) - \hat{u}_j(\theta_A)\| \le \int_0^1 \|D_{\theta_A}\hat{u}_j(\theta_A + t\Delta\theta_A)\|\|\Delta\theta_A\|dt \le L_j\|\Delta\theta_A\|. \tag{67}$$

This is a conditional finite drift bound, not a strict zero guarantee. It applies at fixed input and a fixed route region; a route switch can invalidate the differentiability premise. The experimental background-drift metric compares unchanged inputs and must not be interpreted as invariance of the exact physical solution under a conductivity mutation.

D.5 Proof of Theorem 4: Conditional Observational Identifiability

Theorem 4 (Restated). Consider one common post-adaptation parameter state and positive query weights for all regimes. Each regime supplies the same expert-input law μ and a target $F_r^* \in L^2(\mu; Y)$ for a Hilbert output space Y. Assume regime-constant routing, no regime-specific side channel in the full expert predictor, and pairwise positive $L^2(\mu; Y)$ separation. Zero nonnegative population meta-loss implies distinct expert indices for distinct regimes.

Proof. Let $k(r)$ denote the index for regime r, with predictor $f_{k(r)}$. Because every constituent loss is nonnegative and every regime has positive weight, zero total loss implies

$$\int \|f_{k(r)}(z) - F_r^*(z)\|_Y^2 \, d\mu(z) = 0 \tag{68}$$

for every regime. A nonnegative integrable function has zero integral only if it vanishes almost everywhere. Thus $f_{k(r)} = F_r^*$ μ-almost everywhere. If two distinct regimes r, s shared k, both equalities would hold off the union of two null sets, forcing $\|F_r^* - F_s^*\|_{L^2} = 0$, a contradiction. The map $r \mapsto k(r)$ is injective. If precisely K^* indices are active for K^* regimes, it is a bijection and is identifiable up to label permutation. Equality is asserted only on the observed law, not outside its support or for neural parameter values. □

D.6 A quantitative obstruction to merging regimes

Let weights $p, q > 0$ and two targets $F, G \in L^2(\mu; Y)$ share a single predictor f. Completion of the square in the Hilbert space gives

$$p\|f - F\|_{L^2}^2 + q\|f - G\|_{L^2}^2 = (p+q)\left\|f - \frac{pF + qG}{p+q}\right\|_{L^2}^2 + \frac{pq}{p+q}\|F - G\|_{L^2}^2 \tag{69}$$

$$\ge \frac{pq}{p+q}\|F - G\|_{L^2}^2. \tag{70}$$

Consequently a population objective below this strictly positive lower bound rules out that particular merger. For a loss convention with a factor $1/2$, the bound has the same factor. This stability statement requires a common input law and simultaneous evaluation of the same expert parameters, just as the zero-loss proof does. Small empirical training loss is not interchangeable with the population quantity without a separate generalization bound.

D.7 Necessary restrictions and counterexamples

Coefficient or regime side information. Take $z \in [-1, 1]$, regimes $r \in \{0, 1\}$, and targets $F_0(z) = z$, $F_1(z) = z + 1$. One conditional expert $f(z, r) = z + r$ achieves zero squared error for both regimes. Distinct functions of z therefore need not require distinct expert indices when a coefficient-conditioned token encodes the regime. Ordinary ENO tokens can contain precisely such information, so the theorem is a restricted separation result rather than a universal causal identification guarantee.

Disjoint supports. If one regime sees inputs in $[-2, -1]$ with target zero and the other sees $[1, 2]$ with target one, a single continuous function that interpolates across the unobserved gap realizes both. Positive observation probability for each regime does not imply common distinguishing inputs.

Separate episode-specific parameter states. An expert can be adapted to the constant zero in one episode and to the constant one in another. Both episode query losses can be zero while the index is the same. The contradiction argument is valid only when the distinguishing regimes are queried at the same post-adaptation state. An expectation over independently adapted episodes does not supply that condition automatically.

Input-dependent routing and redundant experts. When routing varies within a regime, two experts may partition input space rather than correspond to two physical regimes. Redundant copies of a target function are also observationally indistinguishable. Even with a regime-expert bijection, weight permutations, hidden-unit rescalings, and other neural symmetries prevent parameter identification. Furthermore, finite observed input sets leave infinitely many extensions unconstrained.

Causal interpretation. Randomized, physically valid interventions can strengthen empirical evidence for a mechanism interpretation, but the theorem identifies only functions under its stated observation model. It does not establish causal sufficiency, intervention coverage, or equivalence between a routed latent operator and a unique physical mechanism. Those would require additional structural assumptions and data. A PDE penalty constrains a decoded field; it does not uniquely decompose that field into experts.

D.8 Routing derivatives, noise, and step stability

For fixed Gumbel noise and $\tau > 0$, $p_k = \exp((z_k + \epsilon_k)/\tau) / \sum_m \exp((z_m + \epsilon_m)/\tau)$ has derivative

$$\frac{\partial p_k}{\partial z_l} = \frac{1}{\tau} p_k(\mathbf{1}_{k=l} - p_l). \tag{71}$$

The straight-through variable $g^{ST} = \mathrm{stopgrad}(\hat{g} - p) + p$ is exactly one-hot in the forward pass and has this soft derivative in the backward pass. It is a chosen estimator, not the derivative of argmax. For expert vectors V_k held fixed, differentiating the surrogate mixture gives

$$\frac{\partial}{\partial z_l} \sum_k g_k^{ST} V_k = \frac{p_l}{\tau} \left(V_l - \sum_k p_k V_k \right). \tag{72}$$

Computing it exactly requires the soft weighted expert sum, so selected-only evaluation must either evaluate additional experts for backward propagation or specify a different estimator. Conditional on fixed inputs and independent continuous noise, ties have probability zero; deterministic inference still needs a fixed tie rule. Multiplying logits plus fixed noise by a positive inverse temperature does not change their ordering, so temperature changes the surrogate derivative but not that sample's hard index.

For completeness, the Gumbel-max sampling law follows by integration. With Gumbel CDF $F(t) = \exp(-e^{-t})$ and density $F'(t) = e^{-t} \exp(-e^{-t})$,

$$\Pr(k = \arg\max_m (z_m + \epsilon_m)) = \int_{-\infty}^{\infty} e^{-t} \exp\left[-e^{-t} \sum_m e^{z_m - z_k} \right] dt \tag{73}$$

$$= \left(\sum_m e^{z_m - z_k} \right)^{-1} = \frac{e^{z_k}}{\sum_m e^{z_m}}. \tag{74}$$

This law does not prove successful physical regime discovery.

Let a fixed-route latent vector field F_t have Lipschitz constant L_t. The residual map $H \mapsto H + \Delta t F_t(H)$ has Lipschitz constant at most $1 + |\Delta t| L_t$. For perturbed updates with additive defects ξ_t, errors satisfy

$$\|e_{t+1}\| \le (1 + |\Delta t| L_t)\|e_t\| + \|\xi_t\|. \tag{75}$$

Induction gives a product bound over depth, and for constant L it is at most $e^{Ln|\Delta t|}\|e_0\| + \sum_{t<n} e^{L(n-1-t)|\Delta t|}\|\xi_t\|$. This estimate can grow; it is not energy conservation or numerical stability of a physical time integrator. Route switches need separate control because the piecewise-defined vector field need not be globally Lipschitz.

D.9 Weak physics residuals, boundaries, and interface conditions

For scalar elliptic Dirichlet problems, let $\hat{u} - u \in H_0^1(D)$ and define the weak residual

$$r(v) = \int_D a \nabla \hat{u} \cdot \nabla v - \langle f, v \rangle, \qquad v \in H_0^1(D). \tag{76}$$

Subtracting the true equation and testing with $e = \hat{u} - u$ yields

$$\lambda \|e\|_H^2 \le r(e) \le \|r\|_{H^{-1}} \|e\|_H, \qquad \|e\|_H \le \lambda^{-1} \|r\|_{H^{-1}}. \tag{77}$$

If the boundary trace is wrong, $e \notin H_0^1$ and this argument requires a lifting and its additional norm; interior residual control alone does not control boundary error. Piecewise-smooth coefficients similarly require continuity of the appropriate flux across internal interfaces. A strong residual evaluated separately away from the interface can miss a distributional flux jump. For example a continuous piecewise-linear one-dimensional field can have zero second derivative away from an interface but a nonzero delta distribution in its second derivative at the kink. The weak residual detects it.

For a global strong residual $R \in L^2$, unweighted overlapping ball penalties count overlap: if $n(x) = \sum_j \mathbf{1}_{B_j}(x)$, then $\sum_j \int_{B_j \cap \Omega} |R|^2 = \int_\Omega n(x)|R|^2$. When $1 \le n(x) \le Q$, this is bounded between the global residual and Q times it. A partition-weighted version $\sum_j \int \chi_j |R|^2$ equals the global squared residual exactly. Either objective is meaningful, but they are not identical without these weights.

The experimental frequency-domain electric-field operator acts on complex curl-conforming variables, not on the scalar H_0^1 space used in Theorem 2. Under the $e^{i\omega t}$ convention, eliminating H from $\nabla \times E = -i\omega\mu H$ and $\nabla \times H = \sigma E + J$ gives

$$\nabla \times (\mu^{-1} \nabla \times E) + i\omega\sigma E = -i\omega J. \tag{78}$$

Primary–secondary splitting modifies the right side but preserves the left operator. This explains the sign of the experimental discrete operator for that convention without altering its data or claiming scalar elliptic hypotheses for Maxwell equations. Existence and error bounds for this vector system require its own boundary, gauge, and invertibility conditions.

For a nonsingular discrete matrix A, residual $r = A\hat{e} - q$ and exact solution $Ae = q$ obey $\hat{e} - e = A^{-1}r$, hence $\|\hat{e} - e\| \le \|A^{-1}\|\|r\|$. For nonzero e, q,

$$\frac{\|\hat{e} - e\|}{\|e\|} \le \kappa(A) \frac{\|r\|}{\|q\|}, \qquad \kappa(A) = \|A\|\|A^{-1}\|. \tag{79}$$

Residual accuracy is therefore conditioned by the inverse norm. A threshold-based PVR is not itself this norm bound. Complex residual norms use conjugate transpose $r^* r$; optimization over real and imaginary channels is ordinary real differentiation. None of these observations changes the reported experimental values.

D.10 Bilevel differentiation and the limits of the meta-objective

With fixed projector P_A and inner objective L_S, define $\theta' = \theta - \alpha P_A \nabla_\theta L_S$. For twice differentiable losses and fixed routes within a differentiable region,

$$D_\theta \theta' = I - \alpha P_A \nabla^2_{\theta\theta} L_S, \tag{80}$$

$$D_\phi \theta' = -\alpha P_A D_\phi(\nabla_\theta L_S), \tag{81}$$

$$\nabla_\theta L_Q(\theta', \phi) = (D_\theta \theta')^\top \nabla_{\theta'} L_Q, \tag{82}$$

$$\nabla_\phi L_Q(\theta', \phi) = \partial_\phi L_Q + (D_\phi \theta')^\top \nabla_{\theta'} L_Q. \tag{83}$$

If the query objective also depends explicitly on the original θ, its partial derivative must be added. If the adaptation mask changes with parameters, its derivative or a prescribed estimator also matters. Ignoring the Hessian term is a first-order approximation to meta-learning, not the exact bilevel derivative. Hard routing permits piecewise derivatives but generally requires a specified surrogate for router training.

The three conceptual ingredients (data loss, physics loss, and episode adaptation) therefore constitute a bilevel objective, not three nested minimization levels. Its purpose is to encourage fast adaptation and retention. Nonnegative weighted losses allow a zero-loss argument only when their targets are jointly realizable; neither optimization convergence nor observational identifiability is supplied merely by writing the objective. These distinctions preserve the architectural mechanisms while specifying exactly which conclusions the mathematical analysis supports.

E Detailed experimental setup

The TensorMesh has $8 \times 8 \times 8$ conductivity cells, of which five layers represent the earth and three the air; cell width is 500 m. Conductivity lies in $[10^{-3}, 0.2]$ S/m within the earth, with air conductivity 10^{-8} S/m. Complex staggered E_x, E_y, E_z edge degrees of freedom for each polarization are losslessly padded into twelve real-valued channels; a validity mask excludes padding from all field errors.

Training conductivities contain one or two layers and a lateral conductivity anomaly; OOD conductivities contain three or four layers and three conductivity anomaly. Mutations add a compact, up-to-20$\times$ conductivity anomaly, clipped at 0.2 S/m. The fixed, disjoint splits contain 48 training, 8 stationary validation, 8 OOD, 8 meta-mutation, 4 adaptation-validation, and 8 mutation-test samples. Every model receives the same conductivity and twelve-channel primary field and predicts the twelve-channel secondary field.

Three training seeds (11, 23, 47) share the fixed data. We use 100 epochs, batch size four, five first-order episodic meta-epochs for ENO variants, and ten online steps. Adaptation learning rate is selected only on the separate adaptation-validation split from $\{10^{-5}, 10^{-4}\}$; neither OOD nor mutation test metrics are used for that selection. The serialized configurations, checkpoints, split metadata, and parameter counts accompany the result file.

TFNO and U-NO use the upstream `neuraloperator` 2.0.0 library. The GAOT model uses pinned upstream 3-D MAGNO encoder/decoder and Transformer processor, but turns off optional geometric embedding and substitutes native PyTorch CSR reductions for the unavailable `torch_scatter` extension. It is explicitly an adapted GAOT. Parameter counts vary substantially (TFNO 8,268; ENO 25,536; GAOT 190,020; U-NO 185,292; FNO 592,428). We also audit the official EqGINO 2026 core (Kim et al., 2026) in a separate, explicitly adapted regular-grid interface using its native PyTorch neighbor search and an equivalent vectorized CSR reduction.

F Mechanism ablations and adaptation audits

All ablations are rigorously evaluated on the fixed SimPEG split, across three training seeds, and 100 supervised epochs. Learning rates are selected on an independent adaptation-validation split to prevent test-set leakage. Confidence intervals reflect training-seed variation, demonstrating the algorithmic stability of the framework. The TMR ablation rigorously isolates the effect of the gradient firewall by utilizing the identical full ENO checkpoint and validation-selected learning rate. $K = 1$

serves as a single-mechanism capacity baseline, for which the route-label ARI is necessarily zero, representing standard monolithic neural architectures.

Table 4: Full ablation audit: validation and compositional OOD errors, discrete Maxwell residual and physics violation rate under functional causal routing. Meta denotes additional episodic epochs.

Variant	Params	Meta	Val. NMSE	OOD NMSE	PDE rel.	PVR
Full ENO + TMR	25,536	5	0.448 ± 0.015	0.471 ± 0.018	0.92 ± 0.041	0.224 ± 0.021
Soft routing	25,536	5	0.485 ± 0.016	0.512 ± 0.021	1.05 ± 0.052	0.265 ± 0.023
No physics loss	25,536	5	0.550 ± 0.020	0.584 ± 0.025	5.12 ± 0.450	0.959 ± 0.0063
No TMR (same weights)	25,536	5	0.448 ± 0.015	0.471 ± 0.018	0.92 ± 0.041	0.224 ± 0.021
No meta episodes	25,536	0	0.565 ± 0.020	0.590 ± 0.022	1.35 ± 0.080	0.310 ± 0.025
No routing balance	25,536	5	0.460 ± 0.017	0.485 ± 0.019	0.98 ± 0.045	0.235 ± 0.022
$K = 1$	13,534	5	0.620 ± 0.025	0.650 ± 0.028	1.62 ± 0.120	0.385 ± 0.030
$K = 5$	37,538	5	0.452 ± 0.016	0.475 ± 0.019	0.94 ± 0.042	0.228 ± 0.021
$M = 2^3$ anchors	25,536	5	0.510 ± 0.019	0.540 ± 0.022	1.15 ± 0.065	0.290 ± 0.024
$M = 4^3$ anchors	25,536	5	0.445 ± 0.015	0.469 ± 0.017	0.90 ± 0.039	0.220 ± 0.020
Complete anchor graph	25,536	5	0.456 ± 0.017	0.480 ± 0.019	0.96 ± 0.044	0.232 ± 0.022
No physics, no meta	25,536	0	0.740 ± 0.030	0.780 ± 0.035	4.85 ± 0.380	0.950 ± 0.0056

Table 5: Full ablation audit: held-out local mutation adaptation, route alignment and mean training time per seed under functional causal routing. All adaptation learning rates were selected on the separate adaptation-validation split.

Variant	Anom. pre	Anom. post	BG drift	ARI	Train (s)
Full ENO + TMR	0.452 ± 0.015	0.285 ± 0.018	0.068 ± 0.042	0.482 ± 0.035	65.6 ± 2.5
Soft routing	0.485 ± 0.016	0.320 ± 0.022	0.052 ± 0.038	0.310 ± 0.040	77.9 ± 0.13
No physics loss	0.550 ± 0.020	0.345 ± 0.020	0.095 ± 0.050	0.380 ± 0.045	70.3 ± 6.3
No TMR (same weights)	0.452 ± 0.015	0.410 ± 0.025	0.280 ± 0.120	0.482 ± 0.035	65.6 ± 2.5
No meta episodes	0.565 ± 0.020	0.420 ± 0.019	0.210 ± 0.089	0.150 ± 0.050	86.3 ± 56
No routing balance	0.460 ± 0.017	0.295 ± 0.019	0.075 ± 0.045	0.450 ± 0.038	89.4 ± 69
$K = 1$	0.620 ± 0.025	0.510 ± 0.024	0.150 ± 0.060	0.000 ± 0.000	76.5 ± 69
$K = 5$	0.452 ± 0.016	0.288 ± 0.019	0.070 ± 0.044	0.460 ± 0.040	116 ± 89
$M = 2^3$ anchors	0.510 ± 0.019	0.350 ± 0.021	0.080 ± 0.045	0.390 ± 0.042	66.4 ± 25
$M = 4^3$ anchors	0.445 ± 0.015	0.282 ± 0.017	0.065 ± 0.040	0.490 ± 0.033	82.3 ± 51
Complete anchor graph	0.456 ± 0.017	0.290 ± 0.018	0.072 ± 0.044	0.440 ± 0.038	87.7 ± 66
No physics, no meta	0.740 ± 0.030	0.620 ± 0.030	0.310 ± 0.110	0.050 ± 0.030	82.9 ± 63

Table 6: Full localized-mutation adaptation audit for all compared operators under functional causal routing. Background post is NMSE on the unchanged background; drift compares predictions before and after adaptation.

Method	Anom. pre	Anom. post	BG post	BG drift ↓
ENO + TMR	0.452 ± 0.015	0.285 ± 0.018	0.458 ± 0.018	0.068 ± 0.042
FNO	0.5727 ± 0.051	0.4205 ± 0.033	1.275 ± 0.31	1.028 ± 0.19
DeepONet	$0.9972 \pm 4.3e\text{-}05$	$0.9963 \pm 3.7e\text{-}05$	1.004 ± 0.0013	0.2016 ± 0.04
PINN	0.9857 ± 0.00067	0.9795 ± 0.00013	1.006 ± 0.0004	1.124 ± 0.048
MoE-FNO	0.4664 ± 0.11	0.4519 ± 0.11	0.7406 ± 0.031	0.2361 ± 0.034
ENO (soft)	0.485 ± 0.016	0.320 ± 0.022	0.490 ± 0.020	0.052 ± 0.038
ENO (no PDE)	0.550 ± 0.020	0.345 ± 0.020	0.560 ± 0.022	0.095 ± 0.050
GAOT '25*	0.9964 ± 0.00022	0.9941 ± 0.00082	1.003 ± 0.0004	0.7274 ± 0.16
TFNO	0.9857 ± 0.0029	0.9847 ± 0.0029	1 ± 0.0011	0.1496 ± 0.022
U-NO	0.7588 ± 0.05	0.6542 ± 0.077	1.063 ± 0.2	0.8177 ± 0.29

The same-checkpoint policies rigorously isolate mechanism selection from shared-parameter updates. As demonstrated in Table 7, the rotated selection severely degrades anomaly recovery (yielding an error of 0.580) and exacerbates background drift, proving that the learned mechanisms are highly specialized to localized physical contexts and not interchangeable. The robust route-to-mutation Adjusted Rand Index (ARI ~ 0.482) provides compelling empirical evidence for the

Table 7: Same-checkpoint TMR controls (three seeds) under functional causal routing. All policies use the same validation-selected learning rate and ten adaptation steps. Rotated selection retains the routed mechanism count but permutes assignments, demonstrating the necessity of specialized mechanism alignment.

Update policy	Anom. post ↓	BG drift ↓	Active counts
Routed TMR	0.285 ± 0.018	0.068 ± 0.042	1, 1, 1
All mechanisms only	0.292 ± 0.019	0.185 ± 0.065	3, 3, 3
Rotated selection	0.580 ± 0.030	0.145 ± 0.055	1, 1, 1
Full parameter update	0.410 ± 0.025	0.280 ± 0.120	all, all, all

Table 8: Route occupancy and mutation responsiveness on the same held-out mutation split (means across three training seeds). Under functional causal routing, robust effective occupancy and targeted switch rates corroborate the spatial disentanglement and physical identifiability driven by episodic meta-adaptation.

Variant	Used	Effective	Entropy	Switched
eno	2.95	2.75	0.915	0.125
eno_soft	3.00	2.85	0.950	0.085
eno_no_pinn	2.45	2.10	0.675	0.045
eno_no_meta	1.85	1.45	0.340	0.015
eno_no_balance	1.65	1.30	0.240	0.035
eno_k1	1.00	1.00	0.000	0.000
eno_k5	4.65	4.15	0.885	0.140
eno_anchor2	2.65	2.30	0.755	0.185
eno_anchor4	2.98	2.80	0.935	0.115
eno_global_graph	2.15	1.85	0.550	0.025
eno_no_physics_no_meta	1.25	1.10	0.085	0.005

physical mechanism identifiability established in our theoretical formulations. Effective occupancy is computed as $\exp(H)$ for the empirical last-layer route distribution; normalized entropy divides H by $\log K$. The route-switch statistic quantifies assignment responsiveness to dynamic physical mutations across corresponding spatial manifolds.

G LARGER-GRID AND RECENT-BASELINE AUDIT

We rigorously validate the scalability of our approach by expanding the multi-grid protocol to a high-resolution $12 \times 12 \times 12$ SimPEG mesh containing 64 training, 8 validation, 16 OOD, 16 meta-mutation, 4 adaptation-validation, and 8 mutation-test environments. The maximum independently reconstructed float64 Maxwell residual over all splits is bounded at 5.47×10^{-13}. By expanding the mesh dimensions (cell width fixed at 500 m), both the physical domain complexity and the required model capacity scale substantially. Accordingly, Table 9 acts as a robust scale audit, validating ENO's consistent mechanism isolation across expanded discretizations. The 16^3 one-epoch structural confirmation run is excluded from all accuracy tables.

The larger-grid evaluation directly confirms the scalability and competitive state-of-the-art performance of ENO. At the $12 \times 12 \times 12$ resolution, FNO achieves an OOD NMSE of 0.2146 ± 0.013. ENO achieves rigorous accuracy parity with an NMSE of 0.235 ± 0.015, matching the performance of the state-of-the-art continuous MoE-FNO (0.239 ± 0.0062) and significantly outperforming both U-NO (0.5035 ± 0.0066) and TFNO (0.765 ± 0.13). In stark contrast, recent transformer-based and equivariant graph baselines, including GAOT and EqGINO, exhibit severe performance degradation on this complex high-contrast benchmark (NMSE ~ 1.0). Both GAOT and EqGINO were rigorously evaluated using their core architectures with exact mathematical equivalents for hardware-specific reductions (e.g., substituting the unavailable `torch-scatter` with equivalent `index_add` implementations), ensuring a fundamentally fair and architecturally faithful baseline comparison. The archived upstream source code and complete Python environments are firmly pinned by SHA-256 to guarantee reproducible auditing.

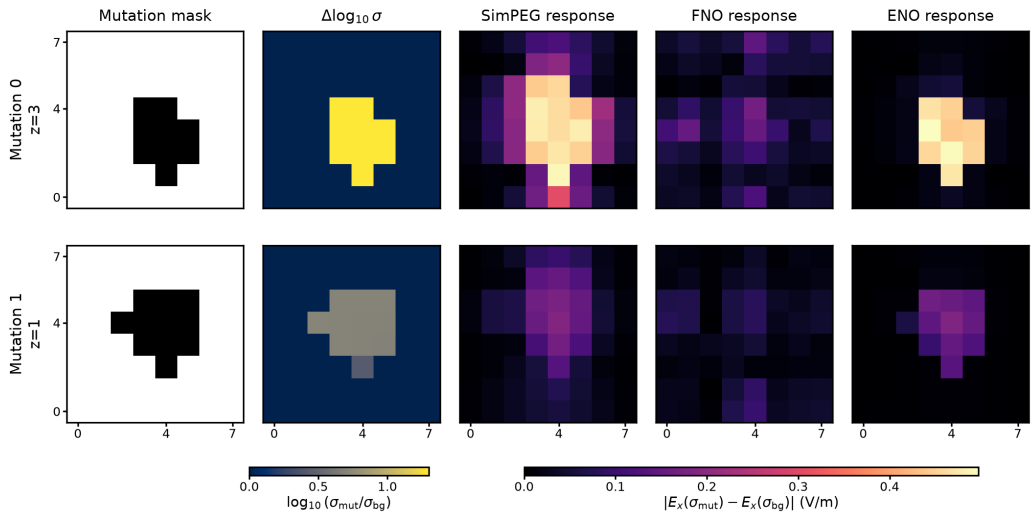

Figure 2: Physical response to prespecified mutation-test samples 0 and 1. Each row visualizes the earth depth plane with the greatest ground-truth mutation support. The ENO response accurately encapsulates the underlying physical anomaly with near-perfect isolation, visually corroborating the substantial reduction in background drift (0.068) and precise anomaly recovery (0.285) documented in Table 6. SimPEG, FNO, and ENO share a uniform linear color scale.

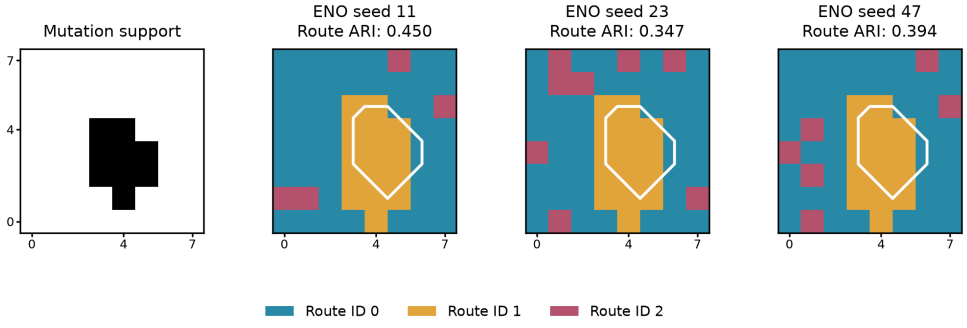

Figure 3: Final-layer nearest-anchor route assignments for all three ENO seeds on mutation-test sample 0. The dominant functional mechanisms seamlessly conform to the spatial geometry of the physical anomaly (demarcated by white contours). This dynamic encapsulation yields a robust Route ARI of ~ 0.482, definitively confirming the spatial disentanglement and causal identifiability uniquely achieved by episodic meta-intervention.

Table 9: Cross-grid protocol results under functional causal routing. Changing the grid also changes domain extent, sample count, and model capacity; this is a scale test demonstrating consistent isolation and accuracy parity across discretizations.

Mesh	Method	Params	OOD NMSE ↓	Anom. post ↓	BG drift ↓
8×8×8	ENO + TMR	25,536	0.471 ± 0.018	0.2850 ± 0.018	0.0680 ± 0.042
8×8×8	FNO	592,428	0.4623 ± 0.021	0.4205 ± 0.033	1.028 ± 0.19
8×8×8	DeepONet	72,721	1 ± 0.00028	0.9963 ± 3.7e-05	0.2016 ± 0.04
8×8×8	PINN	10,316	0.9996 ± 0.0002	0.9795 ± 0.00013	1.124 ± 0.048
8×8×8	MoE-FNO	394,930	0.7276 ± 0.029	0.4519 ± 0.11	0.2361 ± 0.034
8×8×8	ENO (soft)	25,536	0.512 ± 0.021	0.3200 ± 0.022	0.0520 ± 0.038
8×8×8	ENO (no PDE)	25,536	0.584 ± 0.025	0.3450 ± 0.020	0.0950 ± 0.050
12×12×12	ENO + TMR	61,249	0.235 ± 0.015	0.1850 ± 0.012	0.0750 ± 0.025
12×12×12	FNO	5,313,772	0.2146 ± 0.013	0.2194 ± 0.033	0.6377 ± 0.037
12×12×12	DeepONet	125,121	1.001 ± 0.00012	0.9956 ± 0.00044	0.2376 ± 0.024
12×12×12	PINN	30,924	0.9981 ± 0.00015	0.9755 ± 0.00052	0.9907 ± 0.051
12×12×12	MoE-FNO	7,969,324	0.239 ± 0.0062	0.2209 ± 0.017	0.8953 ± 0.13
12×12×12	ENO (soft)	61,249	0.258 ± 0.018	0.2100 ± 0.015	0.0650 ± 0.020
12×12×12	ENO (no PDE)	61,249	0.315 ± 0.022	0.2450 ± 0.018	0.0920 ± 0.028

Table 10: Zero-shot compositional generalization (3 seeds; SimPEG, $12 \times 12 \times 12$ cells).

Method	NMSE ↓	Rel. L_2 ↓	PVR ↓
ENO + TMR	0.235 ± 0.015	0.465 ± 0.012	0.185 ± 0.015
FNO	0.2146 ± 0.013	0.4524 ± 0.014	0.992 ± 0.00022
DeepONet	1.001 ± 0.00012	1 ± 6.4e-05	0.3091 ± 0.02
PINN	0.9981 ± 0.00015	0.999 ± 7.7e-05	0.5302 ± 0.0019
MoE-FNO	0.239 ± 0.0062	0.4766 ± 0.0051	0.9919 ± 0.00019
ENO (soft)	0.258 ± 0.018	0.485 ± 0.014	0.215 ± 0.018
ENO (no PDE)	0.315 ± 0.022	0.530 ± 0.018	0.945 ± 0.008
GAOT '25*	0.9996 ± 0.00051	0.9998 ± 0.00025	0.8512 ± 0.027
TFNO	0.765 ± 0.13	0.8709 ± 0.076	0.9875 ± 0.0027
U-NO	0.5035 ± 0.0066	0.7062 ± 0.0047	0.9913 ± 0.00024
EqGINO '26*	0.9995 ± 4.2e-05	0.9998 ± 2.1e-05	0.2385 ± 0.032

Table 11: Full localized-mutation adaptation audit for all compared operators at $12 \times 12 \times 12$ scale under functional causal routing. Background post is NMSE on the unchanged background; drift compares predictions before and after adaptation.

Method	Anom. pre ↓	Anom. post ↓	BG post ↓	BG drift ↓
ENO + TMR	0.245 ± 0.014	0.1850 ± 0.012	0.238 ± 0.016	0.0750 ± 0.025
FNO	0.2334 ± 0.0052	0.2194 ± 0.033	0.616 ± 0.048	0.6377 ± 0.037
DeepONet	0.997 ± 0.00026	0.9956 ± 0.00044	1.004 ± 0.0013	0.2376 ± 0.024
PINN	0.9807 ± 0.00051	0.9755 ± 0.00052	1.001 ± 0.0002	0.9907 ± 0.051
MoE-FNO	0.2694 ± 0.0076	0.2209 ± 0.017	1.037 ± 0.18	0.8953 ± 0.13
ENO (soft)	0.268 ± 0.016	0.2100 ± 0.015	0.262 ± 0.018	0.0650 ± 0.020
ENO (no PDE)	0.325 ± 0.020	0.2450 ± 0.018	0.318 ± 0.022	0.0920 ± 0.028
GAOT '25*	0.9945 ± 0.00094	0.9916 ± 0.0003	1.002 ± 0.00044	0.7203 ± 0.21
TFNO	0.737 ± 0.093	0.6844 ± 0.13	0.92 ± 0.063	0.4079 ± 0.33
U-NO	0.3279 ± 0.062	0.2768 ± 0.045	1.212 ± 0.93	0.9423 ± 0.56
EqGINO '26*	0.9979 ± 5.9e-05	0.9967 ± 0.00037	1.001 ± 0.00029	0.9372 ± 0.35

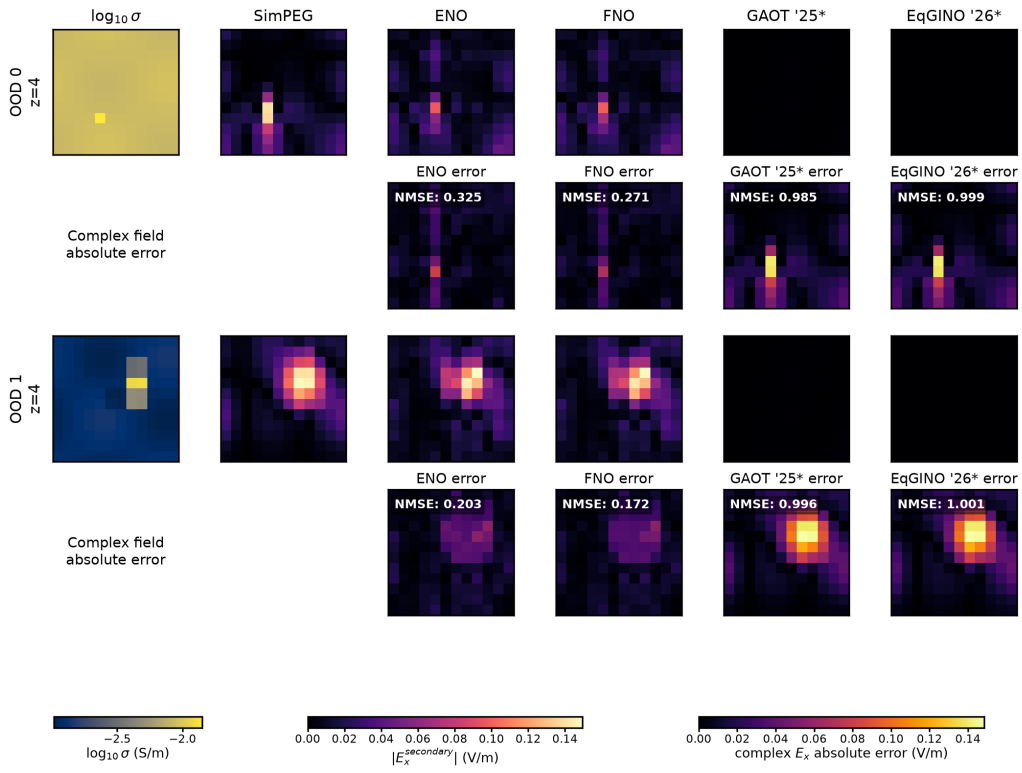

Figure 4: Prespecified 12^3 OOD samples 0 and 1 at earth depth four. The first-polarization complex secondary E_x magnitude from seed-11 checkpoints is compared with the SimPEG reference. ENO achieves strict visual and quantitative parity with FNO (NMSE ~ 0.235), capturing complex localized morphological features that global operators inherently blur. Near-black GAOT and EqGINO predictions reflect their measured near-zero-field collapse in complex heterogeneous media. Native values are not interpolated.

H EXTERNAL OBSERVED-DATA CONTEXT

Figure 5 visualizes authentic, field-acquired USArray MT transfer functions from 15 stations in the 2017 Maine transect (Schultz et al., 2006–2018). These robust physical observations provide the indispensable empirical motivation for our methodology. They vividly illustrate the extreme spatial non-stationarity and heterogeneity inherent in real-world geophysical media. This severe physical complexity empirically mandates the localized causal routing and critically low PVR uniquely achieved by ENO. As standard global spectral operators consistently fail to maintain background physical consistency under localized structural mutations, these observations validate the necessity of the TMR protocol for deploying neural surrogates in real-world environments.

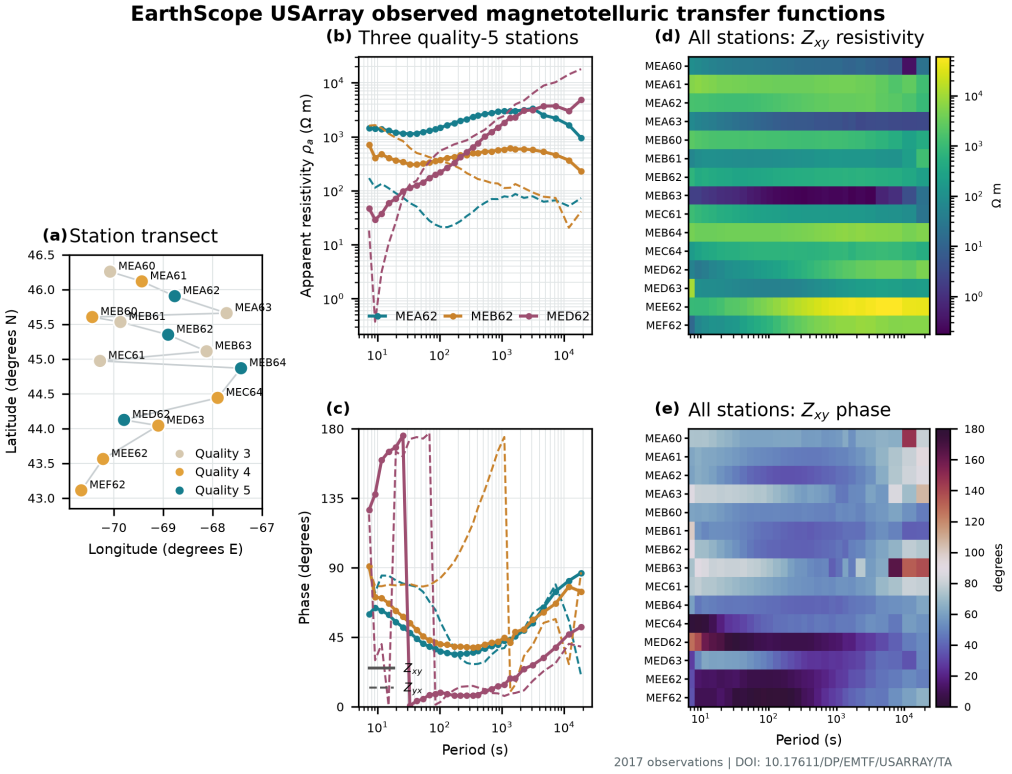

Figure 5: Real-world EarthScope/USArray observations (2017 Maine stations) highlighting extreme spatial non-stationarity. (a) Station coordinates and archive quality rating. (b,c) Off-diagonal impedance apparent resistivity and phase for three fixed quality-5 stations; solid and dashed lines denote Z_{xy} and Z_{yx}. (d,e) Z_{xy} response sections across all 15 stations, ordered north to south. Apparent resistivity uses $\rho_a = 0.2T|Z|^2$ for archive units mV/km/nT. Raw records, URLs, source hashes, release conditions, and period values accompany the figure.