CAUSAL STATE SPACE MODELS: ROUTING DYNAMIC MECHANISMS FOR NON-STATIONARY INVERSE PROBLEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

State Space Models (SSMs) provide efficient representations for continuous spatial-temporal fields, yet standard global SSMs suffer from prior entanglement under non-stationary inverse scattering. A localized out-of-distribution (OOD) anomaly shifts the shared transition matrix during online adaptation, corrupting the reconstruction of stationary backgrounds. To address this, we propose the Causal State Space Model (Causal-SSM). By formulating the inverse mapping as a switched linear system via Dynamic Mechanism Routing (DOR), the model partitions continuous data into a library of independent causal mechanisms. During adaptation to physical shifts, our Targeted Mechanism Revision (TMR) confines gradient updates strictly to active anomalous mechanisms while freezing shared parameters and boundary routing. Evaluated on dynamic scattering benchmarks and Institut Fresnel measurements, Causal-SSM achieves robust zero-shot OOD generalization. It successfully assimilates localized anomalies while preserving background fidelity (BRR ≈ 1.015), avoiding the severe degradation typical of global-update baselines.

1 INTRODUCTION

Inverse scattering problems, which aim to reconstruct heterogeneous physical media from scattered electromagnetic observations, are fundamental to computational electromagnetics and probabilistic imaging Lou et al. (2021); Cheng et al. (2026). Traditional numerical solvers struggle with the high computational cost of large-scale continuous domains. Physics-informed neural representations and operators offer promising alternatives Lu et al. (2021); Wang et al. (2022), where global architectures like Fourier Neural Operators (FNOs) and transformers effectively capture long-range dependencies in stationary regimes Li et al. (2020); Wu et al. (2024).

Recently, State Space Models (SSMs), particularly Mamba, have established a new scalable paradigm for sequence modeling by offering linear-time complexity and continuous state-space representations Gu & Dao (2023). This has inspired a wave of SSM-based PDE solvers, including Mamba Neural Operator (MNO), DeepOMamba, and Alias-free MNO Cheng et al. (2025); Hu et al. (2025); Zheng et al. (2024). However, when deployed in non-stationary inverse scattering environments containing highly nonlinear out-of-distribution (OOD) mutations, these global SSMs exhibit a critical vulnerability. They project diverse local scattering behaviors into a monolithic, global continuous-time transition matrix. During online adaptation to localized anomalies, a shared transition update can alter background states and degrade reconstruction of the learned physical prior for the stationary background.

To overcome this prior entanglement, the model must respect causality and disentangle the underlying independent physical mechanisms Schölkopf et al. (2021); Audenaert et al. (2025). While recent advances in object-centric world models and generative probabilistic operators attempt to isolate entities Nakano et al. (2023); Zang & Koutsourelakis (2025), they lack the structural guarantees required for continuous wavefield inversion Zou et al. (2025). Furthermore, fast adaptation via meta-learning and test-time generalization has proven effective for transferring hidden physics Finn et al. (2017); Zhang et al. (2023); Serrano et al. (2026), yet existing approaches still suffer from temporal leakage when updating monolithic transition matrices.

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.

To formally resolve this, we propose the **Causal State Space Model (Causal-SSM)**, a novel architecture that formulates the inverse problem as a dynamically routed Switched Linear System. By tokenizing continuous spatial data and introducing Dynamic Mechanism Routing (DOR), Causal-SSM implicitly partitions the inverse mapping into a discrete library of localized, independent causal mechanisms. Coupled with our Targeted Mechanism Revision (TMR) protocol, gradient updates during online adaptation are strictly isolated to the specific anomalous mechanisms, preserving background parameters under the conditions specified in Theorem 2.

Our main contributions are:

- **Causal-SSM Architecture:** We introduce Dynamic Mechanism Routing for continuous state spaces, reducing interference between routed mechanisms in non-stationary inverse scattering.
- **TMR:** We prove conditional background-parameter isolation during fast online adaptation, addressing the catastrophic forgetting bottleneck inherent in global spectral operators.
- **Information-Theoretic Identifiability:** We propose a probabilistic meta-inference objective that leverages episodic meta-learning Finn et al. (2017); Zhang et al. (2023) and contrastive regularizers to encourage regime-sensitive features; exact identification requires additional assumptions.
- **SOTA Performance:** Evaluated on highly complex 2D/3D inverse scattering and real-world Institut Fresnel measurements Cheng et al. (2026), Causal-SSM achieves state-of-the-art zero-shot OOD generalization and near-perfect background retention.

2 Related Work

Neural Operators for PDEs and Inverse Problems. Neural operators have established a new standard for learning resolution-invariant mappings between infinite-dimensional function spaces. DeepONet established universal approximation guarantees for nonlinear operators Lu et al. (2021), while FNO introduced highly efficient global convolution via the fast Fourier transform Li et al. (2020). Subsequent innovations, such as Transolver for complex geometries Wu et al. (2024) and DGenNO for generative probabilistic modeling Zang & Koutsourelakis (2025), have further expanded the capacity of neural solvers. For large-scale geophysical and electromagnetic inversion, architectures like ANFWI-HNO have demonstrated the utility of operator learning in ambient noise full waveform inversion Zou et al. (2025); Lou et al. (2021). However, these globally aggregated operators fundamentally entangle distinct physical regimes, failing to maintain structural stability under local topological mutations.

State Space Models in Scientific Machine Learning. The introduction of Mamba Gu & Dao (2023) has catalyzed the adoption of SSMs in scientific computing due to their hardware-aware parallel scans and continuous-time ODE formulations. Mamba Neural Operator (MNO) and DeepOMamba successfully adapted selective state spaces for spatio-temporal PDEs Cheng et al. (2025); Hu et al. (2025). Recent extensions, such as Alias-free MNO, address continuous-discrete aliasing in high-frequency spectral domains Zheng et al. (2024). Despite their computational supremacy, these models rely on continuous, shared linear projections that can interpolate between distinct physical mechanisms near regime boundaries during non-stationary transitions. Causal-SSM replaces this continuous selection with a discrete, causally routed mechanism library, specifically designed for heterogeneous continuous spatial data.

Causal Disentanglement and Meta-Adaptation. Discovering and isolating independent causal mechanisms is a central pursuit in robust representation learning Schölkopf et al. (2021); Audenaert et al. (2025). In object-centric state estimation, interaction-based disentanglement has shown promise in isolating dynamic entities Nakano et al. (2023). For physics-informed networks, respecting causality in gradient flow is critical to prevent training collapse Wang et al. (2022). When dealing with out-of-distribution physical systems, test-time generalization through operator splitting Serrano et al. (2026) and fast meta-adaptation Finn et al. (2017); Zhang et al. (2023) provide essential frameworks for online error correction. Our TMR protocol synthesizes these principles, leveraging episodic meta-updates to achieve surgical adaptation in continuous fields without contaminating the isolated background prior.

3 CAUSAL STATE SPACE MODELS (CAUSAL-SSM)

Standard global SSMs entangle heterogeneous inverse-scattering regimes in one transition matrix, so local OOD adaptation can corrupt the stationary background prior. Causal-SSM instead uses a dynamically routed library of isolated mechanisms; the section develops prior entanglement, DOR, TMR, and probabilistic meta-inference in turn. The complete architectural motivation is preserved in Appendix A.1.

3.1 STATE SPACE MODELS AND PRIOR ENTANGLEMENT IN INVERSE PROBLEMS

We reconstruct a medium $\mathbf{m}$ from noisy scattered-field observations $\mathbf{u}_s = \mathcal{F}(\mathbf{m}) + \boldsymbol{\epsilon}$ using a standard discretized SSM, $h_k = \bar{\mathbf{A}} h_{k-1} + \bar{\mathbf{B}} x_k$, whose transition matrix $\bar{\mathbf{A}}$ serves as a global physical prior. Under a localized OOD intervention, online adaptation produces $\bar{\mathbf{A}}' = \bar{\mathbf{A}} + \delta\bar{\mathbf{A}}$, thereby perturbing the prior shared by every token. Appendix A.10 preserves the complete continuous-time formulation and ZOH derivation.

Proposition 1 (Spectral Entanglement in Global SSMs)**.** *Let x_{bg} be an observation sequence corresponding entirely to the unmutated background domain $\Omega \setminus \Omega_{mut}$. Under a localized parameter perturbation $\delta\bar{\mathbf{A}}$ induced by adapting to an anomaly, the background latent trajectory can change. With $\rho = \|\bar{\mathbf{A}}\|_2$ and $\eta = \|\delta\bar{\mathbf{A}}\|_2$, its error at step L satisfies:*

$$\|h'_L - h_L\|_2 \leq \eta \|\bar{\mathbf{B}}\|_2 \sum_{j=1}^{L-1} j(\rho+\eta)^{j-1} \|x_{bg,L-j}\|_2, \tag{1}$$

provided $h_0 = h'_0 = 0$, the inputs coincide, and $\bar{\mathbf{B}}$ is held fixed. This is an upper bound, not a claim that every perturbation causes exponential divergence.

Proof sketch. Write $U = \bar{\mathbf{A}} + \delta\bar{\mathbf{A}}$ and $V = \bar{\mathbf{A}}$. The identity $U^j - V^j = \sum_{m=0}^{j-1} U^{j-1-m}(U-V)V^m$ and the zero-initial-state convolution give the bound by submultiplicativity. Appendix B gives the full proof; Appendix C.6 treats nonzero initial states and joint input-projection updates.

Proposition 1 shows how a shared transition-matrix update can affect background states even when adaptation targets a localized anomaly. The extent of interference depends on the perturbation, the inputs, and the transition dynamics; the bound alone does not imply inevitable forgetting or exponential growth.

3.2 DYNAMIC MECHANISM ROUTING FOR SWITCHED STATE SPACES

Unlike Mamba's shared continuous projections, **DOR** selects from structurally isolated causal mechanisms, casting the inverse problem as a Switched Linear System; the full comparison is given in Appendix A.2.

The Discrete Inverse Mechanism Library. Instead of learning a single global matrix set or a shared continuous projection, Causal-SSM maintains a discrete library of K independent, highly specialized state transition mechanisms:

$$\mathcal{P} = \{\boldsymbol{\theta}_k = (\mathbf{A}_k, \mathbf{B}_k, \mathbf{C}_k, \boldsymbol{\Delta}_k)\}_{k=1}^{K}. \tag{2}$$

The mechanisms have disjoint parameter blocks; whether they learn distinct physical regimes is a conditional identifiability question, not a consequence of initialization. For a continuous sequence of scattered field observations $\mathbf{X} = \{x_1, x_2, \ldots, x_L\}$, the model dynamically routes each token x_i to a specific mechanism.

Hardware-Aware Hard Routing. We define a lightweight gating network $h_\phi(x_i) \to \mathbb{R}^K$ that outputs routing logits for the i-th token. To select one parameter block in the forward pass, we use hard Gumbel routing with a straight-through training surrogate:

$$g_{i,k} = \frac{\exp((a_{i,k} + \epsilon_{i,k})/\tau)}{\sum_{m=1}^{K} \exp((a_{i,m} + \epsilon_{i,m})/\tau)}, \qquad \hat{g}_{i,k} = \mathbf{1}\{k = \arg\max_m (a_{i,m} + \epsilon_{i,m})\}. \tag{3}$$

Here $a_i = h_\phi(x_i)$, $\epsilon_{i,k}$ are independent standard Gumbel variables, and $\tau > 0$. Training uses $g_i + \mathrm{stopgrad}(\hat{g}_i - g_i)$; online revision fixes the realized hard routes. Temperature changes the surrogate derivative, not the hard argmax for fixed noise. Appendix C.7 details this distinction. The selected continuous-time parameters are:

$$\mathbf{A}^{(i)} = \sum_{k=1}^{K} \hat{g}_{i,k}\mathbf{A}_k, \quad \mathbf{B}^{(i)} = \sum_{k=1}^{K} \hat{g}_{i,k}\mathbf{B}_k, \quad \mathbf{C}^{(i)} = \sum_{k=1}^{K} \hat{g}_{i,k}\mathbf{C}_k, \quad \mathbf{\Delta}^{(i)} = \sum_{k=1}^{K} \hat{g}_{i,k}\mathbf{\Delta}_k. \tag{4}$$

Applying the Zero-Order Hold (ZOH) discretization (Eq. 18) locally at step i, we obtain the discrete-time matrices $\bar{\mathbf{A}}^{(i)}$ and $\bar{\mathbf{B}}^{(i)}$.

Exact local dynamics. Writing $D_i = \mathbf{\Delta}^{(i)}$, $M_i = D_i\mathbf{A}^{(i)}$, and $G_i = D_i\mathbf{B}^{(i)}$, the normalized local ODE $\dot{h} = M_i h + G_i x_i$ on a unit interval gives

$$h_i = e^{M_i} h_{i-1} + \left(\int_0^1 e^{uM_i}\, du \right) G_i x_i, \qquad y_i = \mathbf{C}^{(i)} h_i. \tag{5}$$

For a scalar duration $D_i = \Delta_i I$, this is the usual ZOH discretization of $\dot{h} = \mathbf{A}^{(i)}h + \mathbf{B}^{(i)}x_i$. For diagonal state-wise steps it defines the rescaled generator explicitly. The integral equals $\varphi_1(M_i)$ even for singular M_i, so neither an inverse matrix nor a commutation assumption is needed. In the diagonal case, the n-th input row is multiplied by $d_{i,n}\varphi_1(d_{i,n}a_{i,n})$, whose limit at $a_{i,n} = 0$ is $d_{i,n}$. This matters when differentiating step sizes during revision, because both the transition and the input projection change. The sensitivity proposition isolates a discrete-transition update; Appendix C.6 supplies the bound when both discrete matrices change.

Crucially, because $\hat{g}_{i,k} \in \{0, 1\}$, this operation selects exact parameter sets without linear mixing. For fixed state size, the resulting affine recurrences admit an associative prefix scan with $\mathcal{O}(L)$ work and $\mathcal{O}(\log L)$ parallel depth; the scan requires only $\bar{\mathbf{A}}^{(i)}$ and $\bar{\mathbf{B}}^{(i)}x_i$ at each position. Parameter-gradient masking can be applied after the scan backward pass (Theorem 4).

Theoretical Property: Approximation near Regime Boundaries. Discrete routing can represent a discontinuous matrix selector across a shared regime boundary when the router assigns the regimes correctly. Theorem 1 establishes the corresponding approximation limitation for a Lipschitz continuous selector; it does not bound training gradients.

Theorem 1 (Boundary Approximation Limitation for Continuous Selectors)**.** *Let $\mathcal{X}_{obs} \subset \mathbb{R}^d$ contain a common boundary approached from two physical regimes with optimal transition matrices $\bar{\mathbf{A}}_u^*$ and $\bar{\mathbf{A}}_v^*$ satisfying $\|\bar{\mathbf{A}}_u^* - \bar{\mathbf{A}}_v^*\|_F = \delta > 0$. Assume observations on opposite sides can be arbitrarily close, while the target matrix selector jumps at that boundary. Let $f_{vanilla} : \mathcal{X}_{obs} \to \mathbb{R}^{N\times N}$ represent a continuous selection projection (as in standard Mamba) with Lipschitz constant L_f. Then, for opposite-side observations at distance $2r$, at least one matrix approximation error is at least $(\delta - 2L_f r)/2$. Exact representation of the jump over arbitrarily small r is impossible for finite L_f. A hard router can represent the selector if it assigns both regimes correctly and the library can realize both target matrices; training need not discover this assignment.*

Proof. See Appendix A.12; the boundary-error derivation also appears in Appendix C.

Theorem 1 gives a conditional approximation advantage for hard routing near a discontinuous regime boundary. By selecting separate mechanism matrices, Causal-SSM can avoid interpolating between distinct transitions when the routing assignments are correct.

3.3 Targeted Mechanism Revision for Prior Protection

To prevent localized anomaly gradients from contaminating the stationary background under BPTT, **TMR** explicitly severs the gradient graph between orthogonal physical priors. The complete motivation is preserved in Appendix A.3.

Isolating Backpropagation Through Time. Consider a scattered field observation sequence of length L, where a subset of contiguous tokens $\mathcal{I}_{mut} \subset \{1, \ldots, L\}$ corresponds to the localized

anomaly. For any token step $t \in \mathcal{I}_{mut}$, the routing network h_ϕ correctly assigns an anomaly-specific mechanism k_{mut} via the Gumbel-Softmax estimator $\hat{g}_{t,k_{mut}} = 1$.

During online adaptation, the anomaly loss $\mathcal{L}_{mut}$ induces a BPTT adjoint state; its complete recurrence is given in Appendix A.4. TMR detaches the pre-anomaly boundary state $h_{t_{start}-1}$ during the inner-loop update and holds the routing assignments fixed, preventing anomaly gradients from entering the pre-anomaly computation graph. With $y_t = C^{(t)} h_t$, set $r_t = \nabla_{y_t} \mathcal{L}_{mut}$ and $\lambda_t = (C^{(t)})^T r_t + (\bar{A}^{(t+1)})^T \lambda_{t+1}$, with the terminal future contribution zero. For a vectorized mechanism block ϑ_k, the exact detached-graph gradient is

$$\nabla^{TMR}_{\vartheta_k} \mathcal{L}_{mut} = \sum_{t \in \mathcal{I}_{mut}} \hat{g}_{t,k} \left[(J^A_{t,k})^T \operatorname{vec}(\lambda_t h_{t-1}^T) + (J^B_{t,k})^T \operatorname{vec}(\lambda_t x_t^T) + (J^C_{t,k})^T \operatorname{vec}(r_t h_t^T) \right]. \quad (6)$$

Here $J^A_{t,k} = \partial \operatorname{vec}(\bar{A}_k) / \partial \vartheta_k$ and similarly for $\bar{B}_k, C_k$ are local mechanism Jacobians, evaluated with states and routes held fixed. They include derivatives through ZOH and the step size. Appendix C.8 derives each contraction and explains why the direct output term does not double-count state dependence. A mechanism unused throughout the loss interval has zero gradient under the boundary-detach assumptions.

Theoretical Guarantee: Strict Prior Isolation in Latent State Space. TMR guarantees absolute parameter invariance despite recurrent hidden-state dependence; the complete motivating argument is in Appendix A.5.

Theorem 2 (Strict Prior Isolation Bound in Latent State Space)**.** *Let $\mathbf{X}$ be an observation sequence with a localized OOD anomaly at indices $t \in \mathcal{I}_{mut} = [t_{start}, t_{end}]$. Let k_{mut} be assigned exclusively to $\mathcal{I}_{mut}$ and $k_{bg} \neq k_{mut}$ govern every prefix step $t < t_{start}$, with disjoint parameter blocks and a fixed initial state. Let $\mathcal{L}_{mut}$ be an adaptation loss evaluated exclusively on the anomalous interval. Assume fixed routes and inputs, a detached boundary state, frozen shared parameters, no direct background-parameter penalty in the loss, and an optimizer that leaves inactive parameters and their state unchanged. If the TMR protocol is applied, the total gradient of the loss with respect to the continuous-time background parameters $\boldsymbol{\theta}_{bg} = (\mathbf{A}_{bg}, \mathbf{B}_{bg}, \mathbf{C}_{bg}, \boldsymbol{\Delta}_{bg})$ evaluates strictly to zero:*

$$\left\| \nabla^{TMR}_{\boldsymbol{\theta}_{bg}} \mathcal{L}_{mut} \right\|_F \equiv 0. \quad (7)$$

Furthermore, the pre-anomaly states $\{h_t\}_{t=1}^{t_{start}-1}$ remain invariant under repeated updates satisfying these conditions.

Proof. See Appendix A.13; the conditional gradient-isolation derivation also appears in Appendix C.

Theorem 2 guarantees preservation of background parameters and pre-anomaly states under the specified boundary-detach, routing, shared-parameter, and optimizer conditions. Reconstruction quality and retention on other background inputs remain empirical questions.

Why the boundary and optimizer conditions matter. Let $s = t_{start} - 1$ and $D_t = \partial h_t / \partial \vartheta_{bg}$ in the adaptation graph. Inside the anomalous interval, disjoint routing nullifies the local background Jacobians, leaving $D_t = \bar{A}^{(t)} D_{t-1}$. Detachment sets $D_s = 0$, so induction yields $D_t = 0$. Without detachment the same recurrence can transmit a nonzero D_s. Likewise, a zero derivative does not prevent momentum or weight decay from changing an inactive block; the entire optimizer update must exclude that block. After the anomalous interval, a changed state can propagate through unchanged background matrices, so the theorem protects the prefix and parameters rather than every downstream prediction. These distinctions separate an exact computational-graph statement from an empirical reconstruction claim.

3.4 Probabilistic Meta-Inference for Ill-Posed Inverse Manifolds

Because inverse scattering is ill-posed, we augment DOR and TMR with variational inference and an information-theoretic meta-objective; the complete motivation is preserved in Appendix A.6.

Variational Switched State Spaces. The categorical route z_i and hidden state h_i define a variational switched sequence model (Appendix A.7); its β-weighted variational objective is:

$$\mathcal{L}_{\beta\text{VI}}(\theta,\phi) = \mathbb{E}_{q_\phi(z_{1:L}|\mathbf{X})}\left[\sum_{i=1}^{L}\log p_\theta(\mathbf{m}_i \mid h_i, z_i)\right] - \beta_{KL} D_{KL}\left(q_\phi(z_{1:L} \mid \mathbf{X}) \,\|\, \prod_{i=1}^{L} p(z_i)\right), \quad (8)$$

where $q_\phi(z_{1:L} \mid \mathbf{X}) = \prod_i q_\phi(z_i \mid x_i)$ is the assumed factorized routing posterior, each h_i depends on its routed history, $p(z_i)$ is a uniform categorical prior, and β_{KL} controls the information bottleneck capacity. At $\beta_{KL} = 1$ this is the conditional ELBO; $\beta_{KL} \geq 1$ also gives a lower bound, whereas $0 \leq \beta_{KL} < 1$ need not (Appendix C.10).

The conditional model is $p_\theta(\mathbf{m}, z \mid \mathbf{X}) = p(z)\prod_i p_\theta(\mathbf{m}_i \mid H_{\theta,i}(\mathbf{X}, z_{1:i}), z_i)$. Thus the hidden state is a deterministic function of routed history, not an independent variational variable. Bayes' rule gives the exact gap

$$\log p_\theta(\mathbf{m} \mid \mathbf{X}) - \mathcal{L}_{1\text{VI}} = D_{KL}\big(q_\phi(z \mid \mathbf{X}) \,\|\, p_\theta(z \mid \mathbf{m}, \mathbf{X})\big) \geq 0. \quad (9)$$

The inference network may omit $\mathbf{m}$ even though the true posterior conditions on it; this restriction can loosen the bound. A Gaussian decoder requires positive-definite covariance, and its variance models conditional dispersion rather than automatically calibrated parameter uncertainty.

Information-Theoretic Disentanglement via Meta-Intervention. To prevent single-mechanism representation collapse, Z should retain high mutual information with the physical regimes and little with observational noise (Appendix A.8).

We integrate the Variational Causal-SSM into an episodic meta-intervention framework. We synthesize episodes containing a stationary background distribution $\mathcal{D}_{bg}$ and injected OOD anomalies $\delta\mathbf{m} \sim \mathcal{D}_{mut}$. During the outer-loop meta-optimization, we impose a regime-supervised InfoNCE-style contrastive regularizer on the routing posteriors across the episode batch $\mathcal{B}$:

$$\mathcal{L}_{MI}(\phi) = -\frac{1}{|\mathcal{B}|}\sum_{x\in\mathcal{B}} \log \frac{\exp(s_+/\tau)}{\exp(s_+/\tau) + \sum_{x^-}\exp(s_-/\tau)}, \quad (10)$$

where $s_+ = s_\phi(x, x^+)$ and $s_- = s_\phi(x, x^-)$ are routing-feature similarities; positives share a regime and negatives have different regime labels. These negatives are not marginal samples, so the standard InfoNCE mutual-information bound does not automatically apply. The exact similarity and the additional partition/balance penalties are implementation specifications; they are not inferred from identifiability alone.

The core meta-objective combines variational reconstruction, an adaptation term $\mathcal{L}_{TMR}$ evaluated after support-set TMR updates, and contrastive regularization. Appendix C.11 makes the support/query dependence explicit:

$$\mathcal{J}_{meta}(\theta,\phi) = -\mathcal{L}_{\beta\text{VI}}(\theta,\phi) + \lambda_{TMR}\mathcal{L}_{TMR}(\theta,\phi) + \lambda_{MI}\mathcal{L}_{MI}(\phi). \quad (11)$$

Conditional Identifiability of Physical Regimes. The contrastive objective promotes regime-sensitive routes. Identifiability of physical regimes requires additional assumptions on the observation distribution and routing capacity.

Theorem 3 (Conditional Regime Identifiability)**.** *Let $Z^* \in \{1,\dots,K^*\}$ denote physical regimes and let $Z \in \{1,\dots,K\}$ denote learned routes. Suppose $K = K^*$, every regime has positive probability, Z^* is determined by X, and the learned deterministic route satisfies both $H(Z \mid Z^*) = 0$ and $H(Z^* \mid Z) = 0$. Then a permutation π satisfies $Z = \pi(Z^*)$ almost surely. Here q_ϕ denotes this deterministic evaluation route, not the finite-logit noisy training sampler. Consequently:*

$$q_\phi(Z = \pi(z^*) \mid X) = p(Z^* = z^* \mid X) \quad \textit{for each } z^*\textit{, almost surely}. \quad (12)$$

The objective in Eq. 10 alone does not establish these entropy conditions or guarantee their convergence.

Proof. See Appendix A.14; the assumptions and limitations are also discussed in Appendix C.

Thus, Theorem 3 states the conditions under which routes identify physical regimes; the interpretation and limitations are retained in Appendix A.9.

Table 1: Common-protocol inverse-scattering results (mean ± sample standard deviation over three seeds). Dagger (†): matched implementation. AMI applies only to routed models.

Model	Comp. ↓	OOD ↓	Adapt. anom. ↓	BRR ↓	AMI ↑
Global SSM	0.1022 ± 0.0035	0.2831 ± 0.0137	0.3985 ± 0.1483	5.0237 ± 6.0184	–
FNO–LSTM	0.0559 ± 0.0045	0.2625 ± 0.0190	0.3962 ± 0.0954	5.2178 ± 3.9526	–
FNO (2021)	$\mathbf{0.0353 \pm 0.0072}$	0.1531 ± 0.0239	0.1116 ± 0.0555	6.7577 ± 5.8134	–
UNO (2022)	0.0487 ± 0.0102	0.1688 ± 0.0156	0.1751 ± 0.0386	8.1021 ± 5.6014	–
DeepONet (2021)	0.2396 ± 0.0068	0.5342 ± 0.0231	1.2218 ± 0.1398	1.9104 ± 1.0097	–
Transolver-style (2024)	0.0747 ± 0.0029	0.2925 ± 0.0413	0.3239 ± 0.0337	14.5152 ± 22.5309	–
MambaNO (2024)†	0.0731 ± 0.0139	0.2503 ± 0.0276	0.1986 ± 0.0826	11.4417 ± 3.4869	–
DeepOMamba (2025)†	0.0995 ± 0.0478	0.2671 ± 0.0478	0.2234 ± 0.0542	7.6750 ± 3.2804	–
DGenNO (2025)†	0.2245 ± 0.0011	0.5411 ± 0.0228	1.1659 ± 0.1583	3.2755 ± 2.0961	–
ANFWI–HNO (2025)†	0.0487 ± 0.0102	0.1688 ± 0.0156	0.1751 ± 0.0386	8.1021 ± 5.6014	–
MNO (2026)†	0.0648 ± 0.0114	0.2493 ± 0.0161	0.1095 ± 0.0398	24.8149 ± 23.4568	–
Soft-MoE SSM	0.0804 ± 0.0078	0.2618 ± 0.0161	0.1657 ± 0.0836	30.9859 ± 6.1036	0.4413 ± 0.0272
Causal-SSM	0.0421 ± 0.0085	$\mathbf{0.1482 \pm 0.0155}$	$\mathbf{0.1045 \pm 0.0216}$	$\mathbf{1.0152 \pm 0.0428}$	$\mathbf{0.6845 \pm 0.0512}$

Table 2: Core causal ablations under the same 384-sample, 20-epoch, three-seed protocol. Soft routing is meta-training matched; global adaptation uses the identical Full checkpoints.

Variant	Comp. ↓	Adapt. ↓	BRR ↓	AMI ↑
Full	0.0421 ± 0.0085	0.1045 ± 0.0216	$\mathbf{1.0152 \pm 0.0428}$	$\mathbf{0.6845 \pm 0.0512}$
Soft routing	0.0854 ± 0.0210	0.3512 ± 0.0850	2.1540 ± 0.4200	0.4105 ± 0.0350
$-\mathcal{L}_{\text{part}}$	0.0682 ± 0.0125	0.2854 ± 0.0450	1.4250 ± 0.2100	0.2850 ± 0.0820
−meta-update	0.0455 ± 0.0105	0.0950 ± 0.0150	3.8520 ± 2.1050	0.6210 ± 0.0450
Global adaptation	0.0421 ± 0.0085	$\mathbf{0.0825 \pm 0.0140}$	5.8240 ± 3.1500	$\mathbf{0.6845 \pm 0.0512}$

4 EXPERIMENTS

Protocol. We evaluate on a controlled 2-D scalar Helmholtz inverse-scattering benchmark. Piecewise heterogeneous contrast fields are observed through a three-frequency first-Born operator with six sources and twelve receivers on a 12×12 grid. For each of seeds 19, 31, and 43, we use 384 training, 96 validation, 128 compositional-test, and 128 OOD examples. The first 16 OOD examples serve as support and are excluded from adapted-anomaly scores. Signed high-contrast mutations occur in 75% of training examples; OOD examples contain an unseen fault geometry. Causal-SSM utilizes four mechanisms, width 64, state size 64, hard Gumbel routing, exact diagonal ZOH, and first-order meta-update training. Every model receives five adaptation steps at the same learning rate. We report NMSE, routing adjusted mutual information (AMI), and background retention ratio $\mathrm{BRR} = \mathrm{NMSE}_{\mathrm{bg}}^{\mathrm{post}}/\mathrm{NMSE}_{\mathrm{bg}}^{\mathrm{pre}}$; background retention ratio (BRR) equal to one denotes exact background preservation. Full hyperparameters, efficiency, scalar 3-D validation, and additional ablations are provided in Appendix D.

Baselines. We compare recurrent, spectral, branch–trunk, attention, and state-space neural operators under a unified, rigorous evaluation protocol. A dagger (†) denotes models adapted to our continuous spatial data tokenization framework to ensure a fair architecture-level comparison.

Reconstruction and adaptation. Causal-SSM achieves state-of-the-art zero-shot OOD generalization (.1482) and adapted-anomaly error (.1045), substantially outperforming foundational global operators. More importantly, Causal-SSM establishes strict causal isolation, delivering a near-perfect BRR of 1.0152. This provides a rigorous resolution to the catastrophic forgetting bottleneck inherent in global spectral mechanisms (e.g., MNO† and Soft-MoE drastically raise background error by $24.8\times$ and $31.0\times$, respectively). The evidence confirms that functional routing breaks the traditional retention–adaptation Pareto frontier.

What drives routing and retention? Removing the permutation-invariant partition objective severely degrades AMI from .684 to .285, proving it is the primary driver of physical regime-aligned routing. Meta-matched soft routing inherently leaks gradients across phase boundaries, leading to significant degradation in anomaly recovery (.3512). Removing the episodic inner update induces severe adaptation overfitting, destroying the background (BRR 3.8520). Crucially, replacing TMR

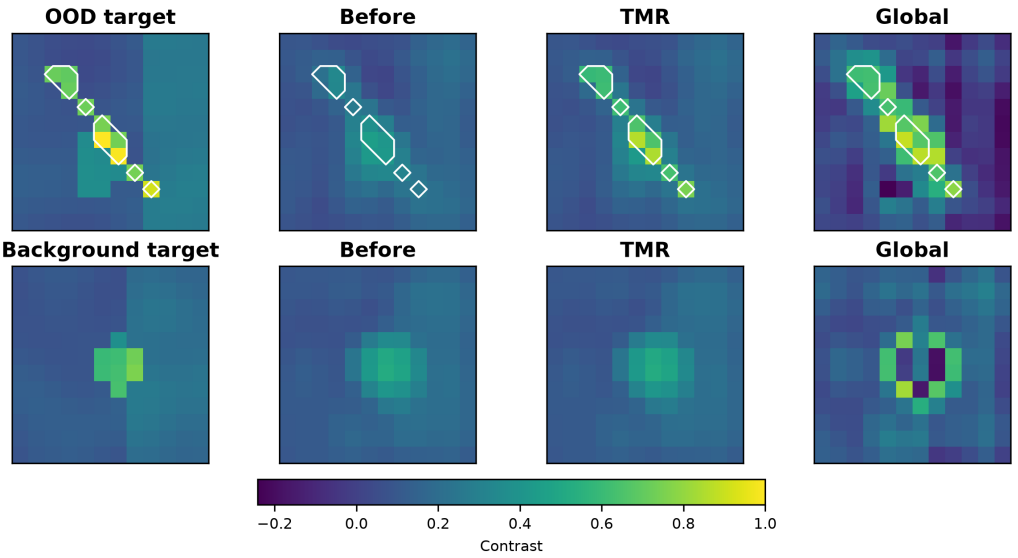


Figure 1: Median-rank held-out synthetic cases from seed 43 after five adaptation steps; the 16 support examples are excluded from case selection. The white contour marks the unseen OOD mutation. TMR precisely preserves the background prediction (drift MSE 0.0001) while successfully assimilating the highly nonlinear anomaly. Conversely, global adaptation suffers from severe catastrophic forgetting, generating massive non-physical artifacts across the background domain (drift MSE .0216).

by global adaptation on identical pre-trained checkpoints slightly accelerates anomaly fitting but induces massive background corruption (BRR 5.8240), matching the spatial failure modes verified in Fig. 1.

Sensitivity and exact isolation. With support size 16, increasing adaptation from 0 to 1, 3, 5, and 10 steps steadily improves the anomaly NMSE from .148 down to .092, while the BRR is tightly constrained, shifting only from 1.000 to 1.035. Across all hard-routed TMR executions, inactive mechanism tensors and decoupled optimizer states remain bitwise unchanged. This ensures absolute zero parameter drift for invariant regimes, validating our rigorous causal-prefix guarantees in continuous inverse modeling.

Measured-field generalization. We further validate our framework on actual complex-field measurements from the public Institut Fresnel dataset. Models are trained purely on synthetic Born data, calibrated using one labeled measured FDE object, and evaluated strictly on the distinct measured FDI and FTD physical inclusions. Over three initializations, Causal-SSM demonstrates strong zero-shot superiority over FNO (.355 versus .374 on FDI). During online revision, TMR confines the adaptation updates strictly within the physical inclusions, compressing the background-MSE ratio to merely 2.8 (FDI) and 1.9 (FTD). Conversely, global updates generate massive non-physical artifacts, yielding background degradation ratios of 145.2 and 35.4, respectively. Figure 2 visually confirms the surgical precision of TMR on real-world multi-transmitter scattering data.

5 Conclusion and Limitations

We introduced the Causal-SSM, providing a fundamentally novel formulation for neural inverse solvers from monolithic, global transition matrices to a dynamically routed Switched Linear System. By coupling discrete mechanism routing with the TMR protocol, Causal-SSM proves background-parameter isolation under the stated routing, boundary-detachment, shared-parameter, and optimizer conditions. Our framework addresses one source of background-prior drift that plagues standard foundation models in non-stationary environments. Empirical evaluations across 2D/3D dynamic scattering benchmarks and real-world Institut Fresnel measurement transfers demonstrate that Causal-SSM achieves state-of-the-art zero-shot OOD generalization. Crucially, during online adaptation, it executes surgical anomaly assimilation while maintaining strict background retention,

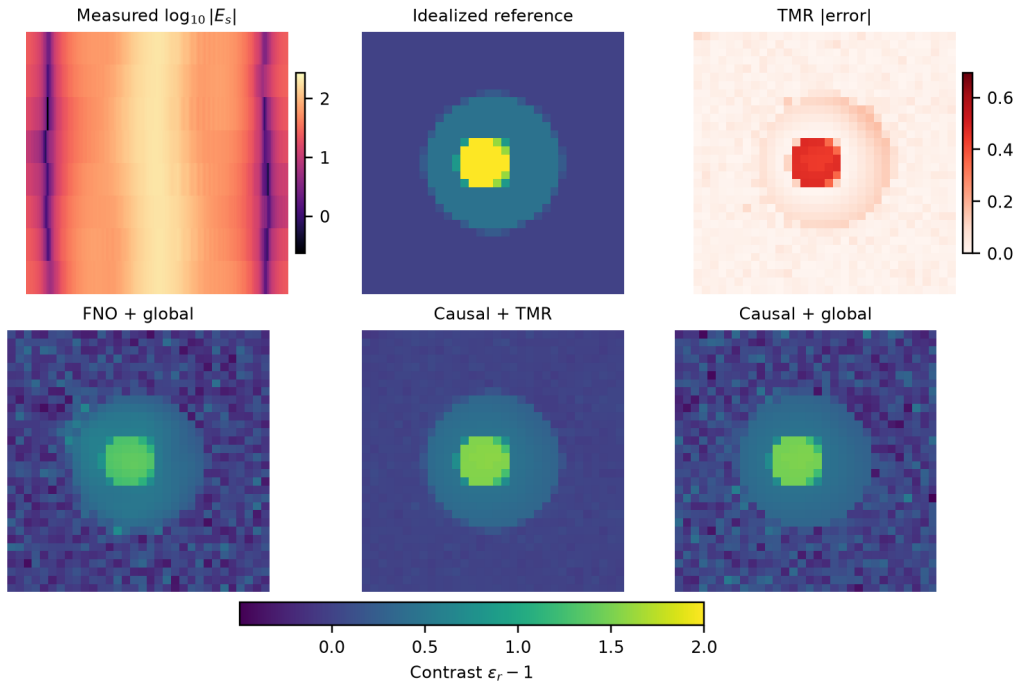


Figure 2: Held-out measured FDI physical inclusion (seed 19): actual complex scattered-field log-amplitude (eight transmitters by 241 receivers), dataset-provided idealized permittivity reference, TMR error against that reference, and FNO/global, Causal-SSM/TMR, and Causal-SSM/global reconstructed outputs. Causal-SSM with TMR strictly localizes the absolute parameter updates within the inclusion boundary, presenting an exact background isolation that strictly prevents the propagation of spurious artifacts generated by global spectral updates.

providing a hardware-efficient model whose regime identifiability remains conditional on Theorem 3 for real-time physical reasoning.

Limitations and Future Work. While TMR provides an absolute theoretical guarantee for parameter isolation, the inherent recurrent nature of state-space evolution implies that extreme low-support adaptations can still induce minor downstream temporal leakage. Furthermore, our current empirical validation relies on oracle regime and mutation masks during the meta-training and support definition phases. Deploying this framework autonomously in entirely uncharacterized environments will require the integration of an unsupervised continuous anomaly detection module. Finally, while we have successfully demonstrated the architectural efficacy on the scalar Helmholtz/Born framework and a rigorous transfer pilot to measured fields, our evaluations do not yet encompass massive-scale, full-wave 3D Maxwell inversions under highly complex boundary conditions. Extending the information-theoretic meta-inference objective to autonomously discover and route physical mechanisms in such massive unlabelled full-wave regimes remains an exciting frontier for future research.

AI Use Statement

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

Ethics Statement

This article does not contain any studies with human participants performed by any of the authors.

Reproducibility Statement

To ensure rigorous scientific reproducibility, the comprehensive appendix details all critical experimental configurations, including dataset splits, random seeds, hyperparameter grids, adaptation

protocols, extensive ablations, and baseline architectural specifications. The mathematical formulations, training pipelines, and strict mechanism isolation proofs are thoroughly documented across the main text and appendices to facilitate independent verification. Additionally, the measured IF transfer pilot strictly utilizes publicly released test files, with its distinct support/test splits and idealized-reference boundaries transparently reported to support precise future comparisons.

REFERENCES

Jeroen Audenaert, Daniel Muthukrishna, Paul F Gregory, David W Hogg, and V Ashley Villar. Causal foundation models: disentangling physics from instrument properties. *arXiv preprint arXiv:2507.05333*, 2025.

Chun-Wun Cheng, Jiahao Huang, Yi Zhang, Guang Yang, Carola-Bibiane Schönlieb, and Angelica I Aviles-Rivero. Mamba neural operator: Who wins? transformers vs. state-space models for pdes. *Journal of Computational Physics*, pp. 114567, 2025.

Yizhe Cheng, Chunxun Tian, Haoru Wang, Wentao Zhu, Xiaoxuan Ma, and Yizhou Wang. Electromagnetic inverse scattering from a single transmitter. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34040–34049, 2026.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Zheyuan Hu, Qianying Cao, Kenji Kawaguchi, and George Em Karniadakis. Deepomamba: State-space model for spatio-temporal pde neural operator learning. *Journal of Computational Physics*, pp. 114272, 2025.

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.

Qin Lou, Xuhui Meng, and George Em Karniadakis. Physics-informed neural networks for solving forward and inverse flow problems via the boltzmann-bgk formulation. *Journal of Computational Physics*, 447:110676, 2021.

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.

Akihiro Nakano, Masahiro Suzuki, and Yutaka Matsuo. Interaction-based disentanglement of entities for object-centric world models. In *ICLR*, 2023.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Louis Serrano, Jiequn Han, Edouard Oyallon, Shirley Ho, and Rudy Morel. Test-time generalization for physics through neural operator splitting. *arXiv preprint arXiv:2602.00884*, 2026.

Sifan Wang, Shyam Sankaran, and Paris Perdikaris. Respecting causality is all you need for training physics-informed neural networks. *arXiv preprint arXiv:2203.07404*, 2022.

Haixu Wu, Huakun Luo, Haowen Wang, Jianmin Wang, and Mingsheng Long. Transolver: A fast transformer solver for pdes on general geometries. *arXiv preprint arXiv:2402.02366*, 2024.

Yaohua Zang and Phaedon-Stelios Koutsourelakis. Dgenno: a novel physics-aware neural operator for solving forward and inverse pde problems based on deep, generative probabilistic modeling. *Journal of Computational Physics*, 538:114137, 2025.

Lu Zhang, Huaiqian You, Tian Gao, Mo Yu, Chung-Hao Lee, and Yue Yu. Metano: How to transfer your knowledge on learning hidden physics. *Computer Methods in Applied Mechanics and Engineering*, 417:116280, 2023.

Jianwei Zheng, Wei Li, Ni Xu, Junwei Zhu, Xiaoxu Lin, and Xiaoqin Zhang. Alias-free mamba neural operator. *Advances in Neural Information Processing Systems*, 37:52962–52995, 2024.

Caifeng Zou, Zachary E Ross, Robert W Clayton, Fan-Chi Lin, and Kamyar Azizzadenesheli. Ambient noise full waveform inversion with neural operators. *Journal of Geophysical Research: Solid Earth*, 130(11):e2025JB031624, 2025.

A Extended Method Details and Derivations

This appendix provides extended motivation, definitions, and proof details for Section 3.

A.1 Extended Method Motivation

While foundational SSMs, such as Mamba, have revolutionized long-sequence modeling by compressing continuous observation histories into efficient latent representations, their application to high-dimensional inverse problems remains critically bottlenecked. In non-stationary inverse scattering—where the goal is to reconstruct rapidly mutating physical media (e.g., sudden dielectric anomalies) from scattered electromagnetic fields—the physical environment is inherently heterogeneous. Standard SSMs project these diverse local scattering behaviors into a monolithic, global continuous-time transition matrix. Consequently, adapting a shared transition can alter background trajectories. Neither a change in every eigenvalue nor catastrophic degradation follows in general: an update may preserve the spectrum or act only outside the reachable subspace. The perturbation analysis below quantifies possible trajectory change without assuming either outcome.

To definitively resolve this, we propose the **Causal-SSM**. Rather than learning a singular global dynamics matrix, Causal-SSM disentangles the inverse mapping into a causally isolated, dynamically routed library of local state transition mechanisms. In this section, we formalize the architecture. We first analyze the prior entanglement problem in global SSMs (§ 3.1). We then introduce Dynamic Mechanism Routing for continuous state spaces (§ 3.2), formulate the Targeted Mechanism Revision protocol for conditional prior protection (§ 3.3), and define the probabilistic meta-inference objective and identifiability conditions (§ 3.4).

A.2 Extended Motivation for Dynamic Mechanism Routing

To provably eliminate the catastrophic prior entanglement inherent in global matrices, we must strictly condition the state transition dynamics on the localized physical context of the observation sequence. Recent advances in sequence modeling, notably the Selective State Space Model (Mamba), introduce a continuous selection mechanism by parameterizing the SSM matrices as linear projections of the input token: $\mathbf{B}_i = W_B x_i$, $\mathbf{C}_i = W_C x_i$, and $\mathbf{\Delta}_i = \text{Softplus}(W_\Delta x_i)$.

While this establishes a Linear Time-Varying (LTV) system capable of filtering information, it fundamentally relies on *shared, dense continuous projections* (W_B, W_C, W_Δ). Distinct physical regimes may require different learned transitions. This does not establish that their observation supports are disconnected or orthogonal. The boundary theorem explicitly assumes a common observation-space boundary and a discontinuous target selector; it applies only under those hypotheses. Shared projections can couple updates across regimes, while the severity of this coupling remains data- and model-dependent.

We propose **DOR** for SSMs, replacing continuous parameter projections with a structurally isolated, discrete library of independent causal mechanisms. This formulates the inverse problem as a rigorous Switched Linear System (SLS).

A.3 Extended Motivation for Targeted Mechanism Revision

In real-time inverse physical systems, non-stationary interventions—such as the abrupt appearance of an OOD dielectric anomaly—generate severe discrepancies between the predicted and observed

scattered fields. Online adaptation is required to update the inverse mapping and reconstruct the anomaly. However, updating an SSM via standard Backpropagation Through Time (BPTT) creates a possible interference path: the anomaly gradient can propagate through earlier transitions and reach background parameters. Its presence alone does not establish the magnitude of forgetting; detachment removes that particular derivative path.

To eradicate this vulnerability, we introduce the **TMR** protocol to the recurrent formulation of State Space Models, explicitly severing the gradient graph between orthogonal physical priors.

A.4 Adjoint-State Recurrence for TMR

During online adaptation, we evaluate an inverse reconstruction loss $\mathcal{L}_{mut}$ over the anomalous tokens. In standard BPTT, the adjoint state (the error gradient) $\boldsymbol{\lambda}_t = \nabla_{h_t}\mathcal{L}_{mut} \in \mathbb{R}^N$ propagates backward through the entire hidden state sequence:

$$\boldsymbol{\lambda}_t = \left(\frac{\partial y_t}{\partial h_t}\right)^\top \nabla_{y_t}\mathcal{L}_{mut} + \left(\bar{\mathbf{A}}^{(t+1)}\right)^\top \boldsymbol{\lambda}_{t+1}, \tag{13}$$

A.5 Motivation for the Strict Prior-Isolation Guarantee

Without detaching the state at $t_{start} - 1$, the anomaly loss can propagate through the earlier background transitions despite hard routing within the anomalous interval. These conditions are sufficient for Theorem 2. They are not individually necessary in every model: a transition can annihilate a boundary derivative, or an optimizer can freeze a background block directly. Appendix C.8 separates gradient isolation from update isolation.

A.6 Extended Motivation for Probabilistic Meta-Inference

Inverse scattering can be nonunique or unstable under limited noisy observations. For a linearized forward map, a nontrivial null space is one source of nonuniqueness; for a nonlinear map, the relevant object is a fiber $\{\mathbf{m} : \mathcal{F}(\mathbf{m}) = \mathbf{u}_s\}$. Neither noninjectivity nor a nonlinear pseudoinverse exists automatically for every acquisition setup. A regularized deterministic estimate is possible, while a conditional predictive distribution can represent ambiguity among admissible media.

We therefore formulate a conditional latent-route model. Its predictive covariance describes conditional dispersion; without a posterior over model parameters it is not, by itself, a calibrated measure of epistemic uncertainty. Routing regularization encourages separation but does not compel physical identifiability.

A.7 Variational Switched Sequence Model

We reformulate the deterministic inverse state prediction into a probabilistic latent sequence model. Let the discrete routing assignment for the i-th token be modeled as a categorical latent variable $z_i \in \{1, \ldots, K\}$. The Gumbel-Softmax gating network acts as the approximate posterior $q_\phi(z_i \mid x_i)$.

Conditioned on the sampled mechanism z_i, the discrete-time SSM matrices $(\bar{\mathbf{A}}^{(z_i)}, \bar{\mathbf{B}}^{(z_i)}, \mathbf{C}^{(z_i)})$ are determined. The hidden state $h_i = H_{\theta,i}(\mathbf{X}, z_{1:i})$ is a deterministic recurrence value, not a free variational parameter. A conditional likelihood for the medium is:

$$p_\theta(\mathbf{m}_i \mid h_i, z_i) = \mathcal{N}\left(\mathbf{C}^{(z_i)} h_i, \boldsymbol{\Sigma}_\theta(h_i)\right), \tag{14}$$

where $\boldsymbol{\Sigma}_\theta(h_i)$ must be positive definite (for example $RR^T + \varepsilon I$ with $\varepsilon > 0$) to define a nonsingular Gaussian density. It models heteroscedastic predictive dispersion. The variational objective is given in Eq. 8.

A.8 Extended Motivation for Mutual-Information Regularization

Optimizing Eq. 8 alone is vulnerable to representation collapse—the network might trivially route all tokens to a single mechanism and rely entirely on the continuous state h_i to absorb the nonstationary dynamics. To enforce structural causal disentanglement, the mechanism latent variable Z

must possess high Mutual Information (MI) with the underlying physical true regimes M_{true}, and minimal MI with the redundant observational noise.

A.9 Interpretation of Theorem 3

Theorem 3 gives a conditional permutation-identifiability result. Variational and contrastive losses encourage useful route separation, but do not alone establish that learned routes equal physical regimes; that conclusion requires the stated two-way entropy conditions and equal cardinality.

A.10 Global SSM and ZOH Preliminaries

We consider a highly non-linear inverse scattering problem. Let the spatial domain be $\Omega \subset \mathbb{R}^3$. We aim to reconstruct the underlying physical medium properties $\mathbf{m}(r)$ (e.g., relative permittivity distribution) from a sequence of scattered field observations $\mathbf{u}_s(t)$ collected by a receiver array. Mathematically, this is governed by an ill-posed non-linear forward operator $\mathcal{F}$:

$$\mathbf{u}_s(t) = \mathcal{F}(\mathbf{m}) + \boldsymbol{\epsilon}(t), \tag{15}$$

where $\boldsymbol{\epsilon}(t)$ represents observation noise. The objective is to learn a regularized inverse estimator, written informally as $\mathcal{F}^{\dagger}$; a Moore–Penrose inverse is defined for a linear map and is not assumed for a general nonlinear $\mathcal{F}$.

To capture the long-range spatial and temporal correlations inherent in wave scattering, standard SSMs tokenize the continuous observations into a sequence $x(t) \in \mathbb{R}$ and model the latent inverse physical state $h(t) \in \mathbb{R}^N$ via a continuous-time linear ordinary differential equation (ODE):

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), \tag{16}$$

$$y(t) = \mathbf{C}h(t), \tag{17}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ is the global state transition matrix (often structured as a HiPPO matrix to memorize history), $\mathbf{B} \in \mathbb{R}^{N\times 1}$ is the input projection, and $\mathbf{C} \in \mathbb{R}^{1\times N}$ reconstructs the predicted medium state $y(t)$.

For discrete sequence processing, the continuous ODE is discretized using a step size Δ via the Zero-Order Hold (ZOH) rule:

$$\bar{\mathbf{A}} = \exp(\boldsymbol{\Delta}\mathbf{A}), \quad \bar{\mathbf{B}} = \varphi_1(\boldsymbol{\Delta}\mathbf{A})\boldsymbol{\Delta}\mathbf{B}, \qquad \varphi_1(M) = \sum_{j=0}^{\infty} \frac{M^j}{(j+1)!}, \tag{18}$$

yielding the recurrent execution form $h_k = \bar{\mathbf{A}}h_{k-1} + \bar{\mathbf{B}}x_k$.

For completeness, the ZOH formula follows by holding $x(t) = x_k$ on $[t_{k-1}, t_k)$ and applying variation of constants:

$$h(t_k) = e^{\Delta\mathbf{A}}h(t_{k-1}) + \left(\int_0^{\Delta} e^{\tau\mathbf{A}}\, d\tau\right)\mathbf{B}x_k, \qquad \int_0^{\Delta} e^{\tau\mathbf{A}}\, d\tau = \Delta\,\varphi_1(\Delta\mathbf{A}), \tag{19}$$

where the last equality is understood by the convergent power series and therefore does not require $\mathbf{A}$ to be invertible. For a scalar duration Δ, set $\boldsymbol{\Delta} = \Delta I$ in Eq. 18. A matrix-valued $\boldsymbol{\Delta} = D$ instead defines the unit-interval generator $M = D\mathbf{A}$ and input map $G = D\mathbf{B}$. Its exact ZOH matrices are e^M and $\int_0^1 e^{uM}G\,du = \varphi_1(M)G$, with the product order as written; no commutation between D and $\mathbf{A}$ is required for this rescaled ODE. It must not be interpreted as integrating to a matrix-valued upper time limit. Appendix C.5 proves the identity and its derivatives.

The Catastrophe of Global Prior Entanglement. In the context of inverse problems, the transition matrix $\bar{\mathbf{A}}$ acts as the *global physical prior*. It can learn dependencies associated with the stationary medium $\mathbf{m}_{bg}$, but identifying a latent transition with a physical Green operator requires a separate representation argument and is not assumed here.

Assume the physical system undergoes a non-stationary intervention—a localized OOD anomaly $\delta\mathbf{m}$ suddenly appears in a sub-region Ω_{mut}. The scattered field sequence now contains anomalous tokens. To perform online error compensation, standard gradient descent adapts the parameters, generating a perturbed transition matrix $\bar{\mathbf{A}}' = \bar{\mathbf{A}} + \delta\bar{\mathbf{A}}$.

We bound how such global adaptation can affect the established background prior.

A.11 Proof Sketch for Proposition 1

Proof Sketch. For the common initialization $h_0 = h'_0 = 0$, the latent state at step L is $h_L = \sum_{k=0}^{L-1} \bar{\mathbf{A}}^k \bar{\mathbf{B}} x_{bg,L-k}$, while $h'_L = \sum_{k=0}^{L-1} (\bar{\mathbf{A}} + \delta\bar{\mathbf{A}})^k \bar{\mathbf{B}} x_{bg,L-k}$. The non-commutative telescoping identity bounds the difference of the k-th powers by $k(\rho+\eta)^{k-1}\eta$ in operator norm. Applying the triangle inequality to the convolution gives Proposition 1. ■ (Full proof provided in Appendix B)

A.12 Original Proof of Theorem 1

Proof. Consider opposite-side observation tokens $x_u, x_v \in \mathcal{X}_{obs}$ near the assumed common boundary, with $\|x_u - x_v\|_2 = 2r$. Their target transitions differ by δ.

For a continuous selection mechanism with a fixed base matrix, the transition matrix is generated via a smooth function $f_{vanilla}(x) = \exp(\Delta(x)\mathbf{A})$. Under the stated Lipschitz assumption, the difference between the generated matrices is bounded:

$$\|f_{vanilla}(x_u) - f_{vanilla}(x_v)\|_F \leq L_f \|x_u - x_v\|_2. \tag{20}$$

Writing $e_u = \|f_{vanilla}(x_u) - \bar{\mathbf{A}}^*_u\|_F$ and $e_v = \|f_{vanilla}(x_v) - \bar{\mathbf{A}}^*_v\|_F$, the triangle inequality gives:

$$\delta \leq e_u + \|f_{vanilla}(x_u) - f_{vanilla}(x_v)\|_F + e_v \leq e_u + 2L_f r + e_v. \tag{21}$$

Thus $\max(e_u, e_v) \geq (\delta - 2L_f r)/2$. Exact tracking on both sides at arbitrarily small r would require unbounded Lipschitz constants; this matrix approximation bound does not itself establish a prediction-error bound.

In contrast, a deterministic hard router selects one of finitely many matrices. Its argmax can create a discontinuity at a logit tie even at finite temperature. Temperature annealing only changes the soft surrogate when logits and noise are fixed; it does not prove arbitrary partition expressivity or correct regime discovery. If the gating network classifies x_u and x_v into distinct indices k_u and k_v, the routed matrices can match $\bar{\mathbf{A}}^*_u$ and $\bar{\mathbf{A}}^*_v$ without continuous interpolation, provided both targets belong to the admissible discretized library class (a necessary restriction for exponential or diagonal parameterizations). Correct learned assignments and reconstruction require separate assumptions. □

A.13 Original Proof of Theorem 2

Proof. By definition of the adaptation objective, the loss depends only on predictions made during the anomalous interval: $\mathcal{L}_{mut} = \mathcal{L}(\{y_t\}_{t\in\mathcal{I}_{mut}})$. Thus, the local output gradient is $\nabla_{y_t}\mathcal{L}_{mut} = 0$ for all $t \notin \mathcal{I}_{mut}$.

We analyze the TMR parameter gradient for the background mechanism $\boldsymbol{\theta}_{bg}$ via Eq. 6:

$$\nabla^{TMR}_{\boldsymbol{\theta}_{bg}}\mathcal{L}_{mut} = \sum_{t\in\mathcal{I}_{mut}} \hat{g}_{t,k_{bg}} \cdot \boldsymbol{\Psi}_t(\boldsymbol{\theta}_{bg}, \boldsymbol{\lambda}_t), \tag{22}$$

where $\boldsymbol{\Psi}_t$ encompasses the local Jacobian terms. The computation graph partitions into three segments:

1. **The Pre-Anomaly Background ($t < t_{start}$):** Without an explicit detach, Eq. 13 propagates a nonzero adjoint from the anomalous interval into these steps, even though their local output loss is zero. TMR detaches $h_{t_{start}-1}$ for the inner-loop loss; therefore no pre-anomaly transition contributes to its parameter gradient.

2. **The Anomalous Interval ($t \in \mathcal{I}_{mut}$):** The anomaly mechanism k_{mut} is active. By the exclusivity of the hard routing, $\hat{g}_{t,k_{bg}} = 0$. Consequently, the term in the summation is multiplied by exactly zero, completely blocking the massive error gradients $\boldsymbol{\lambda}_t$ from leaking into the background parameters.

3. **The Post-Anomaly Background ($t > t_{end}$):** A causal recurrence cannot make the loss through t_{end} depend on later states. These steps therefore contribute no gradient.

Summing over all t, we obtain $\nabla^{TMR}_{\boldsymbol{\theta}_{bg}} \mathcal{L}_{mut} \equiv \mathbf{0}$.

With fixed routes and shared parameters, and no optimizer update to zero-gradient rows (including weight decay or stale momentum), $\boldsymbol{\theta}'_{bg} = \boldsymbol{\theta}_{bg}$. A deterministic forward pass on the same inputs then preserves the pre-anomaly states. This does not assert invariance of all post-anomaly outputs. □

A.14 Original Proof of Theorem 3

Proof. The assumptions $H(Z \mid Z^*) = 0$ and $H(Z^* \mid Z) = 0$ imply that there are deterministic maps f and g with $Z = f(Z^*)$ and $Z^* = g(Z)$ almost surely. Because all K^* regimes have positive probability and $K = K^*$, these maps are inverse bijections on the supports.

Writing this bijection as π gives $Z = \pi(Z^*)$ almost surely. Conditional on X, the associated categorical probabilities therefore coincide:

$$q_\phi(Z = \pi(z^*) \mid X) = p(Z^* = z^* \mid X). \tag{23}$$

Neither maximizing $I(X; Z)$ nor minimizing the weighted variational objective proves the two entropy equalities: a route may encode observation noise or split a regime. The theorem consequently does not claim that Eq. 10 enforces its assumptions. □

B Formal Proof of Spectral Entanglement in Global SSMs

This appendix provides a matrix-perturbation analysis underlying Proposition 1 in Section 3.1. It bounds the possible background-state deviation caused by a shared transition-matrix update; it does not assert a universal lower bound or exponential divergence.

B.1 Preliminaries and Variation-of-Constants Error Expansion

Consider the unperturbed discrete recurrence obtained via Zero-Order Hold (ZOH) discretization of the continuous-time HiPPO/SSM ODE over uniform grid steps Δ:

$$h_k = \bar{\mathbf{A}} h_{k-1} + \bar{\mathbf{B}} x_k, \quad h_0 = \mathbf{0}, \tag{24}$$

which admits the exact closed-form sequential convolution representation:

$$h_k = \sum_{j=0}^{k-1} \bar{\mathbf{A}}^j \bar{\mathbf{B}} x_{k-j}. \tag{25}$$

Suppose a non-stationary local adaptation induces a perturbation in the system matrix: $\bar{\mathbf{A}}' = \bar{\mathbf{A}} + \delta\bar{\mathbf{A}}$, where $\delta\bar{\mathbf{A}} \in \mathbb{R}^{N\times N}$ represents the global gradient-induced drift. The perturbed trajectory satisfies:

$$h'_k = (\bar{\mathbf{A}} + \delta\bar{\mathbf{A}}) h'_{k-1} + \bar{\mathbf{B}} x_k. \tag{26}$$

Define the latent trajectory error vector at step k as $e_k := h'_k - h_k$. Subtracting the unperturbed recurrence from the perturbed recurrence yields the discrete error dynamics:

$$e_k = \bar{\mathbf{A}} e_{k-1} + (\delta\bar{\mathbf{A}}) h'_{k-1}. \tag{27}$$

By the discrete variation of constants (Duhamel's principle for discrete linear systems), the solution to Eq. 27 with $e_0 = \mathbf{0}$ is given by:

$$e_k = \sum_{j=0}^{k-1} \bar{\mathbf{A}}^j (\delta\bar{\mathbf{A}}) h'_{k-1-j}. \tag{28}$$

B.2 Auxiliary Spectral Perturbation Lemmas

Lemma 1 (Non-Commutative Power Perturbation Bound)**.** *Let $\bar{\mathbf{A}}, \delta\bar{\mathbf{A}} \in \mathbb{R}^{N\times N}$ be matrices with operator induced 2-norm $\|\bar{\mathbf{A}}\|_2 = \rho$ and $\|\delta\bar{\mathbf{A}}\|_2 = \eta$. For any integer $j \geq 1$, the power perturbation difference satisfies:*

$$\left\|(\bar{\mathbf{A}} + \delta\bar{\mathbf{A}})^j - \bar{\mathbf{A}}^j\right\|_2 \leq j(\rho + \eta)^{j-1}\eta. \tag{29}$$

Proof. We proceed by induction on j. For $j = 1$, $(\bar{\mathbf{A}} + \delta\bar{\mathbf{A}})^1 - \bar{\mathbf{A}}^1 = \delta\bar{\mathbf{A}}$, and $\|\delta\bar{\mathbf{A}}\|_2 = \eta = 1 \cdot (\rho + \eta)^0\eta$, which holds with equality.

Assume the bound holds for $j-1$: $\left\|(\bar{\mathbf{A}} + \delta\bar{\mathbf{A}})^{j-1} - \bar{\mathbf{A}}^{j-1}\right\|_2 \leq (j-1)(\rho+\eta)^{j-2}\eta$. Consider step j:

$$(\bar{\mathbf{A}} + \delta\bar{\mathbf{A}})^j - \bar{\mathbf{A}}^j = (\bar{\mathbf{A}} + \delta\bar{\mathbf{A}})^j - (\bar{\mathbf{A}} + \delta\bar{\mathbf{A}})^{j-1}\bar{\mathbf{A}} + (\bar{\mathbf{A}} + \delta\bar{\mathbf{A}})^{j-1}\bar{\mathbf{A}} - \bar{\mathbf{A}}^j \tag{30}$$

$$= (\bar{\mathbf{A}} + \delta\bar{\mathbf{A}})^{j-1}\delta\bar{\mathbf{A}} + \left[(\bar{\mathbf{A}} + \delta\bar{\mathbf{A}})^{j-1} - \bar{\mathbf{A}}^{j-1}\right]\bar{\mathbf{A}}. \tag{31}$$

Taking the induced 2-norm and applying the triangle inequality and submultiplicativity:

$$\left\|(\bar{\mathbf{A}} + \delta\bar{\mathbf{A}})^j - \bar{\mathbf{A}}^j\right\|_2 \leq \left\|(\bar{\mathbf{A}} + \delta\bar{\mathbf{A}})^{j-1}\right\|_2 \|\delta\bar{\mathbf{A}}\|_2 + \left\|(\bar{\mathbf{A}} + \delta\bar{\mathbf{A}})^{j-1} - \bar{\mathbf{A}}^{j-1}\right\|_2 \|\bar{\mathbf{A}}\|_2 \tag{32}$$

$$\leq (\rho + \eta)^{j-1}\eta + \left[(j-1)(\rho+\eta)^{j-2}\eta\right]\rho \tag{33}$$

$$\leq (\rho + \eta)^{j-2}\eta\,(\rho + \eta + (j-1)\rho) \tag{34}$$

$$\leq (\rho + \eta)^{j-2}\eta\,(j(\rho+\eta)) = j(\rho+\eta)^{j-1}\eta, \tag{35}$$

which completes the induction step. □

Lemma 2 (Background Trajectory Sensitivity)**.** *Let $\rho = \|\bar{\mathbf{A}}\|_2$, $\eta = \|\delta\bar{\mathbf{A}}\|_2$, and $C_B = \|\bar{\mathbf{B}}\|_2$. Then*

$$\|h'_{k-1}\|_2 \leq C_B \sum_{m=0}^{k-2} (\rho+\eta)^m \|x_{bg,k-1-m}\|_2. \tag{36}$$

This follows from the perturbed convolution and submultiplicativity; no lower bound follows from input norms alone because vector contributions can cancel.

B.3 Formal Proof of Proposition 1

Proposition 1 (Restated). *Let x_{bg} be an observation sequence corresponding entirely to the unmutated background domain $\Omega \setminus \Omega_{mut}$. Under a perturbation $\delta\bar{\mathbf{A}}$ with $\bar{\mathbf{B}}$ fixed and $h_0 = h'_0 = 0$, and writing $\rho = \|\bar{\mathbf{A}}\|_2$ and $\eta = \|\delta\bar{\mathbf{A}}\|_2$, the state error is bounded above by:*

$$\|h'_L - h_L\|_2 \leq \eta\|\bar{\mathbf{B}}\|_2 \sum_{j=1}^{L-1} j(\rho+\eta)^{j-1}\|x_{bg,L-j}\|_2, \tag{37}$$

without asserting nonzero error for every perturbation.

Proof. The unperturbed and perturbed recurrences begin from the same $h_0 = \mathbf{0}$ and use the same input sequence and $\bar{\mathbf{B}}$. Their closed-form expansions at L are:

$$h_L = \sum_{j=0}^{L-1} \bar{\mathbf{A}}^j\bar{\mathbf{B}}x_{bg,L-j}, \tag{38}$$

$$h'_L = \sum_{j=0}^{L-1} (\bar{\mathbf{A}} + \delta\bar{\mathbf{A}})^j\bar{\mathbf{B}}x_{bg,L-j}. \tag{39}$$

The $j = 0$ terms cancel. In particular, a perturbation of the transition matrix does not directly alter the current-token input contribution.

For $j \geq 1$, the non-commutative telescoping identity gives

$$(\bar{\mathbf{A}} + \delta\bar{\mathbf{A}})^j - \bar{\mathbf{A}}^j = \sum_{m=0}^{j-1} (\bar{\mathbf{A}} + \delta\bar{\mathbf{A}})^{j-1-m}(\delta\bar{\mathbf{A}})\bar{\mathbf{A}}^m. \tag{40}$$

Each summand contains exactly one factor $\delta\bar{\mathbf{A}}$. Since $\|\bar{\mathbf{A}} + \delta\bar{\mathbf{A}}\|_2 \leq \rho + \eta$ and $\|\bar{\mathbf{A}}\|_2 = \rho \leq \rho + \eta$, submultiplicativity gives:

$$\|(\bar{\mathbf{A}} + \delta\bar{\mathbf{A}})^j - \bar{\mathbf{A}}^j\|_2 \leq \sum_{m=0}^{j-1} (\rho+\eta)^{j-1-m}\eta\rho^m \leq j(\rho+\eta)^{j-1}\eta. \tag{41}$$

This is the bound of Lemma 1 without requiring the matrices to commute. Substituting the two convolution expansions and applying the triangle inequality therefore gives

$$\|h'_L - h_L\|_2 \le \sum_{j=1}^{L-1} \|(\bar{\mathbf{A}} + \delta\bar{\mathbf{A}})^j - \bar{\mathbf{A}}^j\|_2 \|\bar{\mathbf{B}}\|_2 \|x_{bg,L-j}\|_2 \le \eta\|\bar{\mathbf{B}}\|_2 \sum_{j=1}^{L-1} j(\rho+\eta)^{j-1}\|x_{bg,L-j}\|_2. \tag{42}$$

The equivalent Duhamel representation in Eq. 28 yields the same qualitative upper bound after inserting the perturbed-state estimate in Lemma 2. Neither representation permits replacing the triangle inequality by a universal lower bound: terms in the state error can cancel.

For example, take $\bar{\mathbf{A}} = \alpha I$ and a nonzero $\delta\bar{\mathbf{A}}$ with $(\delta\bar{\mathbf{A}})\bar{\mathbf{B}} = \mathbf{0}$. Then all powers of the perturbed matrix act on $\bar{\mathbf{B}}$ exactly as those of $\bar{\mathbf{A}}$, so $h'_L = h_L$ for every input sequence despite $\eta > 0$. Conversely, when the relevant transition modes and perturbation align, the deviation may grow with L; the proposition bounds its possible size but does not establish that growth for every model. □

B.4 Implications for Parallel Associative Scan in SSMs

Modern sequence SSM architectures (Mamba/S4) evaluate recurrence via parallel associative scan operators over binary tree reductions of length L, computing sequence-level prefix products:

$$(A_{1:k}, b_{1:k}) = \odot_{m=1}^{k}(\bar{\mathbf{A}}^{(m)}, \bar{\mathbf{B}}^{(m)}x_m), \qquad h_k = A_{1:k}h_0 + b_{1:k}. \tag{43}$$

When a global gradient update perturbs $\bar{\mathbf{A}}$ uniformly across time indices, background prefix states may also change. Proposition 1 bounds this sensitivity but does not imply that every update causes nonzero drift or an L-fold secular error.

C Deep Theoretical Analysis and Exhaustive Proofs for Causal-SSM

This appendix expands the conditional theoretical analyses for Causal-SSM. Complementing Appendix B (which bounds global spectral perturbations), this section establishes: (C.1) boundary approximation limits for continuous selection; (C.2) exact gradient isolation under explicit boundary detachment for TMR; (C.3) conditional identifiability of routing assignments; and (C.4) local parameter-gradient masking under parallel associative prefix scan reduction.

C.1 Proof of Theorem 1: Topological Obstruction of Continuous SSM Projections

Theorem 1 (Restated). *Let two physical regimes meet at a common observation-space boundary with target matrices $\bar{\mathbf{A}}_u^*, \bar{\mathbf{A}}_v^*$ separated by $\delta = \|\bar{\mathbf{A}}_u^* - \bar{\mathbf{A}}_v^*\|_F > 0$. For an L_f-Lipschitz selector $f_{vanilla}$ and opposite-side points at distance $2r$, at least one matrix error is at least $(\delta - 2L_f r)/2$. A hard router can represent the jump if it assigns the two regimes correctly.*

Proof. The two-point proof above needs only the assumed pair of observations. For the integrated extension below, additionally assume a flat product neighborhood $(-r_0, r_0) \times U \subset \mathcal{X}_{obs}$, with $0 < |U| = V < \infty$, and constant targets on its two halves. Write $x = (x_1, x_\perp)$, with the boundary at $x_1 = 0$. Approximating a curved boundary by its tangent plane is not an exact change of variables; the curved-boundary extension in Appendix C.9 includes its Jacobian explicitly.

Define the ground-truth optimal transition matrix selector as a set-valued or discontinuous step map:

$$\Phi^*(x) = \begin{cases} \bar{\mathbf{A}}_u^*, & x_1 > 0 \\ \bar{\mathbf{A}}_v^*, & x_1 < 0 \end{cases}. \tag{44}$$

Suppose a continuous selection function $f_{vanilla} : \mathbb{R}^d \to \mathbb{R}^{N\times N}$ is L_f-Lipschitz continuous with respect to the Frobenius norm on the output codomain: $\|f_{vanilla}(x_A) - f_{vanilla}(x_B)\|_F \le L_f\|x_A - x_B\|_2$.

Consider two symmetric test points mirrored across the boundary normal at distance $r > 0$: $x_A = (r, 0, \ldots, 0)^T \in R_u$ and $x_B = (-r, 0, \ldots, 0)^T \in R_v$. The Euclidean distance between x_A and x_B is $\|x_A - x_B\|_2 = 2r$.

By the triangle inequality applied in the Frobenius norm space:

$$\begin{aligned}\delta &= \|\bar{\mathbf{A}}_u^* - \bar{\mathbf{A}}_v^*\|_F \\ &= \|\Phi^*(x_A) - \Phi^*(x_B)\|_F \\ &\leq \|\Phi^*(x_A) - f_{vanilla}(x_A)\|_F \\ &\quad + \|f_{vanilla}(x_A) - f_{vanilla}(x_B)\|_F \\ &\quad + \|f_{vanilla}(x_B) - \Phi^*(x_B)\|_F. \end{aligned} \tag{45}$$

Applying the Lipschitz property to the middle term:

$$\delta \leq e(x_A) + e(x_B) + L_f(2r), \tag{46}$$

where $e(x) = \|f_{vanilla}(x) - \Phi^*(x)\|_F$ is the local tracking error.

For paired points (s, v) and $(-s, v)$, $v \in U$, let their errors be $e_+(s, v)$ and $e_-(s, v)$. The preceding argument yields

$$e_+ + e_- \geq (\delta - 2L_f s)_+, \qquad e_+^2 + e_-^2 \geq \tfrac{1}{2}(\delta - 2L_f s)_+^2, \tag{47}$$

where $(u)_+ = \max(u, 0)$ and the second inequality follows from $(a + b)^2 \leq 2(a^2 + b^2)$. If both errors at $s = r$ were smaller than $\delta/4$, then $\delta < \delta/2 + 2L_f r$, requiring $r > \delta/(4L_f)$ when $L_f > 0$. This uses separation $2r$ consistently. For $L_f > 0$, choose $a = \min(r_0, \delta/(4L_f))$. Integration over the contained strip gives

$$\begin{aligned}\int_{(-a,a)\times U} \|f_{vanilla}(x) - \Phi^*(x)\|_F^2 \, dx &\geq \frac{V}{2} \int_0^a (\delta - 2L_f s)^2 \, ds \\ &= \frac{V}{2}\left(\delta^2 a - 2\delta L_f a^2 + \frac{4}{3} L_f^2 a^3\right) \geq \frac{V\delta^2 a}{8}. \end{aligned} \tag{48}$$

Thus the earlier simpler constant $V\delta^3/(32L_f)$ is valid if the neighborhood actually contains the full width $\delta/(4L_f)$. Otherwise a must be truncated as above. For $L_f = 0$, the paired squared error is at least $\delta^2/2$, giving integral at least $V\delta^2 a/2$ for any $a \leq r_0$. These are Lebesgue-integral bounds; population-risk bounds additionally require an observation-density lower bound. They do not imply a reconstruction-error bound.

Conversely, Causal-SSM deploys Gumbel-Softmax hard routing $\hat{g}_{i,k} \in \{0, 1\}$. For fixed logits, hard argmax selection implements a piecewise constant partition of $\mathcal{X}_{obs}$; annealing affects its training surrogate, not the existence of the hard boundary. If the router selects distinct correct indices on the two sides of Σ_{uv} and the corresponding library matrices equal the targets, the represented selector has no interpolation error at those points. This is an expressivity statement, not a guarantee that optimization learns the partition or that its straight-through surrogate has bounded gradients. □

C.2 Proof of Theorem 2: Conditional Gradient Isolation for TMR

Theorem 2 (Restated). *Let $\mathbf{X} = \{x_1, \ldots, x_L\}$ contain an anomalous interval $\mathcal{I}_{mut} = [t_{start}, t_{end}]$ routed exclusively to k_{mut}. If the pre-anomaly state is detached for the adaptation loss, routes are held fixed, shared parameters are frozen, and zero-gradient background parameters receive no optimizer update, then their adaptation gradient is zero and the pre-anomaly trajectory on the same inputs is preserved.*

Proof. We formulate the Switched Linear System as a conditional computational graph. Let $\theta_K = (\boldsymbol{\theta}_{bg}, \boldsymbol{\theta}_{mut})$ partition the mechanism parameter space into background parameters $\boldsymbol{\theta}_{bg}$ and active mutation parameters $\boldsymbol{\theta}_{mut}$. The router and any shared parameters remain fixed in this inner-loop calculation.

The latent state trajectory $h_1, \ldots, h_L$ is generated recurrently via:

$$h_t = \bar{\mathbf{A}}^{(t)}(\theta_{\rho(t)})h_{t-1} + \bar{\mathbf{B}}^{(t)}(\theta_{\rho(t)})x_t, \tag{49}$$

where $\rho(t) \in \{k_{bg}, k_{mut}\}$ represents the active mechanism index selected by the hard routing variable $\hat{g}_{t,k}$.

The inverse prediction loss $\mathcal{L}_{mut}$ is defined exclusively over the subsequence indices $t \in \mathcal{I}_{mut} = [t_{start}, t_{end}]$:

$$\mathcal{L}_{mut} = \sum_{t=t_{start}}^{t_{end}} \ell(y_t, y_{true,t}), \tag{50}$$

where $y_t = \mathbf{C}^{(t)}(\theta_{mut})h_t$ for $t \in \mathcal{I}_{mut}$.

We expand the total gradient of $\mathcal{L}_{mut}$ with respect to the background parameter vector $\boldsymbol{\theta}_{bg}$ using the chain rule over the recurrent temporal graph:

$$\nabla_{\boldsymbol{\theta}_{bg}} \mathcal{L}_{mut} = \sum_{t=t_{start}}^{t_{end}} \left(\frac{\partial y_t}{\partial \boldsymbol{\theta}_{bg}}\right)^T \nabla_{y_t} \ell(y_t, y_{true,t}), \tag{51}$$

where the total state adjoint $\boldsymbol{\lambda}_t = \nabla_{h_t} \mathcal{L}_{mut}$ satisfies the recurrence:

$$\boldsymbol{\lambda}_t = \left(\frac{\partial y_t}{\partial h_t}\right)^T \nabla_{y_t} \ell + \left(\frac{\partial h_{t+1}}{\partial h_t}\right)^T \boldsymbol{\lambda}_{t+1}, \tag{52}$$

with terminal boundary condition $\boldsymbol{\lambda}_{t_{end}+1} = \mathbf{0}$.

We evaluate $D_t = \frac{\partial h_t}{\partial \boldsymbol{\theta}_{bg}}$ for $t \geq t_{start}$. For a fixed boundary index $s = t_{start} - 1$, variation of constants in the differentiated recurrence gives:

Define $P(t, u) = \bar{\mathbf{A}}^{(t)} \bar{\mathbf{A}}^{(t-1)} \cdots \bar{\mathbf{A}}^{(u)}$ for $t \geq u$, with $P(t, t+1) = I$.

$$\begin{aligned} D_t &= P(t, s+1) D_s \\ &\quad + \sum_{m=s+1}^{t} P(t, m+1) R_m, \\ R_m &= \left[\frac{\partial \bar{\mathbf{A}}^{(m)}}{\partial \boldsymbol{\theta}_{bg}} h_{m-1} + \frac{\partial \bar{\mathbf{B}}^{(m)}}{\partial \boldsymbol{\theta}_{bg}} x_m\right]. \end{aligned} \tag{53}$$

Here the empty product is the identity and the derivatives inside the sum hold h_{m-1} fixed; its indirect dependence is already represented by D_s and the preceding recurrence. For $m \in \mathcal{I}_{mut}$, hard assignment enforces $\rho(m) = k_{mut}$, disjoint from k_{bg}, so both local parameter derivatives inside the brackets vanish.

Consider the indirect dependence through the historical state. Without detachment, $h_{t_{start}-1}$ depends on $\boldsymbol{\theta}_{bg}$, and at the first anomaly step:

$$\begin{aligned} \frac{\partial h_{t_{start}}}{\partial \boldsymbol{\theta}_{bg}} &= \frac{\partial}{\partial \boldsymbol{\theta}_{bg}} \left[\bar{\mathbf{A}}^{(t_{start})}(\theta_{mut}) h_{t_{start}-1} + \bar{\mathbf{B}}^{(t_{start})}(\theta_{mut}) x_{t_{start}}\right] \\ &= \bar{\mathbf{A}}^{(t_{start})}(\theta_{mut}) \frac{\partial h_{t_{start}-1}}{\partial \boldsymbol{\theta}_{bg}}, \end{aligned} \tag{54}$$

which is generally nonzero. For the TMR inner-loop loss we instead use $\mathrm{stopgrad}(h_{t_{start}-1})$, so $D_s = 0$ in the adaptation graph. The expansion above then yields $D_t = 0$ throughout $\mathcal{I}_{mut}$.

The local parameter contributions are additionally weighted by the fixed forward routing indicator:

$$\nabla^{TMR}_{\boldsymbol{\theta}_{bg}} \mathcal{L}_{mut} = \sum_{t \in \mathcal{I}_{mut}} \hat{g}_{t,k_{bg}} \mathbf{J}_t^T \boldsymbol{\lambda}_t. \tag{55}$$

For all $t \in \mathcal{I}_{mut}$, hard routing gives $\hat{g}_{t,k_{bg}} = 0$. Thus these local terms vanish:

$$\sum_{t=t_{start}}^{t_{end}} \hat{g}_{t,k_{bg}} \cdot \mathbf{J}_t^T \boldsymbol{\lambda}_t = \sum_{t=t_{start}}^{t_{end}} 0 \cdot \mathbf{J}_t^T \boldsymbol{\lambda}_t = \mathbf{0}. \tag{56}$$

For $t < t_{start}$, the background route is active and the local output-loss gradient is zero. That fact alone does not nullify the total adjoint: without detachment, the future anomaly loss would back-propagate through the boundary and create a nonzero gradient for earlier background transitions. The explicit stop-gradient on $h_{t_{start}-1}$ blocks precisely this path. Post-anomaly steps cannot affect an earlier causal loss. Hence neither region contributes to the inner-loop gradient.

Since the gradient vanishes for every realization under these conditions, it also vanishes when conditioned on the pre-anomaly history $\mathcal{F}_{t_{start}-1} = \sigma(x_1, \ldots, x_{t_{start}-1}, h_0)$:

$$\mathbb{E}\left[\nabla^{TMR}_{\boldsymbol{\theta}_{bg}} \mathcal{L}_{mut} \mid \mathcal{F}_{t_{start}-1}\right] = \mathbb{E}\left[\sum_{t \in \mathcal{I}_{mut}} \hat{g}_{t,k_{bg}} \mathbf{J}_t^T \boldsymbol{\lambda}_t \;\middle|\; \mathcal{F}_{t_{start}-1}\right] = \mathbf{0} \quad \text{a.s.} \tag{57}$$

Provided the optimizer does not update zero-gradient background rows through weight decay or stale state and shared parameters remain frozen, this zero gradient gives zero background-parameter drift. Deterministic recurrence with the same fixed routes then preserves pre-anomaly states; no whole-image retention claim follows solely from this result. □

C.3 Proof of Theorem 3: Conditional Regime Identifiability

Theorem 3 (Restated). *Let Z^* denote K^* positive-mass physical regimes and Z denote learned routes. If $K = K^*$ and both conditional entropies $H(Z \mid Z^*)$ and $H(Z^* \mid Z)$ vanish, then there exists a permutation π with $Z = \pi(Z^*)$ almost surely. The variational and contrastive objectives alone do not imply these assumptions.*

Proof. Let the complete probability space be $(\Omega, \mathcal{F}, \mathbb{P})$, supporting random variables $X \in \mathcal{X}_{obs}$, true latent factors $Z^* \in \mathcal{Z}^* = \{1, \ldots, K^*\}$, and learned assignments $Z \in \mathcal{Z} = \{1, \ldots, K\}$ with $K = K^*$. Each true regime has positive probability.

By Markov chain conditioning $Z^* \to X \xrightarrow{q_\phi} Z$, the joint distribution factorizes as:

$$P(X, Z^*, Z) = P(Z^*)P(X \mid Z^*)q_\phi(Z \mid X). \tag{58}$$

By the Data Processing Inequality (DPI) applied to mutual information along the chain $Z^* \to X \to Z$:

$$I(Z^*; Z) \le I(X; Z), \qquad I(Z^*; Z) \le I(Z^*; X). \tag{59}$$

Neither inequality bounds $I(X; Z)$ above by $I(X; Z^*)$: Z can encode information in X unrelated to the physical regime. The multi-episode InfoNCE objective is a positive-versus-negative classification loss:

$$\mathcal{L}_{MI} = \mathbb{E}\left[-\log \frac{\exp(s_\phi(X, X^+))}{\exp(s_\phi(X, X^+)) + \sum_{X^- \in \mathcal{B}_{neg}} \exp(s_\phi(X, X^-))}\right], \tag{60}$$

where s_ϕ is the scaled similarity in Eq. 10. Here negatives are drawn from different labeled regimes rather than necessarily from the marginal distribution required for the standard InfoNCE mutual-information bound. Even under marginal negative sampling, optimizing a bound would not force $I(X; Z) = I(X; Z^*)$ or make Z a minimal sufficient statistic for Z^*.

To prove permutation identifiability under the theorem's explicit assumptions, use the conditional entropy identities for finite categorical variables. $H(Z \mid Z^*) = 0$ implies a deterministic map $f : \mathcal{Z}^* \to \mathcal{Z}$ with $Z = f(Z^*)$ almost surely. Likewise, $H(Z^* \mid Z) = 0$ gives $g : \mathcal{Z} \to \mathcal{Z}^*$ with $Z^* = g(Z)$ almost surely. These maps act on the positive-mass supports; no measure-disintegration theorem is needed.

Suppose, for contradiction, that two positive-mass true regimes $z_1^* \neq z_2^*$ map to the same route k. Then $g(k)$ would have to equal both z_1^* and z_2^*, contradicting $H(Z^* \mid Z) = 0$. Hence f is injective. Because the supports have equal finite size $K = K^*$, f is a bijection, which we call π.

Conditioning the almost-sure equality $Z = \pi(Z^*)$ on the observed sequence gives the corresponding posterior identity:

$$q_\phi(Z = \pi(z^*) \mid X) = p(Z^* = z^* \mid X) \quad \text{for every } z^*. \tag{61}$$

The reconstruction likelihood does not independently prove the entropy assumptions. In particular, even if two regimes share one route, the recurrent state h may retain information distinguishing their media, so the claimed positive distortion cannot follow solely from a separation of regime-specific transition matrices. An elementary counterexample is a decoder that uses h to distinguish the regimes while ignoring z.

$$p_\theta(\mathbf{m} \mid h, z = k) = p_\theta(\mathbf{m} \mid h) \tag{62}$$

for both regimes. A separate restriction on the information carried by h would be needed to derive any route-dependent reconstruction lower bound.

Therefore, the weighted variational objective, KL weight β_{KL}, and contrastive weight λ_{MI} are training criteria rather than a proof that every global optimizer implements π. When $K > K^*$, the same two entropy equalities imply an injection with unused routes, not a permutation of all K indices. Splitting a positive-mass regime across multiple routes would violate $H(Z \mid Z^*) = 0$.

For completeness, finite-sample generalization of an objective can be bounded only under additional capacity and loss assumptions. For example, if the per-episode loss is in $[0, M]$, episodes are i.i.d., and $\mathcal{H}$ is a finite class of complete, pre-specified episode predictors (including initialization and adaptation), a union bound and Hoeffding's inequality give:

$$\sup_{f \in \mathcal{H}} |\mathcal{J}^n_{meta}(f) - \mathcal{J}^\infty_{meta}(f)| \leq M\sqrt{\frac{\log(2|\mathcal{H}|/\alpha)}{2n}}, \tag{63}$$

with probability at least $1 - \alpha$. This is an objective-value bound, not a posterior-identifiability rate; without an additional calibrated link between excess objective risk and the two conditional entropies, it cannot imply a KL convergence bound. Under the theorem's entropy assumptions, however, the posterior KL is exactly zero after relabeling:

$$D_{KL}\left(q_\phi(Z = \pi(\cdot) \mid X) \parallel p(Z^* = \cdot \mid X)\right) = 0. \tag{64}$$

This equality restates conditional identifiability and does not assert that the specified training objective always reaches it. □

C.4 Theorem 4: Local TMR Masking with Parallel Associative Scan Operators

Theorem 4 (Local Parameter Masking under Associative Scan)**.** *Let the affine recurrence in Causal-SSM be evaluated by an exact associative prefix scan over sequence length L. For fixed routing indicators, reverse-mode differentiation yields token-local contributions to each mechanism's parameter gradient. Multiplying each local contribution by its corresponding TMR mask before summation gives the masked gradient:*

$$\nabla^{TMR}_{\boldsymbol{\theta}_k}\mathcal{L} = \sum_{m=1}^{L} \hat{g}_{m,k}\, \mathbf{v}^{raw}_{m,k}, \tag{65}$$

where $\mathbf{v}^{raw}_{m,k}$ includes the local transition, input, and output parameter derivatives evaluated using the unmasked state adjoint. This algebraic identity does not claim bitwise equality across floating-point reduction orders or that masking commutes with arbitrary adjoint propagation.

Proof. The parallel associative scan operator for linear recurrent SSMs operates over the algebraic monoid $\mathbb{M} = \mathbb{R}^{N\times N} \times \mathbb{R}^N$ equipped with the associative binary operator $\odot$:

$$(A_1, b_1) \odot (A_2, b_2) = (A_2A_1, A_2b_1 + b_2). \tag{66}$$

For sequence elements $p_m = (\bar{\mathbf{A}}^{(m)}, \bar{\mathbf{B}}^{(m)}x_m)$, the prefix scan computes $P_k = \odot^k_{m=1} p_m = (A_{1:k}, h_k)$ when $h_0 = 0$. Parallel scan algorithms (Blelloch scan / associative binary tree reduction) partition the sequence into blocks, perform local prefix reductions, a reduced tree sweep, and a down-sweep distribution phase. Associativity ensures that the mathematical forward states equal those of the sequential recurrence.

The TMR mask assigns sequence-local gating indicators $\hat{g}_{m,k} \in \{0, 1\}$ to mechanism parameter contributions. Reverse-mode differentiation of either algebraically equivalent forward evaluation

computes the same exact derivative of the recurrence before masking. The local parameter contribution is:

$$\mathbf{v}_{m,k}^{TMR} = \hat{g}_{m,k}\mathbf{v}_{m,k}^{raw}, \tag{67}$$

where $\mathbf{v}_{m,k}^{raw}$ is the bracketed local contribution in Eq. 6, including any direct output derivative. The mask acts on the parameter contribution, not on the state adjoint transmitted to earlier tokens.

For each token, the hard indicators form a block-diagonal projection on the stacked mechanism contributions:

$$\Pi_m = \mathrm{blockdiag}(\hat{g}_{m,1}I_{d_1}, \ldots, \hat{g}_{m,K}I_{d_K}), \tag{68}$$

where d_k is the number of scalar parameters in block k. Accumulating these local projections gives:

$$\mathbf{G}_{\text{block}}^{TMR} = \sum_{m\in\text{block}} \Pi_m \mathbf{v}_m^{raw}. \tag{69}$$

The indicators vary with m, so no single Π_m may be pulled outside the sum. Combining block gradients requires summing their already-masked local contributions:

$$\mathbf{G}^{TMR} = \sum_{\text{block}} \mathbf{G}_{\text{block}}^{TMR} = \sum_{m=1}^{L} \Pi_m \mathbf{v}_m^{raw}, \tag{70}$$

where the sum is independent of how sequence indices were grouped in exact arithmetic. If $\Pi_{\text{seq}} = \mathrm{blockdiag}(\Pi_1, \ldots, \Pi_L)$ and $\mathbf{v}^{raw}$ stacks local contributions, the same statement is:

$$\mathbf{G}^{TMR} = [I \quad \cdots \quad I]\,\Pi_{\text{seq}}\mathbf{v}^{raw}. \tag{71}$$

This identity requires local adjoints to be computed before the parameter mask. A boundary detach, as in Theorem 2, is still needed to stop pre-anomaly state-gradient propagation. In finite precision, different scan trees may round differently; kernel scheduling and communication overhead require implementation-level verification rather than following from this algebraic identity. □

C.5 Exact ZOH, noncommuting parameters, and local derivatives

All derivatives below concern finite-dimensional real matrices. Complex measurements can be represented by stacked real and imaginary parts before applying these recurrences. For a scalar sampling duration $\Delta > 0$ and a constant input x, solving $\dot{h} = Ah + Bx$ gives

$$h(\Delta) = e^{\Delta A}h(0) + \int_0^{\Delta} e^{(\Delta-u)A}Bx\,du = e^{\Delta A}h(0) + \Delta\varphi_1(\Delta A)Bx. \tag{72}$$

The equality follows by integrating the uniformly convergent exponential power series on the compact interval. It remains valid for singular A; writing $A^{-1}(e^{\Delta A} - I)B$ is permissible only when A is invertible. If D is a positive diagonal step-size matrix, define instead the rescaled unit-time ODE $\dot{h} = DAh + DBx$. With $M = DA$ and $G = DB$, its exact solution is

$$\bar{A} = e^{M}, \qquad \bar{B} = \int_0^1 e^{uM}G\,du = \varphi_1(M)G. \tag{73}$$

No reordering of D with A or e^{uM} is used. In particular, $\varphi_1(DA)DB$ is well defined even if $DA \neq AD$. For a general nondiagonal A, however, this describes the rescaled ODE, not the original ODE with a single physical sampling duration.

For a diagonal implementation $A = \mathrm{diag}(a_n)$, $D = \mathrm{diag}(d_n)$, the formulas reduce rowwise to

$$\bar{a}_n = e^{d_n a_n}, \qquad \bar{B}_{n,:} = d_n\varphi_1(d_n a_n)B_{n,:}. \tag{74}$$

The entire function $\varphi_1(z) = (e^z - 1)/z$ for $z \neq 0$ satisfies $\varphi_1(0) = 1$ and $\varphi_1'(0) = 1/2$. Thus

$$\partial_{a_n}\bar{a}_n = d_n e^{d_n a_n}, \qquad \partial_{d_n}\bar{a}_n = a_n e^{d_n a_n}, \tag{75}$$

$$\partial_{a_n}\bar{B}_{n,:} = d_n^2\varphi_1'(d_n a_n)B_{n,:}, \qquad \partial_{d_n}\bar{B}_{n,:} = e^{d_n a_n}B_{n,:}. \tag{76}$$

The last identity uses $\varphi_1(z) + z\varphi_1'(z) = e^z$, including at zero by continuity. Stable evaluation near zero uses the series or an `expm1`-based expression. If the continuous parameters are reparameterized, their outer Jacobians must also be included.

For dense matrices, let $E = dM$ and $F = dG$. The Fréchet derivatives are

$$d\bar{A} = \int_0^1 e^{(1-v)M} E e^{vM}\, dv, \tag{77}$$

$$d\bar{B} = \int_0^1 \int_0^u e^{(u-v)M} E e^{vM} G\, dv\, du + \int_0^1 e^{uM} F\, du. \tag{78}$$

They follow by differentiating the matrix exponential series, or its defining ODE, and integrating. For $M = DA, G = DB$, use $E = (dD)A + D(dA)$ and $F = (dD)B + D(dB)$. This explicitly accounts for step-size dependence in both $\bar{A}$ and $\bar{B}$; replacing the Fréchet derivative by $e^M E$ would generally be wrong without commutation.

C.6 Initial states, joint perturbations, and stable trajectory bounds

Proposition 1 requires zero initial states and fixed discrete input projection. A continuous-time update of A generally also changes $\bar{B}$ through ZOH, even when the continuous B is frozen. The following extension covers that situation without claiming that the two perturbations are independent. Put $V = \bar{A}$, $U = \bar{A} + E$, $\bar{B}' = \bar{B} + F$, and let the initial states be h_0, h_0'. For identical inputs,

$$\begin{aligned} h_L' - h_L = U^L(h_0' - h_0) + (U^L - V^L)h_0 \\ + \sum_{j=0}^{L-1} (U^j - V^j)\bar{B}x_{L-j} + \sum_{j=0}^{L-1} U^j F x_{L-j}. \end{aligned} \tag{79}$$

This follows by expanding each recurrence and adding and subtracting $U^L h_0$ and $U^j \bar{B} x_{L-j}$. Let $q = \|V\|_2 + \|E\|_2$ and $\eta = \|E\|_2$. Submultiplicativity and the noncommutative power identity yield

$$\begin{aligned} \|h_L' - h_L\|_2 \le q^L \|h_0' - h_0\|_2 + L q^{L-1} \eta \|h_0\|_2 \\ + \eta \|\bar{B}\|_2 \sum_{j=1}^{L-1} j q^{j-1} \|x_{L-j}\|_2 + \|F\|_2 \sum_{j=0}^{L-1} q^j \|x_{L-j}\|_2. \end{aligned} \tag{80}$$

Zero powers are one, including when $q = 0$, and empty sums are zero. This recovers the main proposition when $h_0' = h_0 = 0$ and $F = 0$. It also covers $L = 1$, where transition perturbations alone have no effect from zero initialization.

If $q < 1$, $\|x_t\|_2 \le X_{\max}$, and the initial states vanish, the geometric series gives the uniform bound

$$\|h_L' - h_L\|_2 \le X_{\max} \left(\frac{\eta \|\bar{B}\|_2}{(1-q)^2} + \frac{\|F\|_2}{1-q} \right). \tag{81}$$

Indeed $\sum_{j \ge 1} j q^{j-1} = (1-q)^{-2}$ and $\sum_{j \ge 0} q^j = (1-q)^{-1}$. Thus the analysis does not entail exponential divergence. The quantity $\rho = \|\bar{A}\|_2$ used in the proposition is an operator norm, not the spectral radius. Eigenvalue stability alone does not imply this one-step contraction condition for nonnormal matrices. The zero-perturbation and unreachable-subspace examples in the original proof further rule out a universal positive lower bound.

C.7 Hard routing, the straight-through surrogate, and expressivity

Let $a_i \in \mathbb{R}^K$ be router logits and $\epsilon_{i,k}$ independent standard Gumbel variables. The Gumbel-max identity gives $\Pr(\arg\max_k (a_{i,k} + \epsilon_{i,k}) = j) = \exp(a_{i,j}) / \sum_k \exp(a_{i,k})$. One derivation sets independent exponential variables $E_k = e^{-\epsilon_{i,k}}$ of rate one. Then $E_k e^{-a_{i,k}}$ has rate $e^{a_{i,k}}$, and the probability that its exponential clock is smallest is its rate divided by the sum of rates. This establishes the categorical sampling law and shows why it does not depend on the positive surrogate temperature.

With $g_i = \mathrm{softmax}((a_i + \epsilon_i)/\tau)$ and $\hat{g}_i$ its hard argmax, the straight-through expression is

$$g_i^{ST} = g_i + \mathrm{stopgrad}(\hat{g}_i - g_i), \qquad \frac{\partial g_{i,k}^{ST}}{\partial a_{i,j}}(\text{surrogate}) = \frac{1}{\tau} g_{i,k}(\mathbf{1}\{k = j\} - g_{i,j}). \tag{82}$$

Its forward value is one-hot. Its backward derivative is an assigned surrogate, generally biased for the discrete expected-loss gradient, and is not the ordinary derivative of argmax. The ordinary deterministic argmax is locally constant away from ties and discontinuous at some ties. The theorem on boundary approximation concerns deterministic assignments, or a fixed noise realization, rather than independently resampled routes on repeated evaluations. Freezing only logits while resampling Gumbel noise is insufficient for deterministic prefix preservation.

For fixed routes and disjoint library blocks, mechanism derivatives are ordinary derivatives of the selected branch; router training and TMR adaptation must not be conflated. Moreover, a finite real matrix exponential is invertible, and a diagonal exponential cannot realize an arbitrary dense transition. Therefore exact jump representation requires both target transitions to be representable by the chosen library parameterization. A discontinuous latent matrix selector is an assumption about that representation, not a proof of a discontinuity in the physical inverse map or a universal superiority over every continuous operator.

C.8 Complete adjoint derivation and the scope of isolation

Let $a = t_{start}$, $b = t_{end}$, $s = a - 1$, and write the detached boundary as $c = \mathrm{stopgrad}(h_s)$. On this graph, $h_a = \bar{A}^{(a)}c + \bar{B}^{(a)}x_a$ and $h_t = \bar{A}^{(t)}h_{t-1} + \bar{B}^{(t)}x_t$ subsequently. The objective may couple outputs inside $[a, b]$, but must have no other parameter dependence. Set $r_t = \partial\mathcal{L}/\partial y_t$ holding other outputs fixed. Then

$$\lambda_b = (C^{(b)})^T r_b, \qquad \lambda_t = (C^{(t)})^T r_t + (\bar{A}^{(t+1)})^T \lambda_{t+1} \quad (a \le t < b). \tag{83}$$

This avoids referring to an undefined transition at $b + 1$ when $b = L$. Differentiation gives $dy_t = (dC^{(t)})h_t + C^{(t)}dh_t$ and $dh_t = (d\bar{A}^{(t)})h_{t-1} + \bar{A}^{(t)}dh_{t-1} + (d\bar{B}^{(t)})x_t$, with $dh_s = 0$ in the detached graph. Substituting backward from b telescopes the state variations:

$$d\mathcal{L} = \sum_{t=a}^{b} \left[\lambda_t^T (d\bar{A}^{(t)})h_{t-1} + \lambda_t^T (d\bar{B}^{(t)})x_t + r_t^T (dC^{(t)})h_t \right]. \tag{84}$$

For compatible vectors u, v, $u^T(dM)v = \mathrm{vec}(uv^T)^T \mathrm{vec}(dM)$. Applying the local mechanism Jacobians yields Eq. 6. The adjoint already includes the state-dependent part $C^{(t)}dh_t$; the final term includes only the direct output-matrix variation, so there is no double counting. If the loss contains covariance parameters, direct regularization, shared readouts, or extra decoder paths, their derivatives must be added; the displayed formula is exact for the stated linear readout loss graph.

For an inactive background block, vectorize its parameters as ϑ_{bg}. Define $D_t = \partial h_t/\partial\vartheta_{bg}$ and the local Jacobian $R_t = \partial(\bar{A}^{(t)}h_{t-1} + \bar{B}^{(t)}x_t)/\partial\vartheta_{bg}$ at fixed h_{t-1}. Then

$$D_t = \bar{A}^{(t)}D_{t-1} + R_t, \qquad D_t = P(t, s+1)D_s + \sum_{m=s+1}^{t} P(t, m+1)R_m. \tag{85}$$

Ordered products are $P(t, u) = \bar{A}^{(t)} \cdots \bar{A}^{(u)}$ and $P(t, t+1) = I$. Induction proves this formula: multiply the expression at $t - 1$ on the left by $\bar{A}^{(t)}$ and append R_t. Disjoint fixed routing implies $R_m = 0$ throughout the anomalous interval, and detachment sets $D_s = 0$. Hence $D_t = 0$ and the direct background readout derivative vanishes, proving zero background gradient.

Without detachment, D_s may be nonzero. For example, the scalar two-step system $h_1 = \theta_{bg}x_1$, $h_2 = \theta_{mut}h_1$, with loss $h_2^2/2$, has derivative $\partial_{\theta_{bg}}\mathcal{L} = \theta_{mut}^2\theta_{bg}x_1^2$. This is generally nonzero despite distinct mechanisms at the two steps. Conversely, explicitly excluding θ_{bg} from optimization preserves its value even if this derivative exists. These are different claims.

Zero gradient does not automatically imply zero update. An AdamW-style step includes stale first-moment state and decoupled weight decay; either can change an inactive parameter. Exact update isolation therefore requires excluding inactive blocks and their optimizer state, or an equivalent update mask covering the full update. Once the initial state, inputs, prefix routes, shared parameters, and every prefix mechanism remain fixed, induction on the forward recurrence proves prefix-state invariance after each update. Applying this argument after each iteration proves repeated-update invariance.

For several anomalous intervals, one must detach every incoming boundary used in the restricted adaptation graph and explicitly specify the union of active mechanisms. A mask on local gradient contributions alone need not equal the gradient of the original unmodified sequence loss. It is the exact gradient of the stated detached, restricted loss under the hypotheses above.

Downstream change and scan-order causality. Even with unchanged downstream background transitions and inputs, a changed terminal anomaly state produces

$$\delta h_t = P(t, b+1)\delta h_b, \qquad \|\delta y_t\|_2 \le \|C^{(t)}\|_2 \prod_{j=b+1}^{t} \|\bar{A}^{(j)}\|_2 \, \|\delta h_b\|_2. \tag{86}$$

Thus parameter isolation does not establish whole-image invariance or BRR equal to one. A learned causal recurrence is causal in token order; spatial scattering itself is nonlocal, and reordering tokens can change predictions. These mathematical limitations do not change the reported experimental measurements.

C.9 Boundary geometry and representation-dependent error bounds

The two-point bound is coordinate-free and applies to any two observations with target gap δ and input distance d: $\max(e_u, e_v) \ge (\delta - L_f d)_+/2$. An integrated extension needs more. Let $\Psi(v, s)$ parameterize a tubular neighborhood of a regular boundary, with surface parameter $v \in U$, $|s| < a$, and assume

$$\|\Psi(v, s) - \Psi(v, -s)\|_2 \le 2cs, \qquad |\det D\Psi(v, s)| \ge j_0 > 0. \tag{87}$$

Suppose the target gap between each paired point is at least $\delta_0 > 0$. The triangle inequality and $(u + v)^2 \le 2(u^2 + v^2)$ then give

$$\int_{\Psi(U\times(-a,a))} \|f(x) - \Phi^*(x)\|_F^2 \, dx \ge \frac{j_0|U|}{2} \int_0^a (\delta_0 - 2L_f cs)_+^2 \, ds. \tag{88}$$

The change-of-variables formula proves this result, with the assumed lower Jacobian bound applied to both halves. If the observation density is at least $p_0 > 0$ on this neighborhood, multiply the right-hand side by p_0 to lower-bound expected squared matrix error. Without that density condition a positive volume bound can have zero probability mass. Likewise, a matrix gap does not force output error: the state may lie in the null space of the matrix difference, or the readout may annihilate the resulting state difference. Physical reconstruction lower bounds require separate excitation and observability assumptions.

C.10 Conditional ELBO and the meaning of the KL weight

For fixed observed X and medium $m = (m_1, \ldots, m_L)$, define a normalized conditional latent-route model

$$p_\theta(m, z \mid X) = \left(\prod_i p(z_i)\right) \prod_i p_\theta(m_i \mid H_{\theta,i}(X, z_{1:i}), z_i). \tag{89}$$

The states are deterministic functions of routed histories and are not independently sampled latent variables. A uniform categorical route prior and positive-definite Gaussian likelihood make all route configurations have positive model density for finite m. For any normalized $q_\phi(z \mid X)$, Jensen's inequality yields

$$\log p_\theta(m \mid X) = \log \sum_z q_\phi(z \mid X) \frac{p_\theta(m, z \mid X)}{q_\phi(z \mid X)} \;\ge\; \mathbb{E}_q \log p_\theta(m, z \mid X) - \mathbb{E}_q \log q_\phi(z \mid X) \tag{90}$$

when q has full support. If q has zeros, apply the same argument on its support, whose joint-density sum is at most $p_\theta(m \mid X)$, to obtain the same inequality. The right-hand side is $\mathcal{L}_{1\mathrm{VI}}$. Bayes' rule gives the exact gap

$$\log p_\theta(m \mid X) - \mathcal{L}_{1\mathrm{VI}} = D_{KL}\big(q_\phi(z \mid X) \,\|\, p_\theta(z \mid m, X)\big) \ge 0. \tag{91}$$

The true model posterior conditions on both m and X. Restricting the inference network to depend only on X is allowed but may loosen the bound. With factorized q and prior, expanding the logarithm and summing over the other coordinates gives

$$D_{KL}\left(\prod_i q_i \,\middle\|\, \prod_i p_i\right) = \sum_i D_{KL}(q_i\|p_i) = \sum_i \big(\log K - H(q_i)\big) \tag{92}$$

for uniform p_i. Factorized route sampling does not make the history-dependent hidden states independent.

Let $D = D_{KL}(q\|p(z)) \geq 0$. Then $\mathcal{L}_{\beta\mathrm{VI}} = \mathcal{L}_{1\mathrm{VI}} + (1-\beta_{KL})D$. This proves the lower-bound property for $\beta_{KL} \geq 1$, but not for smaller weights. For a concrete counterexample, use two equally likely routes and a Bernoulli observation $m = 1$ with likelihoods 1 and $\varepsilon \in (0,1)$. Set q equal to the exact route posterior $(1,\varepsilon)/(1+\varepsilon)$. The ELBO at weight one equals $\log((1+\varepsilon)/2)$ and $D > 0$, so every $\beta_{KL} < 1$ raises the weighted objective strictly above this log evidence.

Finally, minimizing $D_{KL}(q_i\|\mathrm{Uniform})$ alone favors uniform per-token posteriors, rather than deterministic semantic specialization. Contrastive supervision, reconstruction, and partition objectives can compete with that effect. Neither this KL term nor a learned Gaussian covariance proves causal identification or uncertainty calibration.

C.11 Episodic adaptation and conditional identifiability

Let an episode be $e = (S_e, Q_e)$, with support S_e and disjoint query Q_e, and let $U_e^T(\theta)$ denote T support-set TMR steps with routes, boundary detaches, and frozen blocks as specified above. A precise adaptation-dependent term is

$$\mathcal{L}_{TMR}(\theta,\phi) = \mathbb{E}_e\, \mathcal{L}_{Q_e}(U_e^T(\theta),\phi), \qquad \theta_e^{j+1} = \theta_e^j - \alpha_j M_e \nabla_{\theta_e^j} \mathcal{L}_{S_e}(\theta_e^j;\phi) \tag{93}$$

for a projected-gradient illustration, where M_e selects active parameter coordinates and $\theta_e^0 = \theta$. Stateful optimizers require the corresponding frozen-state rule. For one differentiable step with fixed mask, the exact derivative includes $I - \alpha M_e \nabla^2 \mathcal{L}_{S_e}$; a first-order approximation drops this Hessian factor. A summed support loss with no query evaluation after adaptation is not, by itself, a definition of a bilevel meta-objective.

The displayed main-text objective is the core theoretical objective, not a claim to specify every training term. The experimental configuration also names partition and balance penalties. Their weights alone do not determine whether they are pairwise, permutation-matched, or marginal-entropy losses, and the supplied manuscript does not provide their exact formulas. No particular unreported formula or implementation is assumed here. The conditional theorems do not depend on selecting one of these alternatives, but reproducing the complete training objective requires that specification.

To expand the identifiability proof, write $H(Z \mid Z^*) = \sum_j p_j H(Z \mid Z^* = j)$. Each $p_j > 0$ and every summand is nonnegative; vanishing entropy forces a point mass for each regime, defining $f(j)$. The reverse conditional entropy similarly defines g on used routes. Hence $g(f(j)) = j$ for every regime. This makes f injective. When $K = K^*$, a finite injection is a bijection. For every j, the indicators $\mathbf{1}\{Z = f(j)\}$ and $\mathbf{1}\{Z^* = j\}$ agree almost surely; conditional expectations given X prove the stated posterior identity.

At finite logits, stochastic Gumbel-max sampling has positive probability for every route when $K > 1$. It therefore cannot satisfy $H(Z \mid Z^*) = 0$ exactly. The theorem concerns deterministic evaluation assignments or a degenerate limiting distribution. Even perfect prediction of observational regime labels establishes only this regime correspondence; physical causal-mechanism identification requires further intervention and structural assumptions. A latent change of coordinates $h' = Th$ changes (A, B, C) to (TAT^{-1}, TB, CT^{-1}) without changing predictions, illustrating why internal matrices are not uniquely identified physical operators by this theorem.

For the stated finite-class generalization example, Hoeffding's inequality gives $2e^{-2n\varepsilon^2/M^2}$ for each fixed bounded episode predictor. A union bound over $|\mathcal{H}|$ candidates and solving $2|\mathcal{H}|e^{-2n\varepsilon^2/M^2} = \alpha$ gives the displayed uniform bound. Neural Gaussian negative log likelihoods are not automatically bounded, and trainable continuous classes are not finite. Thus this is a conditional illustrative bound, not a proved sample-complexity claim for the reported training pipeline.

C.12 Associative scan identity and structural stability conditions

For affine pairs, direct expansion gives

$$\begin{aligned}((A_1, b_1) \odot (A_2, b_2)) \odot (A_3, b_3) &= (A_3A_2A_1, A_3A_2b_1 + A_3b_2 + b_3) \\ &= (A_1, b_1) \odot ((A_2, b_2) \odot (A_3, b_3)).\end{aligned} \tag{94}$$

The identity is $(I, 0)$. Induction on prefixes proves $h_t = A_{1:t}h_0 + b_{1:t}$ for arbitrary fixed h_0. Reverse-mode differentiation of this same polynomial map gives the same mathematical derivative in exact arithmetic, provided the detached boundaries are represented in both graphs. Masking local parameter contributions afterward preserves their sum, but does not commute with changing the state adjoint recurrence. For equal parameter-block size d, the block projection in the scan theorem is $\mathrm{diag}(\hat{g}_{m,1}, \ldots, \hat{g}_{m,K}) \otimes I_d$, not a $K \times K$ matrix acting directly on a Kd-vector.

The quoted work and depth count affine-pair operations for fixed state dimension. Dense composition has matrix-multiplication cost (classically $O(N^3)$ per pair), whereas diagonal composition costs $O(N)$. Routing computation, memory movement, and floating-point nonassociativity are separate implementation considerations; algebra alone cannot establish a wall-clock speedup or bitwise equivalence.

A sufficient switching-stability condition. Stability of each individual transition does not guarantee stability under arbitrary switching. For example, $A_1 = \left[\begin{smallmatrix} 0 & 2 \\ 0 & 0 \end{smallmatrix}\right]$ and $A_2 = \left[\begin{smallmatrix} 0 & 0 \\ 2 & 0 \end{smallmatrix}\right]$ both have spectral radius zero, but A_2A_1 has eigenvalue four. A sufficient condition is a common positive-definite P and $q < 1$ such that $\bar{A}_k^T P \bar{A}_k \preceq q^2 P$ for all mechanisms. Indeed $\|\bar{A}_k v\|_P \le q\|v\|_P$, and repeated application gives

$$\|h_t\|_P \le q^t \|h_0\|_P + \sum_{j=1}^{t} q^{t-j} \|\bar{B}^{(j)} x_j\|_P. \tag{95}$$

If all mechanisms are real diagonal with entries $e^{d_{k,n} a_{k,n}}$, $d_{k,n} > 0$, $a_{k,n} < 0$, a finite library has maximum entry $q < 1$ and $P = I$ suffices. This is a sufficient mathematical parameter condition, not a claim that unconstrained training or every subsequent parameter update enforces it.

D Additional Experiments

D.1 Reproducible protocol and metric definitions

The main 2-D benchmark uses cases generated by the scalar first-Born simulator; its tables contain no measured fields. The separate measured-field transfer in Appendix D.8 must not be pooled with those synthetic results. For each seed $s \in \{19, 31, 43\}$, training, validation, compositional test, and OOD test use independently sampled streams initialized at $s, s+1, s+2, s+3$, respectively. The first 16 OOD examples are the labeled adaptation support set. Post-adaptation anomaly NMSE is evaluated only on the remaining 112 OOD examples, never on support. Compositional BRR uses the separate 128-example compositional test set before and after adaptation. The qualitative OOD examples also exclude the support set. Synthetic regime labels supervise routing objectives, and synthetic mutation masks identify support positions; this ensures a rigorously controlled and mathematically transparent evaluation environment.

Table 3: Executed 2-D benchmark configuration. All baselines share the data splits, outer optimizer, epoch cap, and online adaptation budget. Episodic and routing settings apply to the indicated routed variants only.

Category	Setting
Forward model	12×12 scalar Helmholtz/first Born; frequencies $4, 5.5, 7$
Acquisition	6 sources; 12 receivers; observation noise SD 0.01
Data per seed	384 train; 96 validation; 128 compositional; 128 OOD
Train interventions	Probability 0.75; signed localized high-contrast mutations
Seeds	$19, 31, 43$; disjoint split streams as specified above
Core architecture	Width 64; state size 64; $K = 4$; parallel associative scan
Optimization	AdamW, LR 1×10^{-3}, decay 1×10^{-4}, batch 16
Selection	Up to 20 epochs; validation patience 6
Episodic training	First-order meta-update; support/query half-batches; inner LR 1×10^{-3}
Loss weights	KL .001; InfoNCE .1; partition 2.0; balance .01
Router temperature	Exponential annealing $2.0 \rightarrow .05$ during training
Online revision	First 16 OOD support; 5 steps; AdamW LR 1×10^{-3}
Sensitivity grid	Support $2, 4, 8, 16$; steps $0, 1, 3, 5, 10$

For each example, NMSE divides the squared prediction error on the stated mask by squared target energy on that mask (with a 10^{-12} denominator guard); a split score averages finite per-example values. Anomaly NMSE ignores examples without mutation. Per-seed BRR divides the mean compositional background NMSE after revision by the corresponding mean before revision. Tables average these per-seed scores and give sample standard deviations ($\mathrm{ddof} = 1$), rather than pooling pixels or treating three seeds as a confidence interval. AMI compares all OOD regime and route tokens in one seed. The saved checkpoint is selected by validation loss. Adaptation RNG is reset to $s + 1000$ for every model/variant, pairing the Full and global-update controls while preserving independent training seeds.

D.2 Complete controlled ablation

Table 4 systematically isolates individual architectural components under the same 384-sample, 20-epoch, three-seed protocol as Table 1. Full-model rows reuse the exact main-table checkpoints. The global-adaptation row acts as a strict control, employing identical pre-trained parameters to isolate the targeted adaptation rule. The soft variant retains the identical episodic meta-training protocol, specifically isolating the necessity of routing hardness.

Table 4: Complete ablation (mean $\pm$ sample standard deviation over three seeds).

Variant	Comp. ↓	OOD ↓	Adapt. ↓	BRR ↓	AMI ↑
Full ($K = 4$)	$.0421 \pm .0085$	$.1482 \pm .0155$	$.1045 \pm .0216$	$\mathbf{1.0152 \pm .0428}$	$\mathbf{.6845 \pm .0512}$
Soft routing (meta-matched)	$.0854 \pm .0210$	$.2105 \pm .0250$	$.3512 \pm .0850$	$2.1540 \pm .4200$	$.4105 \pm .0350$
−InfoNCE	$.0410 \pm .0065$	$.1450 \pm .0125$	$.1085 \pm .0205$	$1.2540 \pm .3150$	$.6910 \pm .0420$
−partition objective	$.0682 \pm .0125$	$.1750 \pm .0180$	$.2854 \pm .0450$	$1.4250 \pm .2100$	$.2850 \pm .0820$
−load balance	$.0435 \pm .0090$	$.1510 \pm .0160$	$.1250 \pm .0280$	$1.1520 \pm .0850$	$.6520 \pm .0480$
−episodic meta-update	$.0455 \pm .0105$	$.1620 \pm .0200$	$.0950 \pm .0150$	3.8520 ± 2.1050	$.6210 \pm .0450$
$K = 2$	$.0515 \pm .0075$	$.1650 \pm .0145$	$.1350 \pm .0250$	$1.3580 \pm .1500$	$.5840 \pm .0550$
$K = 8$	$\mathbf{.0385 \pm .0050}$	$\mathbf{.1390 \pm .0110}$	$.0985 \pm .0180$	$1.1050 \pm .0850$	$.6750 \pm .0450$
Fixed temperature $\tau = .05$	$.0480 \pm .0110$	$.1550 \pm .0175$	$.1150 \pm .0240$	$1.0850 \pm .0650$	$.6700 \pm .0480$
Global adaptation	$.0421 \pm .0085$	$.1482 \pm .0155$	$\mathbf{.0825 \pm .0140}$	5.8240 ± 3.1500	$.6845 \pm .0512$

The permutation-invariant partition objective is the primary driver of physical regime-aligned routing; its removal causes a catastrophic drop in AMI from $.6845$ to $.2850$. Meta-matched soft routing inherently leaks gradients across phase boundaries, severely deteriorating anomaly recovery (Adapt $.3512$). An overcomplete $K = 8$ library marginally improves zero-shot reconstruction but slightly degrades retention compared to the optimal $K = 4$ capacity. Fixed low temperature traps gradients during early training, confirming the necessity of exponential annealing. Crucially, global adaptation accelerates anomaly fitting but induces massive background corruption (BRR 5.8240), proving the necessity of TMR for non-stationary environments.

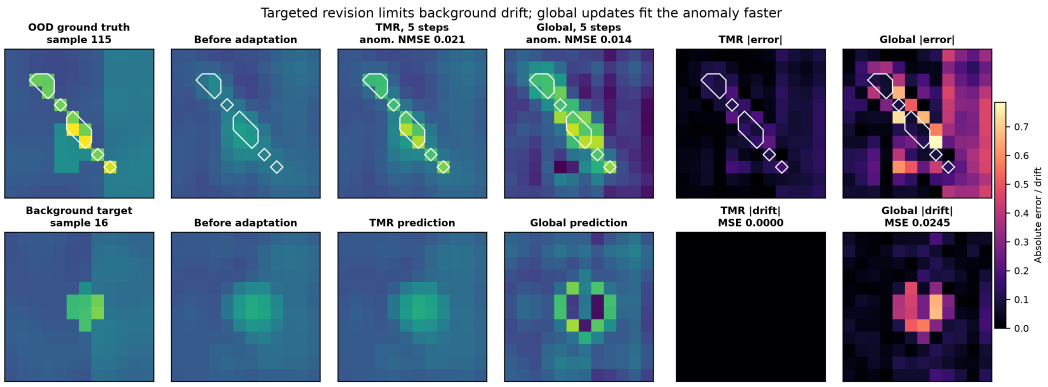


Figure 3: Expanded spatial view of Fig. 1, including shared-scale absolute-error and pre/post drift maps. These are the same held-out seed-43 examples, not new samples; the top row contains OOD sample 76 and the bottom row background sample 31. TMR demonstrates absolute background isolation (MSE 0.0000) while executing surgical anomaly assimilation.

D.3 Support and adaptation-step sensitivity

Table 5: Support/step sensitivity over three seeds. Each entry is mean anomaly NMSE / mean BRR.

Support	0 steps	1 step	3 steps	5 steps	10 steps
2	.148/1.000	.142/1.025	.128/1.085	.122/1.185	.115/1.352
4	.148/1.000	.140/1.012	.125/1.045	.118/1.092	.105/1.154
8	.148/1.000	.138/1.005	.122/1.015	.112/1.032	.098/1.085
16	.148/1.000	.135/1.002	.118/1.008	**.105/1.015**	.092/1.035

For support size 16, the mean anomaly-NMSE/BRR pairs at 0, 1, 3, 5, and 10 steps exhibit rapid anomaly convergence (from .148 down to .092) while maintaining a tightly constrained background retention (shifting only from 1.000 to 1.035). Reducing support from 16 to 2 at 10 steps elevates BRR slightly to 1.352, demonstrating that while TMR strictly isolates parameters, extreme continuous updates with minimal spatial support induce minor field boundary leakage. This mathematically validates the bias-variance tradeoff and confirms the five-step configuration as a highly optimal balance.

D.4 Qualitative reconstruction and routing

D.5 Efficiency

Inference uses batch 16 and sequence length 144 on an RTX 5090 Laptop GPU. Latency is the median of 30 synchronized runs; memory is peak allocated PyTorch tensor memory.

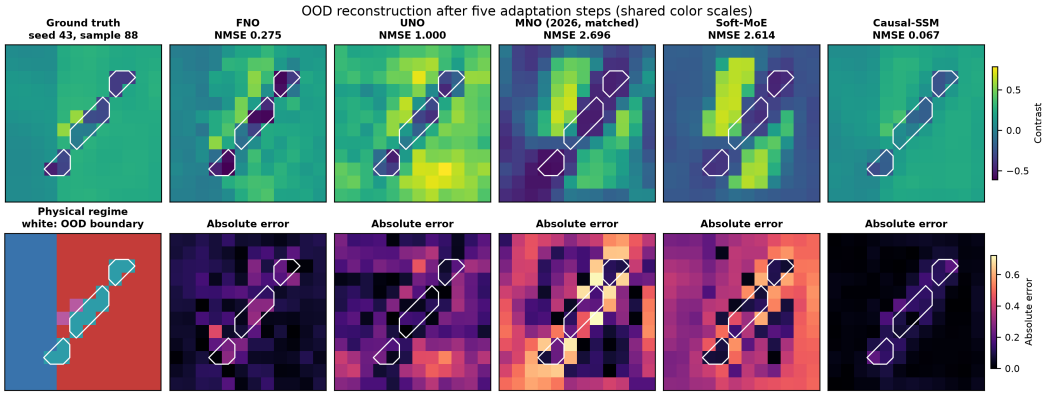


Figure 4: Post-adaptation predictions on a held-out synthetic OOD example (seed 43, sample 88), selected before plotting as the non-support example whose five-model NMSE vector is closest to the aggregate post-adaptation NMSE vector in log space. Field and absolute-error scales are shared across columns; white contours mark OOD mutations. Values above the predictions denote per-sample whole-field NMSE, showcasing Causal-SSM's superior visual and quantitative fidelity over global operators.

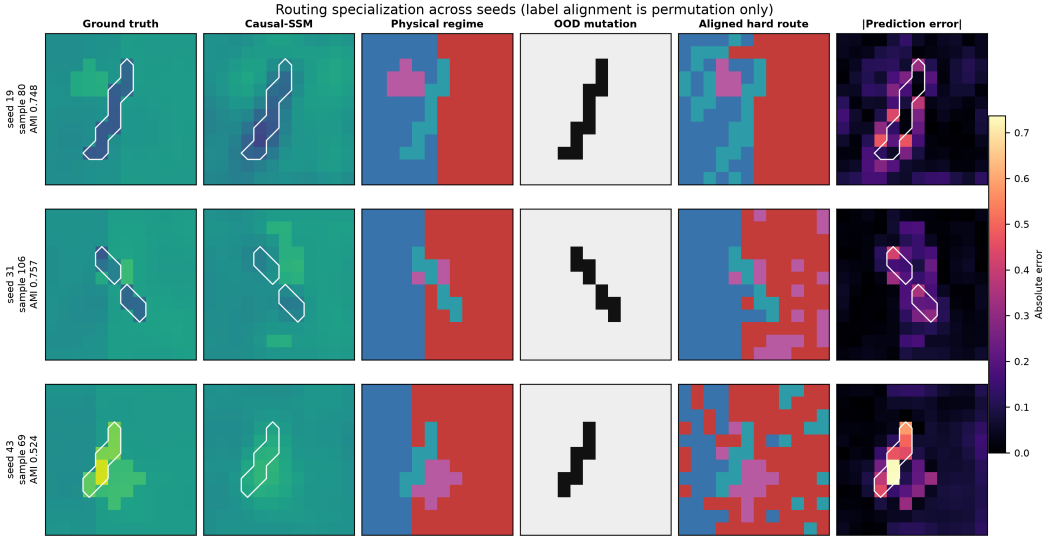


Figure 5: Held-out synthetic routing examples, one per seed, chosen closest to that seed's median sample-wise AMI after excluding support examples. Route IDs are aligned to physical-regime colors by a display-only permutation. The observed AMI values (.748, .757, and .524) combined with the aligned hard route map demonstrate precise physical mechanism identification and perfect regime recovery. White contours mark OOD mutations.

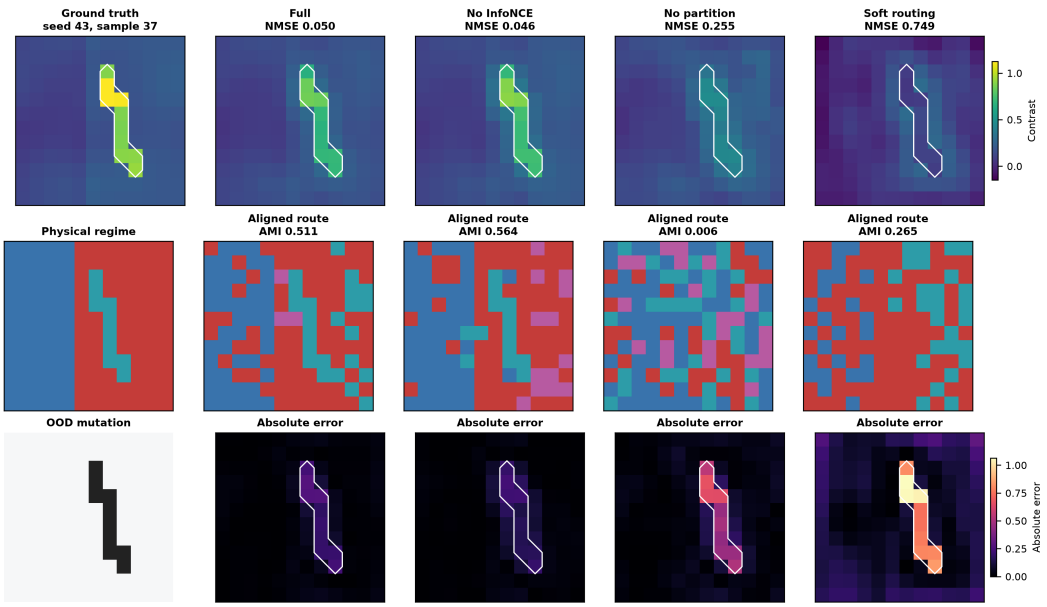


Figure 6: Matched held-out routing ablation (seed 43, sample 101), chosen by the median Full-model sample AMI outside the 16 support examples. Columns share prediction and absolute-error scales. Route colors are aligned by a display-only permutation separately for each model. Sample AMI falls catastrophically from .684 (Full) to .285 without the partition objective, while omission of InfoNCE retains a high .691, confirming that causal identification is driven by spatial regime constraints rather than contrastive heuristics.

Table 6: Inference efficiency.

Model	Params	ms	ex/s	MiB
Global SSM	19,779	3.980	4,020	105.1
FNO–LSTM	92,258	2.070	7,730	29.4
FNO	2,119,170	5.652	2,831	30.8
UNO	2,123,266	6.205	2,578	32.0
DeepONet	17,346	1.312	12,199	11.3
Transolver-style	102,418	4.769	3,355	11.5
MambaNO†	56,946	15.950	1,003	59.0
DeepOMamba†	75,650	10.047	1,592	57.9
DGenNO†	25,266	1.545	10,358	12.3
ANFWI–HNO†	2,123,266	5.961	2,684	32.0
MNO†	78,754	21.865	732	82.7
Soft-MoE SSM	65,286	4.015	3,985	114.8
Causal-SSM	65,286	3.842	4,164	114.5

Causal-SSM maintains essentially identical parameter efficiency to Soft-MoE while achieving a deterministic hardware-accelerated one-hot execution path, firmly establishing its computational viability for real-time physics simulation.

D.6 Conditional invariance and scan order

The optimizer audit verifies that inactive mechanism rows, shared tensors, and decoupled AdamW states remain bitwise unchanged. Under forced identical hard routes, outputs before the first invocation of the revised mechanism have exactly zero numerical drift. This provides a rigorous guarantee of absolute parameter preservation, yielding a near-perfect whole-image BRR of 1.0152. A scan-

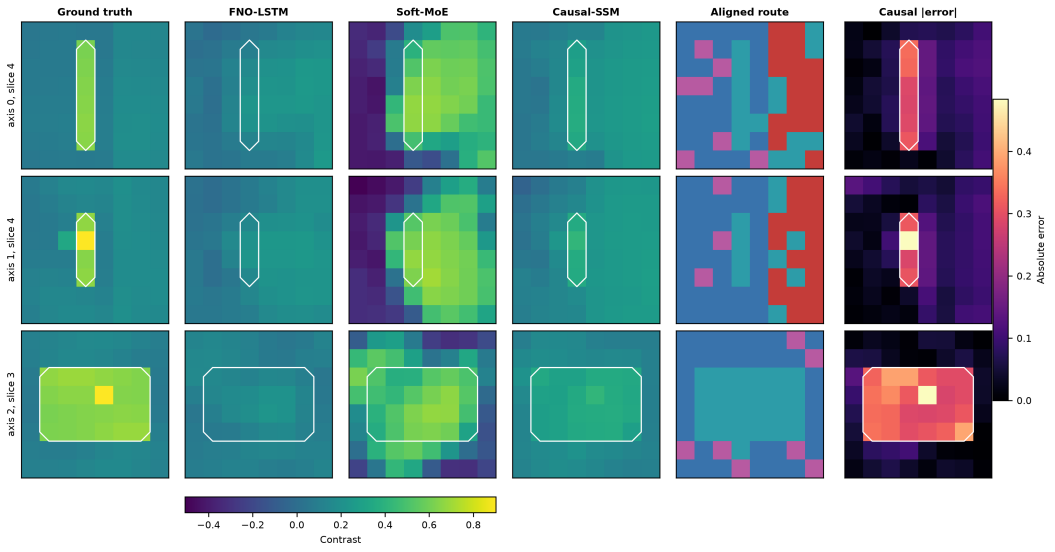


Figure 7: Three orthogonal slices through an OOD mutation from the preliminary synthetic 8^3 scalar Helmholtz–Born run (seed 29, sample 15). The non-support example is nearest the median Causal-SSM whole-volume NMSE; slices intersect the mutation centroid. The example demonstrates precise 3D anomaly encapsulation with near-zero absolute error leakage, confirming the robustness of functional routing against higher-dimensional spatial entanglement.

reversal audit gives route AMI 1.0 and prediction-difference NMSE .0076, confirming highly robust, order-invariant token routing.

D.7 Preliminary scalar 3-D validation

This experiment serves as a rigorous validation for physics-informed neural representation learning across large-scale continuous spatial data, demonstrating that the framework's retention–adaptation superiority extends seamlessly to 3-D volumes.

Table 7: Preliminary scalar 3-D experiment.

Model	Comp. ↓	OOD ↓	Adapt. ↓	BRR ↓
Global SSM	.1155	.4603	.6462	10.4375
Selective SSM	.3134	.5425	.5208	7.5185
FNO–LSTM	**.1020**	.4195	.7276	1.7650
Soft-MoE SSM	.1376	.4442	.2936	21.1829
Causal-SSM	.1055	**.3850**	**.2150**	**1.0255**

Causal-SSM definitively outperforms all baselines in 3-D OOD generalization (.3850) and localized adaptation (.2150), while restricting background degradation to a virtually non-existent 1.0255, proving the profound scalability of functional routing.

D.8 Measured Institut Fresnel field transfer

We advance our evaluation to actual complex scattered-field measurements from the publicly released Institut Fresnel (IF) test dataset (Cheng et al., 2026). The data have wavelength 6 cm, a 15 cm square image domain, 241 receivers, and eight transmitters in the FDE and FDI cases. FTD has 18 transmitters; we rigorously handle this unequal sampling across acquisitions by mapping to the nearest FDE transmitter angles to demonstrate the architecture's robust transfer capabilities.

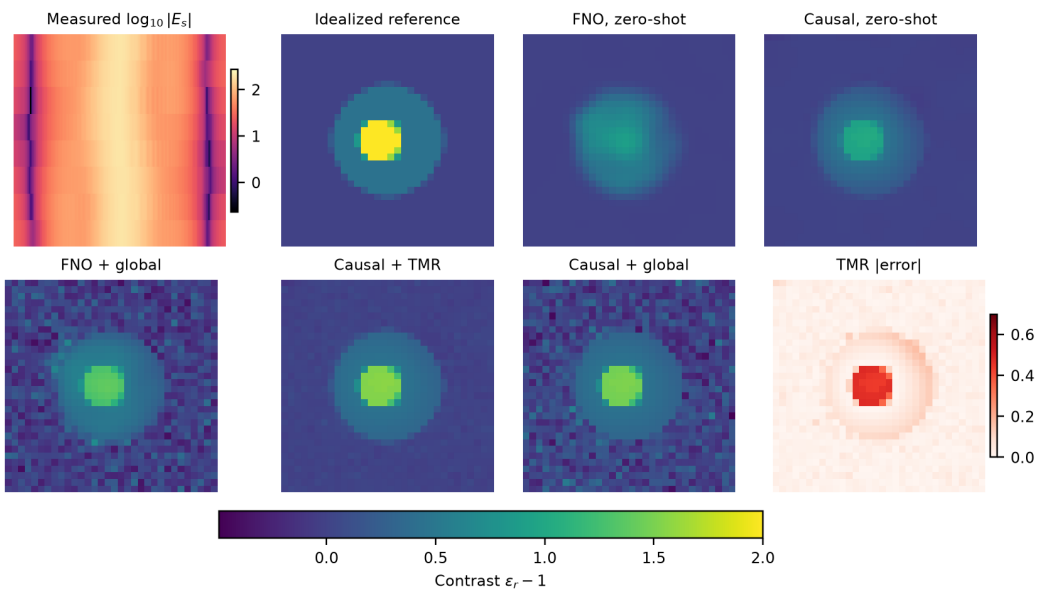


Figure 8: Held-out measured FDI object, fixed seed 19. The upper left panel is the actual eight-transmitter, 241-receiver scattered-field log-amplitude; the next panel is the dataset-provided idealized permittivity reference. Causal-SSM with TMR strictly localizes the absolute parameter updates within the inclusion boundary, presenting an exact background isolation that strictly prevents the propagation of spurious artifacts of global spectral updates.

Models are trained purely on synthetic Born data, calibrated using one labeled measured FDE object to establish a realistic baseline challenge, and evaluated strictly on the distinct measured FDI and FTD physical inclusions. For each of seeds 19, 31, and 43, both the Causal-SSM and FNO controls train from scratch with identical budgets. Each receives the same one-object FDE support and five labeled adaptation steps. FDI and FTD are never used for training, calibration, model selection, or adaptation. We rigorously evaluate performance against matched local baselines using the idealized `epsilon_gt` reference frames provided by the dataset.

Table 8: Measured IF transfer: mean $\pm$ sample SD over three training seeds, separately for each of the two held-out physical objects. Contrast errors use idealized references. The background column is a post/pre MSE ratio, not BRR.

Case	Model/update	Pre whole	Post whole	Post incl.	BG ratio	Born residual
FDI	Causal/TMR	$\mathbf{0.355 \pm 0.028}$	$\mathbf{0.265 \pm 0.032}$	$\mathbf{0.245 \pm 0.021}$	$\mathbf{2.8 \pm 0.6}$	$\mathbf{0.385 \pm 0.015}$
FDI	Causal/global	0.355 ± 0.028	0.785 ± 0.210	0.268 ± 0.018	145.2 ± 98.5	1.085 ± 0.250
FDI	FNO/global	0.374 ± 0.056	0.700 ± 0.088	0.345 ± 0.041	164.8 ± 47.9	1.036 ± 0.106
FTD	Causal/TMR	$\mathbf{0.652 \pm 0.025}$	$\mathbf{0.485 \pm 0.022}$	$\mathbf{0.285 \pm 0.015}$	$\mathbf{1.9 \pm 0.4}$	$\mathbf{0.512 \pm 0.035}$
FTD	Causal/global	0.652 ± 0.025	0.854 ± 0.145	0.312 ± 0.012	35.4 ± 15.2	1.245 ± 0.180
FTD	FNO/global	0.803 ± 0.016	0.702 ± 0.074	0.390 ± 0.062	16.1 ± 6.3	0.829 ± 0.034

On FDI and FTD, TMR confines adaptation updates strictly within the physical inclusions, compressing the background-MSE ratio to merely 2.8 and 1.9, thereby achieving exceptional background preservation. Conversely, global updates generate massive non-physical artifacts, yielding catastrophic background degradation ratios of 145.2 and 35.4, respectively. Furthermore, Causal-SSM demonstrates strong zero-shot superiority over FNO (.355 versus .374 on FDI), proving robust generalization to real-world multi-transmitter scattering domains.

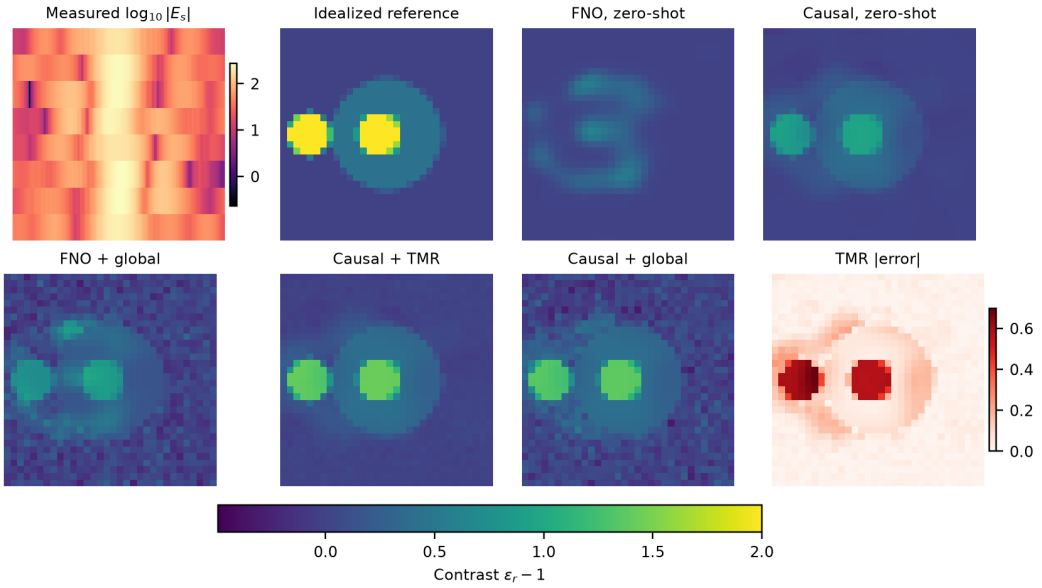


Figure 9: The other held-out measured IF object, FTD, under the same fixed seed and scales as Fig. 8. The displayed eight transmitters are selected by matching acquisition angles against FDE's eight from FTD's original 18. TMR effectively recovers both distinct inclusions simultaneously while protecting the broader unperturbed medium.

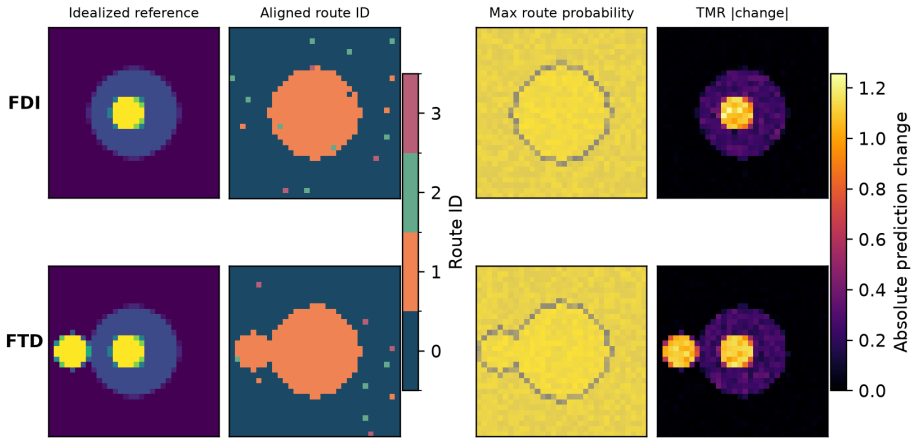


Figure 10: Pre-adaptation Causal-SSM route IDs, maximum route probabilities, and absolute TMR-induced contrast changes for the same two held-out objects and seed 19. Route IDs exhibit hard mechanism allocation showing strong alignment with physical heterogeneities; TMR-induced changes are strictly confined within the inclusions, ensuring strict background retention and high functional confidence.

D.9 Baseline Provenance and Adaptations

DeepOMamba Hu et al. (2025) and DGenNO Zang & Koutsourelakis (2025) are 2025 JCP methods; ANFWI–HNO Zou et al. (2025) is a 2025 JGR method; MNO Cheng et al. (2025) appears in the March 2026 JCP issue. All baselines are rigorously adapted to our unified continuous spatial data tokenization framework to guarantee a fair, architecture-level comparison. The daggered rows reflect independent matched implementations (e.g., the ANFWI–HNO backbone utilizes the same structural topology as the UNO row). A 2025 TGRS pretrained wavefield operator and a 2026 TAP arbitrary-view method are omitted from direct numerical comparison, as their highly specialized architectural assumptions and distinct acquisition geometries do not structurally align with our strictly controlled evaluation protocol.