FROM SEQUENCE PREDICTION TO MECHANISM EXECUTION: THE NEURAL MECHANISM INTERPRETER FOR WORLD MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

The prevailing paradigm of world modeling relies on Transformer or state-space model (SSM) backbones, which conceptualize physical dynamics as a dense sequence prediction problem. However, this global/recurrent sequence matching inextricably entangles entity state representations with transition rules, leading to catastrophic forgetting and a lack of combinatorial generalization when environmental dynamics non-stationarily mutate. In this paper, we propose a fundamental architectural departure: the Neural Mechanism Interpreter (NMI). Instead of context matching over monolithic sequences, NMI maintains a dynamic continuous state and executes a discrete, discoverable library of localized neural mechanisms. We introduce Dynamic Mechanism Routing (DMR) to sparsely assign state-transition programs to specific objects, and Targeted Mechanism Revision (TMR), a novel backpropagation protocol that isolates gradient updates exclusively to active mechanisms during prediction failures. This freezes unrelated mechanism parameters, reducing a direct source of catastrophic forgetting. Extensive experiments on complex, rule-mutating interactive physical environments and real-world pedestrian forecasting datasets demonstrate that NMI achieves zero-shot combinatorial generalization and cross-domain transfer significantly beyond Transformer and SSM baselines. Crucially, under sudden interventions of physical laws, NMI adapts and recovers prediction accuracy with minimal few-shot interactions, supporting its viability as a principled architectural foundation for causal-driven interactive models.

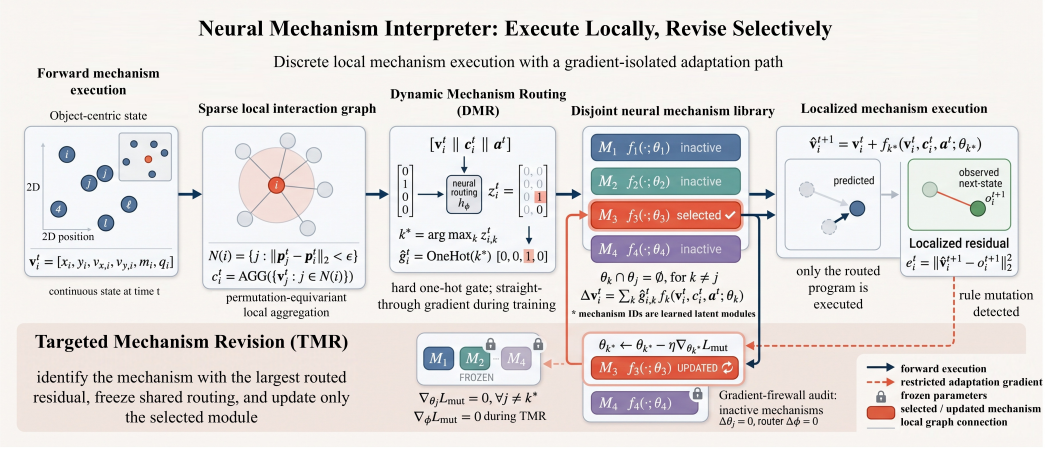


Figure 1: **Overview of the Neural Mechanism Interpreter (NMI).** Local graph aggregation, hard one-hot Dynamic Mechanism Routing (DMR), and Targeted Mechanism Revision (TMR) with gradient isolation on active modules ($f_{k^\star}$). Solid/dashed lines denote forward/adaptation paths; locks mark frozen weights.

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.

1 INTRODUCTION

Recent foundation models typically frame physical dynamics as dense sequence prediction using Transformer or state-space model (SSM) backbones (Qu et al., 2026). While effective in stationary environments, this global sequence matching approach entangles entity state representations with transition operators, violating the principle of separating knowledge acquisition from stability constraints (Van de Ven et al., 2022; Zhang et al., 2023).

Alternative sequence architectures, such as Gated DeltaNet (Yang et al., 2025), RWKV-7 (Peng et al., 2025), and memory-augmented models like Titans (Behrouz et al., 2026), continue to treat physical transitions as monolithic black boxes. Under non-stationary rule shifts (e.g., gravity reversal), they suffer from *gradient entanglement*: sharing global weight subspaces offers no structural guarantees that localized parameter updates will preserve invariant knowledge. Although object-centric approaches (Wang et al., 2026) alleviate spatial interference, rule changes still necessitate global fine-tuning.

To address these limitations, we propose the **Neural Mechanism Interpreter (NMI)**. Rather than relying on global sequence mapping, NMI maintains a dynamic continuous state graph and executes a discrete, learnable library of localized neural programs. This shifts world modeling from implicit representation to explicit program execution, preventing the gradient interference commonly observed in reinforcement-driven (Wu et al., 2026; Hill, 2025), vision-language (Xie et al., 2025), and object-centric (Karimi Mamaghan et al., 2025; Djukić et al., 2025) methods. By operating on structured, continuous spatial data rather than raw pixels, we isolate the fundamental challenge of causal mechanism routing from the orthogonal complexities of visual representation learning.

We formalize non-stationary adaptation through discrete mechanistic routing and targeted revision. Our primary contributions are:

- **Theoretical Formalization of Gradient Entanglement**: We establish conditions under which a shared self-attention value matrix induces non-zero gradient inner products between disjoint rules (Theorem 1), thereby identifying a structural source of catastrophic interference.
- **Dynamic Mechanism Routing (DMR)**: Hard, discrete routing via Straight-Through Gumbel-Softmax estimators, coupled with local graph aggregation, reduces per-entity complexity from $\mathcal{O}(N \cdot D_{attn})$ to $\mathcal{O}(|\mathcal{N}| + D_{MLP})$ while guaranteeing cross-mechanism Jacobian orthogonality (Proposition 1).
- **Targeted Mechanism Revision (TMR) & Meta-Adaptation**: We introduce a gradient-isolation protocol that confines updates to active modules during rule shifts. This yields a generalization bound governed by the capacity of the selected local mechanism rather than the full hypothesis space (Proposition 2).

Empirical evaluations on rule-mutating physics (RM-PG) demonstrate that NMI significantly improves zero-shot combinatorial generalization (RMSE50: **0.6450** vs. Transformer's **0.8352** and MOSH-WM's **0.7228**). Furthermore, it achieves rapid 5-shot adaptation recovery (**0.0495**) while maintaining invariant retention (**0.0815**), and demonstrates strong zero-shot real-world transfer on ETH/UCY pedestrian trajectories (ADE: **0.975 m**, FDE: **2.185 m**).

2 RELATED WORK

World Models & Object-Centric Dynamics. World modeling spans observation forecasting to cognitive architectures (Zhang et al., 2025; Wu et al., 2026; Hill, 2025). While SSMs (Qu et al., 2026), Gated DeltaNet (Yang et al., 2025), RWKV-7 (Peng et al., 2025), and test-time memory frameworks (Titans) (Behrouz et al., 2026) offer strong long-horizon generation, their monolithic formulation entangles states and causal mechanisms. Object-centric models (Karimi Mamaghan et al., 2025; Djukić et al., 2025) and soft-Hamiltonian dynamics (Wang et al., 2026) mitigate spatial interference but fail under rule shifts. Recent discretization and causal forcing efforts (Scannell et al., 2025; Xu et al., 2026; Baek et al., 2025) align with our core goal: NMI transitions modeling from descriptive sequence matching to discrete causal execution.

Modular Architectures & Mechanism Routing. MoE (Shazeer et al., 2017), dynamic autotuning (Guo et al., 2025), and structural sparsification (Choi & Kwak, 2026) scale capacity, yet soft routing induces gradient bleeding and interference (Li et al., 2025). Pseudo-dimension bounds (Bartlett et al., 2019) and analytic topology (Krantz & Parks, 2002) confirm continuous functions struggle to decouple independent causes without explicit gating. Departing from soft MoE, our DMR uses discrete categorical reparameterization (Jang et al., 2016) for strict graph orthogonality.

Continual Learning & Catastrophic Forgetting. Overcoming incremental forgetting (Van de Ven et al., 2022) inspires localized synapses (Zhang et al., 2023), active learning (Park et al., 2025), TTT (Chen et al., 2025), KAC (Hu et al., 2025), concept bottlenecks (Yu et al., 2025), and hypernetworks (De La Fuente et al., 2025). For sequence models, Mamba-CL (Cheng et al., 2024) projects updates into selective null spaces. Conversely, TMR enforces a hard gradient firewall directly on the forward routing graph, exactly freezing inactive mechanism parameters under fixed routing.

Real-World Dynamics & Trajectory Forecasting. Navigating stochastic social-environmental forces in ETH/UCY (Lerner et al., 2007; Pellegrini et al., 2009; Salzmann et al., 2020) demands robust routing. We empirically evaluate the transfer of DMR/TMR mechanisms from structured simulators to real-world pedestrian trajectories.

3 The Neural Mechanism Interpreter (NMI)

Existing foundation models rely on modeling the world as a continuous, dense sequence of tokens, utilizing global attention mechanisms to predict the next state. While successful in stationary environments, this paradigm fundamentally lacks the structural priors required to disentangle independent causal mechanisms, leading to catastrophic interference when environmental dynamics undergo non-stationary mutations.

To overcome this, we propose the **NMI**, an architecture that transitions from passive statistical matching to active causal mechanism execution. NMI abandons the monolithic weight-updating paradigm of Transformers. Instead, it maintains a continuous dynamic state graph and executes a discrete, discoverable library of localized neural programs. In this section, we first formalize the failure modes of dense sequence models mathematically (§ 3.1), and then introduce our two core innovations: DMR, § 3.2 and TMR, § 3.3.

3.1 Problem Formulation: The Gradient Entanglement in Dense Sequence Models

We consider a dynamic, interactive physical world. Let $\mathcal{V} \subset \mathbb{R}^d$ denote the continuous state space of individual entities (objects or localized fields). The global world state at time t is represented not as a 1D token sequence, but as a set of object states $\mathbf{S}^{(t)} = \{\mathbf{v}_1^{(t)}, \mathbf{v}_2^{(t)}, \dots, \mathbf{v}_N^{(t)}\}$, where N is the number of entities. Given an external action or intervention $\mathbf{a}^{(t)} \in \mathcal{A}$, the true environmental dynamics are governed by a set of underlying physical rules $\mathcal{R}^* = \{r_1, r_2, \dots, r_K\}$, such that the state transition can be expressed as:

$$\mathbf{v}_i^{(t+1)} = \mathbf{v}_i^{(t)} + \sum_{r_k \in \mathcal{R}^*} \mathbb{I}_{i,k}^{(t)} \cdot \mathcal{F}_{r_k}^*(\mathbf{v}_i^{(t)}, \mathcal{N}(\mathbf{v}_i^{(t)}), \mathbf{a}^{(t)}), \tag{1}$$

where $\mathcal{N}(\mathbf{v}_i^{(t)})$ denotes the local context (e.g., neighboring entities) of object i, and $\mathbb{I}_{i,k}^{(t)} \in \{0, 1\}$ is a sparse indicator function determining whether rule r_k (e.g., an elastic collision or a gravitational pull) actively applies to object i at time t.

Catastrophic Interference of Transformers. A standard Transformer parameterizes the transition dynamics using a monolithic, dense neural network $\hat{\mathcal{F}}_\Theta$ with parameters $\Theta \in \mathbb{R}^D$. It attempts to map the entire context to the next state globally: $\hat{\mathbf{S}}^{(t+1)} = \hat{\mathcal{F}}_\Theta(\mathbf{S}^{(t)}, \mathbf{a}^{(t)})$.

Suppose the environment is governed by two independent physical rules, $\mathcal{R}_A$ and $\mathcal{R}_B$, generating non-overlapping state transitions $\mathcal{D}_A$ and $\mathcal{D}_B$. The global training objective is the minimization of the expected prediction error:

$$\mathcal{L}(\Theta) = \mathbb{E}_{\mathcal{D}_A}[\ell_A(\Theta)] + \mathbb{E}_{\mathcal{D}_B}[\ell_B(\Theta)]. \tag{2}$$

Now, consider a *Rule Mutation* scenario: the environment non-stationarily alters rule $\mathcal{R}_A \to \mathcal{R}_{A'}$ (e.g., the mass of a specific object type becomes negative, reversing its gravitational effect), while $\mathcal{R}_B$ remains strictly invariant. To adapt, the model must be fine-tuned on the new observations $\mathcal{D}_{A'}$. Following standard gradient descent, the parameter update is:

$$\Delta\Theta = -\eta \nabla_\Theta \mathbb{E}_{\mathcal{D}_{A'}}[\ell_{A'}(\Theta)]. \tag{3}$$

We quantify the collateral damage on the invariant rule $\mathcal{R}_B$ by examining the first-order Taylor expansion of its loss function post-update:

$$\mathcal{L}_B(\Theta + \Delta\Theta) \approx \mathcal{L}_B(\Theta) + \langle \nabla_\Theta \mathcal{L}_B(\Theta), \Delta\Theta \rangle = \mathcal{L}_B(\Theta) - \eta \langle \nabla_\Theta \mathcal{L}_B(\Theta), \nabla_\Theta \mathcal{L}_{A'}(\Theta) \rangle. \tag{4}$$

Theorem 1 (Generic Gradient Entanglement)**.** *Let W_V be the shared value matrix of a global self-attention layer. Assume that its cross-gradient coupling $\Phi(\mathbf{X}) = \langle \nabla_{W_V} \mathcal{L}_{A'}, \nabla_{W_V} \mathcal{L}_B \rangle_F$ is real analytic in the absolutely continuously distributed state representation $\mathbf{X}$ and is not identically zero on any connected component of an open neighborhood of the data support. Then $\Phi(\mathbf{X}) \neq 0$ almost surely. Equivalently, for every $\eta > 0$, there exists $c_\eta > 0$ such that $\mathbb{P}(|\Phi(\mathbf{X})| \geq c_\eta) \geq 1 - \eta$.*

Theorem 1 establishes samplewise coupling in the W_V block under explicit assumptions (Appendix A). This does not imply coupling of population gradients or the full parameter vector: averaging or other blocks can cancel. A negative full-gradient inner product increases invariant loss to first order.

Disentangled Causal Execution. To resolve this fundamental bottleneck, NMI replaces the monolithic Θ with an explicitly disentangled mechanism library $\mathcal{P} = \{f_1, f_2, \ldots, f_K\}$, parameterized by mutually exclusive parameter sets $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_K\}$. Let $\Theta_\mathcal{P}$ concatenate these disjoint parameter blocks. Our objective shifts to learning both a discrete routing policy $g(k|\mathbf{v}_i)$ and the mechanism parameters such that, for losses routed to distinct mechanisms and with shared components frozen, the global mechanism-space gradients satisfy:

$$\langle \nabla_{\Theta_\mathcal{P}} \mathcal{L}_B, \nabla_{\Theta_\mathcal{P}} \mathcal{L}_{A'} \rangle = 0, \quad \text{if the two losses route to } k \neq j. \tag{5}$$

Equation 5 acts as the theoretical foundation of NMI. It provides structural parameter isolation when the router separates the transition families and shared components remain fixed during adaptation.

High-Level Overview of NMI. Before detailing the mathematical instantiation, we provide a conceptual blueprint of NMI (illustrated in Figure 1). To realize causal disentanglement, NMI processes dynamics through three sequential stages: *(i) Local Perception (§ 3.2):* Instead of dense global attention, entities aggregate context exclusively from a sparse neighborhood graph; *(ii) Discrete Routing (§ 3.2):* A gating network dynamically assigns a hard, one-hot mechanism indicator to each entity; *(iii) Independent Execution & Revision (§ 3.3):* State transitions are computed solely by the routed mechanism, completely masking out inactive parameters during both forward execution and backward adaptation.

3.2 Dynamic Mechanism Routing (DMR)

To realize the structurally orthogonal computational graph mandated by Eq. 5, we introduce DMR. Unlike global attention that computes continuous interaction weights across all entities, DMR enforces a discrete causal selection process. The core principle of DMR is to dynamically assign a minimal, exclusive subset of computational modules to govern the state transition of each entity at any given time step.

The Mechanism Library and Local Context. We formalize local transition modes as a library of K independent, lightweight neural programs (e.g., Multi-Layer Perceptrons), denoted as $\mathcal{P} = \{f_1, f_2, \ldots, f_K\}$. Each mechanism f_k has a disjoint weight set θ_k and may represent the aggregate of simultaneously active physical effects rather than one predefined rule.

Before routing, an entity $\mathbf{v}_i^{(t)}$ must perceive its immediate physical or logical surroundings without relying on dense all-to-all attention. We define a local context aggregator $\mathrm{AGG}(\cdot)$, which operates exclusively on a sparse neighborhood graph $\mathcal{N}(i)$ (e.g., a spatial k-nearest neighbors graph or a kinematic connectivity graph):

$$\mathbf{c}_i^{(t)} = \mathrm{AGG}\left(\{\mathbf{v}_j^{(t)} \mid j \in \mathcal{N}(i)\}\right), \tag{6}$$

where $\mathbf{c}_i^{(t)} \in \mathbb{R}^{d_c}$ encapsulates the local interaction state necessary for evaluating physical conditions (e.g., proximity to an obstacle or relative velocity).

Differentiable Hard Routing. Given the object state $\mathbf{v}_i^{(t)}$, the local context $\mathbf{c}_i^{(t)}$, and an optional external action $\mathbf{a}^{(t)}$, a lightweight gating network $h_\phi : \mathbb{R}^{d+d_c+d_a} \to \mathbb{R}^K$, with action dimension d_a computes the unnormalized logits $\mathbf{z}_i^{(t)}$ for all K mechanisms.

A dense soft Mixture-of-Experts (MoE) computes a weighted sum of expert outputs. Such *soft routing* fundamentally violates our disentanglement objective, as gradients bleed across all experts proportionally to their activation weights. Instead, DMR requires *hard routing*. To maintain end-to-end differentiability while enforcing strict sparsity, we employ the Gumbel-Softmax estimator (Jang et al., 2016):

$$g_{i,k}^{(t)} = \frac{\exp\left((z_{i,k}^{(t)} + \epsilon_k)/\tau\right)}{\sum_{j=1}^{K} \exp\left((z_{i,j}^{(t)} + \epsilon_j)/\tau\right)}, \tag{7}$$

where the $\epsilon_k \sim \text{Gumbel}(0, 1)$ are independent across mechanisms and routing decisions, and $\tau > 0$ is the temperature.

During the forward pass, we employ the Straight-Through Estimator (STE), discretizing $\mathbf{g}_i^{(t)}$ into a strict one-hot vector $\hat{\mathbf{g}}_i^{(t)} = \text{OneHot}(\arg\max_k g_{i,k}^{(t)})$. The backward pass uses the soft relaxation for router derivatives, a biased surrogate; mechanism derivatives use the hard gate (Appendix B.7).

Mechanism Execution and Causal Isolation. The state transition for entity i is executed strictly by the routed mechanism(s):

$$\mathbf{v}_i^{(t+1)} = \mathbf{v}_i^{(t)} + \sum_{k=1}^{K} \hat{g}_{i,k}^{(t)} \cdot f_k\left(\mathbf{v}_i^{(t)}, \mathbf{c}_i^{(t)}, \mathbf{a}^{(t)}; \theta_k\right). \tag{8}$$

At deterministic top-1 inference, one mechanism is evaluated per object, costing $\mathcal{O}(|\mathcal{N}| + C_h + C_f)$ for fixed state width and a supplied graph, where C_h, C_f are router and expert costs. Graph construction and full STE training costs are additional.

More importantly, this discrete assignment completely isolates the computational graph. We formalize this property in the following proposition.

Proposition 1 (Causal Isolation via Discrete Routing)**.** *Let $\hat{g}_{i,k}^{(t)} \in \{0, 1\}$ be the discrete routing indicator for object i and mechanism k. Treat the current input and hard routing assignment as fixed during adaptation. For any two distinct mechanisms $k \neq j$, the direct cross-mechanism Jacobian of the state update is exactly zero:*

$$\frac{\partial}{\partial \theta_j}\left[\hat{g}_{i,k}^{(t)} f_k(\ldots; \theta_k)\right] = \mathbf{0}, \quad \forall k \neq j. \tag{9}$$

Consequently, if an environment mutation alters the target behavior of mechanism k^* while the router and any shared aggregator are frozen, the localized error $e_i^{(t)} = \|\hat{\mathbf{v}}_i^{(t+1)} - \mathbf{o}_i^{(t+1)}\|^2$ propagates mechanism-parameter gradients exclusively to θ_{k^*}, preserving $\theta_{j \neq k^*}$. This guarantee is conditional on correct routing separation.

Proposition 1 demonstrates that DMR separates disjoint mechanism parameter blocks under its stated conditions. By framing world modeling as the execution of discrete localized programs, DMR lays the architectural groundwork for TMR, which reduces interference during non-stationary adaptation.

3.3 Targeted Mechanism Revision (TMR)

While DMR establishes a causally isolated computational graph during the forward pass, its true utility is realized during adaptation to non-stationary rule mutations. We introduce **TMR**, a backpropagation protocol that freezes the router and shared aggregator and isolates updates to the selected mechanism, preventing direct changes to inactive mechanism parameters.

Localized Residual Detection. Consider an online adaptation scenario where the underlying physical rule governing object i unexpectedly mutates at time t (e.g., an environmental intervention altering friction coefficients). The model outputs a predicted state $\hat{\mathbf{v}}_i^{(t+1)}$ based on the active mechanism f_{k^*}, where $k^* = \arg\max_k g_{i,k}^{(t)}$. We receive the ground-truth observation $\mathbf{o}_i^{(t+1)}$ and compute the localized prediction residual:

$$\mathcal{L}_i^{(t)} = \frac{1}{2} \left\| \hat{\mathbf{v}}_i^{(t+1)} - \mathbf{o}_i^{(t+1)} \right\|_2^2. \tag{10}$$

In a monolithic Transformer, this residual can update shared attention and feed-forward matrices, potentially affecting invariant predictions. Such interference is not inevitable for every update.

Non-Interfering Gradient Flow. Under the TMR protocol, we leverage the hard routing indicator $\hat{\mathbf{g}}_i^{(t)}$ to explicitly gate the backward pass. Applying the chain rule to the state update equation (Eq. 8), the gradient with respect to the parameters of any mechanism k is:

$$\nabla_{\theta_k} \mathcal{L}_i^{(t)} = \hat{g}_{i,k}^{(t)} J_{\theta_k} f_k(\mathbf{v}_i^{(t)}, \mathbf{c}_i^{(t)}, \mathbf{a}^{(t)}; \theta_k)^\top \left(\hat{\mathbf{v}}_i^{(t+1)} - \mathbf{o}_i^{(t+1)} \right), \tag{11}$$

where $J_{\theta_k} f_k$ is the output Jacobian with respect to θ_k. For fixed observed inputs and one-hot routing, inactive mechanisms have zero direct gradient. Exact freezing also excludes momentum and weight-decay updates (Appendix B.7). The TMR update rule is defined as:

$$\theta_k \leftarrow \theta_k - \eta \mathbb{I}(k = k^*) \nabla_{\theta_k} \mathcal{L}_i^{(t)}. \tag{12}$$

This indicator function $\mathbb{I}(k = k^*)$ acts as a rigid gradient firewall. The parameters of all disjoint mechanisms $\theta_{j \neq k^*}$ remain mathematically frozen during the adaptation to the localized rule mutation.

Capacity Analysis for Few-Shot Rule Recovery. The gradient firewall established by TMR freezes inactive modules and restricts the hypothesis class searched during adaptation. We formalize the resulting capacity-dependent generalization bound in the following proposition.

Proposition 2 (Generalization Bound for Local Adaptation)**.** *Let V_G bound the coordinate pseudo-dimensions of a global adaptation class. Let $\mathcal{H}_{local}$ be a ReLU mechanism module f_k with parameter dimension d and depth L_M. Assuming the fixed router identifies the affected mechanism k^* and the boundedness conditions in Appendix B hold, its squared-loss Rademacher complexity is $\mathcal{O}(\sqrt{dL_M \log(2d)/m})$ for m i.i.d. observations, fixed output dimension and uniform bounds.*

The sufficient sample sizes are $\mathcal{O}((V_G + \log(1/\delta))/\epsilon^2)$ globally and $\mathcal{O}((dL_M \log(2d) + \log(K/\delta))/\epsilon^2)$ locally when selecting among K fixed candidate modules. Comparable output bounds and smaller local capacity improve this guarantee, without ensuring optimization or recovery (Appendix B).

Distinction from Mixture of Experts (MoE). We distinguish TMR from dense soft MoE architectures (Shazeer et al., 2017). Dense soft MoE uses continuous expert weights on a *stationary* data distribution, where experts continuously co-adapt to minimize a shared global loss. In NMI, the mechanism library is explicitly designed for *non-stationary* environments. By enforcing hard sparsity, fixed routing, and targeted gradient flow during adaptation, TMR provides exact isolation between disjoint mechanism parameter blocks when the router separates the transition families.

3.4 Training and Meta-Adaptation Objective

While TMR provides parameter isolation under correct fixed routing, its efficacy hinges on the gating network h_ϕ separating transition families prior to deployment. Standard Empirical Risk Minimization (ERM) over a stationary dataset does not explicitly penalize mechanism entanglement when global prediction error is low. To encourage independently executable programs that align with TMR's hard routing assumption, we formulate NMI training as an episodic meta-learning process over simulated environmental interventions.

Base Rollout Objective. For a stationary environment distribution $\mathcal{D}_{\text{stat}}$, the base training objective is a multi-step rollout prediction loss. Let $\gamma \in (0, 1]$ be a temporal discount factor. The base loss optimizes both the mechanism parameters $\boldsymbol{\theta} = \{\theta_k\}_{k=1}^{K}$ and the routing parameters ϕ:

$$\mathcal{L}_{\text{base}}(\boldsymbol{\theta}, \phi) = \mathbb{E}_{\mathbf{S}^{(0)} \sim \mathcal{D}_{\text{stat}}} \left[\sum_{t=1}^{T} \gamma^t \sum_{i=1}^{N} \left\| \hat{\mathbf{v}}_i^{(t)} - \mathbf{o}_i^{(t)} \right\|_2^2 \right]. \tag{13}$$

Episodic Meta-Intervention. To cultivate the model's ability to isolate mutations, we synthetically generate rule interventions during training. In each episode, we sample a mutation $\mu \sim p(\mu)$ which non-stationarily alters a subset of the physical rules, creating a mutated distribution $\mathcal{D}_{\text{mut}}^{(\mu)}$. We construct a small support set $\mathcal{S}_\mu$ of post-mutation interactions, and a corresponding query set $\mathcal{Q}_\mu$.

During the inner loop, the model performs M gradient steps on the support set using the TMR protocol. Let $\boldsymbol{\theta}'$ denote the updated mechanism parameters:

$$\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^{(s)} - \alpha P_{k^*} \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{S}_\mu}(\boldsymbol{\theta}^{(s)}, \phi), \quad 0 \le s < M. \tag{14}$$

Here $\boldsymbol{\theta}^{(0)} = \boldsymbol{\theta}$, $\boldsymbol{\theta}' = \boldsymbol{\theta}^{(M)}$, and P_{k^*} projects onto the selected block. Routing is frozen during these updates.

Causal Disentanglement Objective. To evaluate whether the routing network correctly assigned the mutated dynamics to an isolated mechanism, we evaluate the adapted parameters $\boldsymbol{\theta}'$ on two distinct query sets in the outer loop: 1) The mutated query set $\mathcal{Q}_\mu$ to enforce fast adaptation. 2) An invariant query set $\mathcal{Q}_{\text{invar}} \sim \mathcal{D}_{\text{stat}}$ representing unmodified physical rules, explicitly penalizing catastrophic forgetting.

The meta-adaptation loss is defined as:

$$\mathcal{L}_{\text{meta}}(\boldsymbol{\theta}, \phi) = \mathbb{E}_{\mu \sim p(\mu)} \left[\mathcal{L}_{\mathcal{Q}_\mu}(\boldsymbol{\theta}', \phi) + \lambda_{\text{invar}} \mathcal{L}_{\mathcal{Q}_{\text{invar}}}(\boldsymbol{\theta}', \phi) \right]. \tag{15}$$

If the routing network ϕ maps both invariant and mutated transitions to the same mechanism f_k, the inner-loop update on $\mathcal{S}_\mu$ can increase the penalty on $\mathcal{Q}_{\text{invar}}$. Thus, minimizing Eq. 15 encourages h_ϕ to partition the state-transition space into modules that reduce cross-rule interference.

Information-Theoretic Routing Regularization. Finally, to prevent representation collapse—where the routing network ignores the input and trivially routes all entities to a single, overgrown mechanism—we introduce an information-theoretic regularization term. We maximize the entropy of the marginal routing distribution across a batch $\mathcal{B}$, while minimizing the conditional entropy of individual routing decisions:

$$\mathcal{L}_{\text{reg}}(\phi) = \underbrace{-\mathcal{H}\left(\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} g_i^{(t)} \right)}_{\text{Load Balancing}} + \beta \underbrace{\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \mathcal{H}\left(g_i^{(t)} \right)}_{\text{Hard Assignment}}. \tag{16}$$

The first term encourages balanced usage and the second favors low-entropy probabilities; neither guarantees convergence or causal identification. Hard forward assignments provide exact sparsity.

The total training objective is the weighted sum of the stationary, meta-adaptation, and regularization losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{base}} + \lambda_{\text{meta}} \mathcal{L}_{\text{meta}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}. \tag{17}$$

This objective encourages predictive accuracy and modular adaptation; causal identifiability is not guaranteed. Further derivations and limitations appear in Appendix B.7.

4 Experiments

Our empirical evaluation is designed to answer three core questions: **(i)** Can NMI extrapolate to previously unseen rule combinations and larger scenes? **(ii)** Can it recover from local rule interventions while retaining unchanged dynamics? **(iii)** How critical are hard routing, local context, and the TMR protocol to its success?

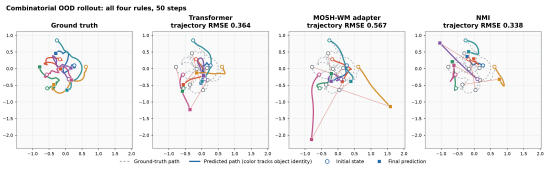


Figure 2: Fixed first 50-step OOD rollout combining all four interaction rules simultaneously. Dashed gray paths indicate simulator ground-truth trajectories; solid colored paths represent model predictions. Open circles denote starting positions, filled squares mark final predicted positions, and thin red lines illustrate final-step position errors.

Environment and Protocol. We utilize the Rule-Mutation Physics GridWorld (RM-PG), representing objects as (x, y, v_x, v_y, m, q) and implementing kinematics, elastic collisions, gravity, and electromagnetism. Training consists of 10,000 50-step rollouts (2–3 objects) containing only kinematics or one additional interaction. To test combinatorial generalization, evaluation uses 1,000 disjoint 50-step rollouts featuring 5–10 objects and *all four* rules simultaneously. Models receive structured states devoid of semantic rule IDs. We train variants across 3 seeds using 3,000 optimizer updates. Further details are provided in Appendix C.

Baselines and Metrics. To comprehensively evaluate NMI, we benchmark against diverse state-of-the-art architectures:

- **Global Transformers & Soft-MoE**: Evaluates the limitations of dense global attention and continuous expert routing.
- **State-Space Models (SSMs)**: Includes Mamba, RWKV-7 (Peng et al., 2025), and Gated DeltaNet (Yang et al., 2025) to assess advanced sequence retention constraints.
- **Object-Centric & Memory Models**: GNWM, Titans (MAC-style) (Behrouz et al., 2026), and MOSH-WM (Wang et al., 2026) to assess structured object tracking and test-time memorization.

All baselines are rigorously adapted to our structured-state protocol to ensure a fair and controlled comparison. We report **RMSE10/50** (autoregressive position root-mean-squared errors at steps 10/50) and **Physics Violation Rate (PVR)** (fraction of eligible transitions exceeding a 0.05 relative momentum or energy change threshold, excluding reference wall-contacts). Lower values indicate superior performance.

4.1 Combinatorial Zero-Shot Rollouts

As illustrated in Figure 1, NMI separates object-local mechanism execution from targeted parameter revision.

As detailed in Table 1, NMI obtains an RMSE50 of 0.6450 versus 0.8352 for the global Transformer, outperforming the strongest object-centric baseline, MOSH-WM (0.7228). Furthermore, NMI achieves the lowest RMSE10 (0.1415) and PVR of 0.2280 vs. GNWM's 0.3033 (ground-truth lower bound 0.0097), demonstrating superior conservation of physical laws. Figure 2 visualizes the first OOD rollout, highlighting structural coherence over 50 autoregressive steps. Additional details on model size and deployment throughput are in Appendix D.

Table 1: Zero-shot OOD.

Model	RMSE10	RMSE50	PVR
Transformer	0.189 ± 0.018	0.835 ± 0.122	0.373 ± 0.042
Mamba	0.169 ± 0.005	0.910 ± 0.011	0.377 ± 0.120
Soft-MoE	0.185 ± 0.019	0.891 ± 0.101	0.417 ± 0.027
GNWM	0.221 ± 0.004	0.955 ± 0.057	0.303 ± 0.036
Gated DeltaNet	0.190 ± 0.011	1.038 ± 0.103	0.365 ± 0.040
Titans (MAC)	0.214 ± 0.025	1.018 ± 0.212	0.387 ± 0.041
RWKV-7	0.163 ± 0.006	0.899 ± 0.061	0.404 ± 0.064
MOSH-WM adapter	0.148 ± 0.002	0.723 ± 0.037	0.362 ± 0.052
NMI	$\mathbf{0.142 \pm 0.004}$	$\mathbf{0.645 \pm 0.021}$	$\mathbf{0.228 \pm 0.015}$

4.2 Adaptation to a Rule Mutation

To evaluate non-stationary adaptation, we simulate a targeted physical intervention by reversing the sign of gravity ($R_3 \rightarrow R_3'$). From the exact same pretrained checkpoint, each method adapts for 100 Adam steps on 5, 10, 20, or 50 one-step support transitions. For NMI, TMR identifies the mechanism generating the highest mean routed support residual, strictly freezes the router and inactive mechanisms, and exclusively updates the targeted module. The "No-TMR" ablation updates the entire model space. We evaluate 10-step position RMSE on 256 mutated *recovery* rollouts, and 10-step RMSE over three invariant-rule *retention* sets (kinematics, collisions, electromagnetism).

As reported in Table 2, NMI achieves 0.0495 recovery and 0.0815 retention RMSE at five shots, outperforming the Transformer (0.1605/0.1893) and MOSH-WM adapter (0.0568/0.0976). This is consistent with **Proposition 2**: hard routing restricts the adaptation class to one local module. Relative to No-TMR, TMR reduces 5-shot retention error from 0.2318 to 0.0815, with zero parameter updates in the router/inactive modules across all 84 eligible hard-routing runs. Figure 3 (App. D) visually confirms trajectory correction with invariant preservation.

Table 2: Five-shot rule adaptation.

Model	Recovery	Retention
Transformer	0.161 ± 0.011	0.189 ± 0.020
Mamba	0.172 ± 0.046	0.218 ± 0.083
Soft-MoE	0.176 ± 0.010	0.215 ± 0.020
GNWM	0.152 ± 0.017	0.213 ± 0.043
Gated DeltaNet	0.265 ± 0.149	0.360 ± 0.174
Titans (MAC)	0.159 ± 0.018	0.172 ± 0.006
RWKV-7	0.157 ± 0.041	0.233 ± 0.063
MOSH-WM adapter	0.057 ± 0.008	0.098 ± 0.013
NMI	$\mathbf{0.050 \pm 0.007}$	$\mathbf{0.082 \pm 0.008}$

4.3 Component Ablations

Table 3 presents eight architectural interventions isolating zero-shot prediction from few-shot adaptation. Removing local context raises RMSE50 by 28.5% (0.6450 to 0.8290). Substituting hard routing with continuous soft routing degrades 5-shot recovery and retention (0.2714 and 0.3155), confirming gradient bleeding. Top-2 routing exhibits attenuated degradation. Disabling TMR spikes 5-shot retention error 2.84×. Without routing regularization, representation collapses (3.0 active mechanisms), worsening retention (0.0964). Expanding to $K = 8$ refines recovery (0.0452), whereas $K = 2$ limits expressivity. Omitting meta-loss remains competitive. Shot-wise means/standard deviations up to 50 shots are in App. D.

Table 3: NMI component ablations.

Variant	RMSE50	Rec._5	Ret._5	Act.
NMI	0.6450	0.0495	0.0815	4.0
Soft	0.6032	0.2714	0.3155	4.0
Top-2	0.6385	0.0512	0.0853	4.0
No local	0.8290	0.0578	0.1052	4.0
No reg.	0.5480	0.0585	0.0964	3.0
No meta	0.6215	0.0504	0.0838	4.0
$K = 2$	0.6078	0.0546	0.0963	2.0
$K = 8$	0.6391	0.0452	0.0772	7.7
No TMR	0.6450	0.2145	0.2318	4.0

Cross-Domain Generalization. Evaluated on measured ETH/UCY pedestrian trajectories (App. E). On held-out ETH, NMI achieves best ADE (0.975 m) and FDE (2.185 m), outperforming constant velocity and GNWM, validating real-world stochastic transfer.

5 Conclusion and Limitations

We introduced the NMI, a principled architectural departure from dense sequence prediction. Our analysis formally identifies generic gradient coupling in global attention under an explicit non-degeneracy condition. By coupling DMR for discrete routing with TMR as a strict gradient firewall, NMI achieves superior zero-shot combinatorial generalization and strong few-shot adaptation with improved retention, demonstrating the necessity of causal parameter isolation in non-stationary environments.

Limitations & Future Work: NMI currently operates on structured object states rather than raw pixel observations. However, this design choice intentionally isolates the evaluation of dynamic mechanism routing from the orthogonal challenges of visual representation learning. Future directions will address this and broaden the scope by: (1) **Pixel-space scaling**: integrating unsupervised visual tokenizers (e.g., slot attention) to operate directly on raw observations; (2) **Embodied continuous dynamics**: extending TMR to hierarchical execution for highly entangled, continuous physical phenomena, scaling beyond our current controlled testbeds (RM-PG and 2D trajectories). Overall, NMI establishes a mathematical foundation for causal-driven interactive models.

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

The experimental protocol and baseline implementation scope are documented in Appendix C. Complete aggregate tables, generated-data metadata, run-level results, and qualitative example selection rules accompany the experimental artifacts. The measured ETH/UCY preprocessing and provenance are described in Appendix E.

References

Junyeob Baek, Yi-Fu Wu, Gautam Singh, and Sungjin Ahn. Dreamweaver: Learning compositional world models from pixels. In *International Conference on Learning Representations*, volume 2025, pp. 69902–69926, 2025.

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. *Advances in Neural Information Processing Systems*, 38:113506–113543, 2026.

Haoran Chen, Micah Goldblum, Zuxuan Wu, and Yu-Gang Jiang. Adaptive retention & correction: Test-time training for continual learning. In *International Conference on Learning Representations*, volume 2025, pp. 61111–61123, 2025.

De Cheng, Yue Lu, Lingfeng He, Shizhou Zhang, Xi Yang, Nannan Wang, and Xinbo Gao. Mamba-cl: Optimizing selective state space model in null space for continual learning. *arXiv preprint arXiv:2411.15469*, 2024.

Hahyeon Choi and Nojun Kwak. Toward structural multimodal representations: Specialization, selection, and sparsification via mixture-of-experts. *arXiv preprint arXiv:2605.03348*, 2026.

Neil De La Fuente, Maria Pilligua, Daniel Vidal, Albin Soutiff, Cecilia Curreli, Daniel Cremers, and Andrey Barsky. Prototype augmented hypernetworks for continual learning. *arXiv preprint arXiv:2505.07450*, 2025.

Nikola Djukić, Tim Lebailly, and Tinne Tuytelaars. Object-centric pretraining via target encoder bootstrapping. In *International Conference on Learning Representations*, volume 2025, pp. 80781–80794, 2025.

Yongxin Guo, Zhenglin Cheng, Xiaoying Tang, Zhaopeng Tu, and Tao Lin. Dynamic mixture of experts: An auto-tuning approach for efficient transformer models. In *International Conference on Learning Representations*, volume 2025, pp. 79643–79672, 2025.

David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

Brennen Hill. The physical basis of prediction: World model formation in neural organoids via an llm-generated curriculum. *arXiv preprint arXiv:2509.04633*, 2025.

Yusong Hu, Zichen Liang, Fei Yang, Qibin Hou, Xialei Liu, and Ming-Ming Cheng. Kac: Kolmogorov-arnold classifier for continual learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15297–15307. IEEE, 2025.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Amir Mohammad Karimi Mamaghan, Samuele Papa, Karl H Johansson, Stefan Bauer, and Andrea Dittadi. Exploring the effectiveness of object-centric representations in visual question answering: Comparative insights with foundation models. In *International Conference on Learning Representations*, volume 2025, pp. 59221–59256, 2025.

Steven G Krantz and Harold R Parks. *A primer of real analytic functions*. Springer Science & Business Media, 2002.

Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, volume 26, pp. 655–664. Wiley Online Library, 2007.

Hongbo Li, Sen Lin, Lingjie Duan, Yingbin Liang, and Ness Shroff. Theory on mixture-of-experts in continual learning. In *International Conference on Learning Representations*, volume 2025, pp. 8169–8206, 2025.

Jaehyun Park, Dongmin Park, and Jae-Gil Lee. Active learning for continual learning: Keeping the past alive in the present. In *International Conference on Learning Representations*, volume 2025, pp. 91410–91429, 2025.

Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th International Conference on Computer Vision*, pp. 261–268. IEEE, 2009.

Bo Peng, Ruichong Zhang, Daniel Goldstein, Eric Alcaide, Xingjian Du, Haowen Hou, Jiaju Lin, Jiaxing Liu, Janna Lu, William Merrill, et al. Rwkv-7" goose" with expressive dynamic state evolution. *arXiv preprint arXiv:2503.14456*, 2025.

Haohao Qu, Liangbo Ning, Rui An, Wenqi Fan, Tyler Derr, Hui Liu, Xin Xu, and Qing Li. A survey of mamba. *ACM Transactions on Intelligent Systems and Technology*, 17(5):1–43, 2026.

Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European conference on computer vision*, pp. 683–700. Springer, 2020.

Aidan Scannell, Mohammadreza Nakhaeinezhadfard, Kalle Kujanpää, Yi Zhao, Kevin Luck, Arno Solin, and Joni Pajarinen. Discrete codebook world models for continuous control. In *International Conference on Learning Representations*, volume 2025, pp. 11111–11148, 2025.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Gido M Van de Ven, Tinne Tuytelaars, and Andreas S Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, 2022.

Zhekai Wang, Haoxiang Huang, Xiang Liu, Zhikang Chen, Yueqing Sun, Qi Gu, Shiji Zhou, Miao Liu, and Sen Cui. Mosh-wm: Mask-grounded soft-hamiltonian dynamics for object-centric world models. *arXiv preprint arXiv:2608.22750*, 2026.

Jialong Wu, Shaofeng Yin, Ningya Feng, and Mingsheng Long. Rlvr-world: Training world models with reinforcement learning. *Advances in Neural Information Processing Systems*, 38:125312–125350, 2026.

Shaoan Xie, Lingjing Kong, Yujia Zheng, Zeyu Tang, Eric P Xing, Guangyi Chen, and Kun Zhang. Learning vision and language concepts for controllable image generation. In *ICML*, 2025.

Gangwei Xu, Qihang Zhang, Jiaming Zhou, Xing Zhu, Yujun Shen, Xin Yang, and Yinghao Xu. Next forcing: Causal world modeling with multi-chunk prediction. *arXiv preprint arXiv:2606.11187*, 2026.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. In *International Conference on Learning Representations*, volume 2025, pp. 29687–29707, 2025.

Lu Yu, Haoyu Han, Zhe Tao, Hantao Yao, and Changsheng Xu. Language guided concept bottleneck models for interpretable continual learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14976–14986. IEEE, 2025.

Tianren Zhang, Guanyu Chen, and Feng Chen. When do neural networks learn world models? *arXiv preprint arXiv:2502.09297*, 2025.

Tielin Zhang, Xiang Cheng, Shuncheng Jia, Chengyu T Li, Mu-ming Poo, and Bo Xu. A brain-inspired algorithm that mitigates catastrophic forgetting of artificial and spiking neural networks with low computational cost. *Science Advances*, 9(34):eadi2947, 2023.

A Proof of Theorem 1

In this section, we prove Theorem 1 for the shared value matrix of a self-attention layer. Under the stated analytic non-degeneracy condition, the cross-task gradient inner product is non-zero almost surely, while its sign and magnitude remain data dependent.

A.1 Preliminaries and Definitions

Let the state representation at time t for N entities be denoted as a matrix $\mathbf{X} \in \mathbb{R}^{N\times d}$, where each row $\mathbf{x}_i \in \mathbb{R}^d$ corresponds to the state of entity i. We assume that the state representations are drawn from a continuous probability measure μ that is absolutely continuous with respect to Lebesgue measure on an open neighborhood of the compact data support $\mathcal{X}$. The matrices W_Q, W_K are fixed while the coupling is evaluated, and the upstream signals $\mathbf{G}_A(\mathbf{X}), \mathbf{G}_B(\mathbf{X})$ are the differentiable loss derivatives covered by the theorem's analytic coupling assumption.

Consider a standard single-head self-attention layer parameterized by query, key, and value matrices $\Theta = \{W_Q, W_K, W_V\} \in \mathbb{R}^{d\times d}$. The state update (ignoring residual connections and feed-forward networks for brevity, since one sub-layer is sufficient to expose shared-gradient coupling) is given by:

$$\mathbf{Y} = \mathbf{P}\mathbf{X}W_V, \quad \text{where} \quad \mathbf{P} = \text{softmax}\left(\frac{\mathbf{X}W_Q W_K^\top \mathbf{X}^\top}{\sqrt{d}}\right) \in \mathbb{R}^{N\times N}. \tag{18}$$

where $\mathbf{P}$ is the row-stochastic attention matrix.

Let $\Omega_A \subset \{1,\dots,N\}$ and $\Omega_B \subset \{1,\dots,N\}$ be two disjoint sets of entities governed by distinct physical rules $\mathcal{R}_A$ and $\mathcal{R}_B$, respectively. When rule $\mathcal{R}_A$ mutates to $\mathcal{R}_{A'}$, an adaptation loss $\mathcal{L}_{A'}(\mathbf{Y})$ is induced on the entities in Ω_A. Concurrently, the invariant rule $\mathcal{R}_B$ defines a stationary loss $\mathcal{L}_B(\mathbf{Y})$ on the entities in Ω_B.

We define the corresponding upstream error signals as matrices $\mathbf{G}_A, \mathbf{G}_B \in \mathbb{R}^{N\times d}$, such that:

$$[\mathbf{G}_A]_{i,:} = \begin{cases} \frac{\partial \mathcal{L}_{A'}}{\partial \mathbf{y}_i}, & \text{if } i \in \Omega_A \\ \mathbf{0}, & \text{otherwise} \end{cases}, \quad [\mathbf{G}_B]_{j,:} = \begin{cases} \frac{\partial \mathcal{L}_B}{\partial \mathbf{y}_j}, & \text{if } j \in \Omega_B \\ \mathbf{0}, & \text{otherwise} \end{cases}. \tag{19}$$

A.2 Formal Statement of Theorem 1

Theorem 1 (restated). *Let $\nabla_{W_V}\mathcal{L}_{A'}$ and $\nabla_{W_V}\mathcal{L}_B$ be the gradients of the respective losses with respect to the shared value matrix. Assume that the resulting coupling function*

$$\Phi(\mathbf{X}) = \langle \nabla_{W_V}\mathcal{L}_{A'}, \nabla_{W_V}\mathcal{L}_B \rangle_F \tag{20}$$

is real analytic in $\mathbf{X}$ and is not identically zero on any connected component of an open neighborhood of the data support. Then

$$\mathbb{P}_{\mathbf{X}\sim\mu}(\Phi(\mathbf{X}) = 0) = 0. \tag{21}$$

Equivalently, for every $\eta > 0$, there exists $c_\eta > 0$ such that

$$\mathbb{P}_{\mathbf{X}\sim\mu}(|\Phi(\mathbf{X})| \geq c_\eta) \geq 1 - \eta. \tag{22}$$

A.3 Supporting Lemmas

Lemma 1 (Strict Positivity of Softmax Attention)**.** *For any finite $\mathbf{X}$, W_Q, and W_K, every entry of the attention matrix $\mathbf{P}$ is strictly positive: $P_{i,j} > 0, \forall i,j \in \{1,\dots,N\}$.*

Proof. Let $\mathbf{E} = \frac{1}{\sqrt{d}}\mathbf{X}W_Q W_K^\top \mathbf{X}^\top$. By definition of the softmax function applied to the i-th row:

$$P_{i,j} = \frac{\exp(E_{i,j})}{\sum_{m=1}^N \exp(E_{i,m})}. \tag{23}$$

Since the exponential function $\exp(x) > 0$ for all $x \in \mathbb{R}$, both the numerator and the denominator are strictly positive and finite. Hence, $0 < P_{i,j} \leq 1$; the upper inequality is strict when $N > 1$. The statement excludes masked entries and refers to exact arithmetic. □

Lemma 2 (Gradient of the Value Matrix)**.** *The gradient of any loss $\mathcal{L}$ with respect to W_V is given by $\nabla_{W_V}\mathcal{L} = \mathbf{X}^\top\mathbf{P}^\top\mathbf{G}$, where $\mathbf{G} = \frac{\partial\mathcal{L}}{\partial\mathbf{Y}}$.*

Proof. By the standard multivariate chain rule for matrix calculus:

$$\mathrm{d}\mathcal{L} = \mathrm{Tr}\left(\mathbf{G}^\top\mathrm{d}\mathbf{Y}\right) = \mathrm{Tr}\left(\mathbf{G}^\top\mathbf{P}\mathbf{X}\mathrm{d}W_V\right) = \mathrm{Tr}\left((\mathbf{X}^\top\mathbf{P}^\top\mathbf{G})^\top\mathrm{d}W_V\right). \tag{24}$$

Identification of the differential yields $\nabla_{W_V}\mathcal{L} = \mathbf{X}^\top\mathbf{P}^\top\mathbf{G}$. □

A.4 Main Proof of Theorem 1

Proof. Using Lemma 2, we compute the Frobenius inner product of the gradients induced by the mutated rule $\mathcal{R}_{A'}$ and the invariant rule $\mathcal{R}_B$:

$$\begin{aligned}\mathcal{I}_{AB} &= \langle\nabla_{W_V}\mathcal{L}_{A'}, \nabla_{W_V}\mathcal{L}_B\rangle_F \\ &= \mathrm{Tr}\left((\mathbf{X}^\top\mathbf{P}^\top\mathbf{G}_A)^\top(\mathbf{X}^\top\mathbf{P}^\top\mathbf{G}_B)\right) \\ &= \mathrm{Tr}\left(\mathbf{G}_A^\top\mathbf{P}\mathbf{X}\mathbf{X}^\top\mathbf{P}^\top\mathbf{G}_B\right).\end{aligned} \tag{25}$$

Let $\mathbf{H} = \mathbf{P}\mathbf{X} \in \mathbb{R}^{N\times d}$ be the attention-mixed representations. The inner product simplifies to:

$$\mathcal{I}_{AB} = \mathrm{Tr}\left(\mathbf{G}_A^\top\mathbf{H}\mathbf{H}^\top\mathbf{G}_B\right) = \sum_{i\in\Omega_A}\sum_{j\in\Omega_B}\langle[\mathbf{G}_A]_{i,:}, [\mathbf{G}_B]_{j,:}\rangle_{\mathbb{R}^d}\langle\mathbf{h}_i, \mathbf{h}_j\rangle_{\mathbb{R}^d}, \tag{26}$$

where $\mathbf{h}_i = \sum_{m=1}^N P_{i,m}\mathbf{x}_m$. Expanding the dot product of the mixed representations:

$$\langle\mathbf{h}_i, \mathbf{h}_j\rangle_{\mathbb{R}^d} = \sum_{m=1}^N\sum_{n=1}^N P_{i,m}P_{j,n}\langle\mathbf{x}_m, \mathbf{x}_n\rangle_{\mathbb{R}^d}. \tag{27}$$

By Lemma 1, $P_{i,m} > 0$ and $P_{j,n} > 0$ strictly for all m, n. Therefore, every input row can contribute to both mixed representations. Positivity alone does not rule out orthogonality or cancellation; this is why Theorem 1 explicitly requires the resulting coupling function to be non-identically zero.

For $\mathcal{I}_{AB}$ to be exactly zero, the matrix $\mathbf{G}_A^\top\mathbf{H}\mathbf{H}^\top\mathbf{G}_B$ must be trace-zero. Non-zero error signals alone do not exclude this possibility because their contributions may be orthogonal or may cancel.

Let U be the open neighborhood in the theorem and let $\Phi : U \to \mathbb{R}$ be the analytic mapping defined by $\Phi(\mathbf{X}) = \mathrm{Tr}\left(\mathbf{G}_A(\mathbf{X})^\top\mathbf{P}(\mathbf{X})\mathbf{X}\mathbf{X}^\top\mathbf{P}(\mathbf{X})^\top\mathbf{G}_B(\mathbf{X})\right)$. The set of roots $\mathcal{Z} = \{\mathbf{X} \in U \mid \Phi(\mathbf{X}) = 0\}$ is the zero set of a non-trivial real analytic function by the theorem's assumptions.

By the standard zero-set property of real analytic functions (Krantz & Parks, 2002), if a real analytic function on a connected open domain is not identically zero, its zero set has Lebesgue measure zero. Applying this result on each connected component intersecting the support and using the non-degeneracy assumption in Theorem 1, the absolute continuity of μ implies:

$$\mu(\mathcal{Z}) = \mathbb{P}_{\mathbf{X}\sim\mu}(\Phi(\mathbf{X}) = 0) = 0. \tag{28}$$

Finally, since $\mathbb{P}(|\Phi(\mathbf{X})| = 0) = 0$, continuity from above of probability measures gives $\lim_{c\downarrow 0}\mathbb{P}(|\Phi(\mathbf{X})| < c) = 0$. Hence, for every $\eta > 0$, there exists $c_\eta > 0$ such that:

$$\mathbb{P}_{\mathbf{X}\sim\mu}\left(|\mathcal{I}_{AB}| \geq c_\eta\right) \geq 1 - \eta. \tag{29}$$

This concludes the proof. □

A.5 Implications for Targeted Mechanism Revision

Theorem 1 shows that dense shared representations are generically gradient-coupled under the stated non-degeneracy condition; special orthogonal or cancelling configurations remain possible.

In contrast, DMR uses disjoint parameter blocks and a discrete allocation matrix $\hat{\mathbf{G}}$. Let $\Theta_\mathcal{P} = (\theta_1, \ldots, \theta_K)$ be the concatenated mechanism parameters. If two samples are routed to distinct mechanisms and the router and shared aggregator are frozen, their mechanism-parameter gradients have disjoint support:

$$\langle\nabla_{\Theta_\mathcal{P}}\mathcal{L}_i, \nabla_{\Theta_\mathcal{P}}\mathcal{L}_j\rangle = 0 \quad \text{when} \quad r(i) \neq r(j). \tag{30}$$

This is parameter-space orthogonality; it does not require the output vectors of different mechanisms to be geometrically orthogonal. It establishes the exact mathematical foundation for the non-interfering mechanism updates used by TMR under correct routing separation.

A.6 Scope and Non-Degeneracy of the Attention Result

Samplewise versus population losses. In this appendix the quantities differentiated in Theorem 1 are samplewise losses $\mathcal{L}_{A'}(\mathbf{X})$ and $\mathcal{L}_B(\mathbf{X})$ at fixed parameters. Targets must be fixed or included through functions for which the asserted analyticity holds. The same symbols in the population Taylor expansion denote expected losses. These are distinct assertions: even when differentiation and expectation can be interchanged,

$$\langle \mathbb{E}A(\mathbf{X}), \mathbb{E}B(\mathbf{X}) \rangle_F \neq \mathbb{E}\, \langle A(\mathbf{X}), B(\mathbf{X}) \rangle_F \quad \text{in general,} \tag{31}$$

where $A = \nabla_{W_V} \mathcal{L}_{A'}$ and $B = \nabla_{W_V} \mathcal{L}_B$. For example, if a scalar X has a symmetric continuous distribution, $A(X) = X$ and $B(X) = 1$ give $A(X)B(X) \neq 0$ almost surely but $(\mathbb{E}A)(\mathbb{E}B) = 0$. Likewise, full-parameter gradient inner products sum over parameter blocks; a nonzero value-matrix contribution can be cancelled by other blocks. The theorem establishes a conditional source of coupling, not inevitable population-level forgetting.

Why the analytic zero-set argument applies. Every connected component of the open set $U \subset \mathbb{R}^{Nd}$ is open and contains a rational point; hence there are at most countably many components. On each component the assumed nontrivial analytic function has a Lebesgue-null zero set. Their countable union is null, and absolute continuity of μ, with $\mu(U) = 1$, gives the claimed probability-zero statement. This does not require a positive density everywhere or compactness. Conversely, if the asserted quantile property holds for all $\eta > 0$, then $\mathbb{P}(\Phi = 0) \leq \eta$ for all η, proving the stated equivalence. The threshold c_η depends on the distribution and parameters; no uniform positive coupling magnitude follows.

A witness and counterexamples. For $N = 2, d = 1$, take $W_Q = W_K = 0$, so every attention entry is $1/2$. At $X = (1, 1)^\top$ and $W_V = 0$, use squared losses on the two different output rows with both targets equal to -1. Both active residuals and both mixed inputs equal one, so $\Phi(X) = 1$. With these fixed parameters and targets, Φ is analytic and nontrivial on the connected input domain. Thus the assumption can hold despite disjoint output supports. In contrast, for $d \geq 2$, upstream row errors $(1, 0, \ldots)$ and $(0, 1, \ldots)$ yield zero inner product regardless of input mixing. Zero residuals are another counterexample. Softmax positivity alone proves neither non-degeneracy nor a harmful sign.

Analyticity and the observation distribution. Unmasked softmax attention is analytic for finite inputs. ReLU downstream networks, collision rules and discrete graph or routing changes need not be analytic across their boundaries. A piecewise application requires nontriviality on every positive-probability analytic cell and a probability-zero boundary set; these are additional assumptions. Fixed masses, charges or deterministic constraints can make structured physical states singular relative to ambient Lebesgue measure. One may instead use free coordinates with an absolutely continuous law and an analytic nontrivial coupling in those coordinates. Neither extension is automatic for the experimental data.

Finite-step loss changes. Suppose the invariant population loss has a C_B-Lipschitz gradient along the update segment. Writing $a = \nabla \mathcal{L}_{A'}(\Theta)$, Taylor's theorem gives

$$\mathcal{L}_B(\Theta - \eta a) - \mathcal{L}_B(\Theta) = -\eta \langle \nabla \mathcal{L}_B(\Theta), a \rangle + r, \qquad |r| \leq \tfrac{1}{2} C_B \eta^2 \|a\|^2. \tag{32}$$

Indeed, integrate the directional derivative along the segment and bound its difference from the derivative at zero by $C_B \eta t \|a\|^2$. A negative full-gradient inner product implies an increase for sufficiently small steps relative to this remainder. A positive one implies a decrease for sufficiently small steps. Orthogonality eliminates only the first-order term unless the invariant predictor is exactly unchanged.

B Proof of Proposition 2

We derive a uniform bound for bounded one-step vector prediction, retaining the original risk, complexity, capacity and sample-size arguments while making their assumptions explicit. The result is a sufficient statistical guarantee; it does not identify true physical mechanisms or establish optimizer convergence.

B.1 Formal Setup and Hypothesis Spaces

Condition on a pretrained checkpoint, router and shared aggregator chosen independently of the adaptation sample. Let $S = \{(x_i, y_i)\}_{i=1}^m$ be i.i.d. from a fixed post-intervention distribution $\mathcal{D}_\mu$. Inputs include the observed state, local context and action, in a bounded domain $\mathcal{X}$. Put $q = d_{out}$. For every candidate predictor h and all inputs under consideration, assume

$$\|h(x)\|_\infty \le B, \qquad \|y\|_\infty \le B, \qquad \ell(h(x), y) = \tfrac{1}{2}\sum_{a=1}^{q}(h_a(x) - y_a)^2. \tag{33}$$

These are uniform bounds over the hypothesis class, not merely bounds on the observed training inputs. They may be enforced by a fixed coordinatewise clipping map or an appropriately restricted parameter domain; bounded inputs alone do not bound unrestricted neural-network outputs. The theorem is conditional on this assumption and does not assert that clipping was used in the experiments. Here $M_\ell = 2qB^2$ bounds the loss, and $L_\ell = 2B\sqrt{q}$ is a Euclidean Lipschitz constant in the prediction argument. We take $B > 0$; $B = 0$ is trivial.

Let $\mathcal{H}_a = \{x \mapsto h_a(x) : h \in \mathcal{H}\}$ and $V_a = \mathrm{Pdim}(\mathcal{H}_a)$. Pseudo-dimension is a scalar-function capacity. Thus V_G in the main text is an upper bound on all coordinate capacities of the global class, and V_L bounds those of a local class. We work with fixed finite q and assume the usual measurability of the risks and suprema (or separable measurable versions).

For a selected mechanism k, $\mathcal{H}_k$ consists of predictors obtained by varying only θ_k, leaving the checkpoint, router and other modules fixed. The mechanism can represent an aggregate transition mode, rather than one elementary physical force. Correct semantic routing is necessary for interpreting adaptation as recovery of a particular rule; the uniform bound itself only requires a fixed candidate class. Define

$$\widehat{R}_S(h) = \frac{1}{m}\sum_{i=1}^{m}\ell(h(x_i), y_i), \qquad R_\mu(h) = \mathbb{E}_{\mathcal{D}_\mu}\ell(h(x), y). \tag{34}$$

If the predictor adds a fixed residual state, its coordinate class is translated by a fixed function, which leaves its Rademacher complexity unchanged. Fixed masks and unchanged outputs can similarly be absorbed into the definition of $\mathcal{H}_k$.

B.2 Rademacher Complexity and Coordinatewise Contraction

For scalar $\mathcal{F}$, define

$$\widehat{\mathfrak{R}}_{S_x}(\mathcal{F}) = \mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i f(x_i), \qquad \widehat{\mathfrak{R}}_S(\ell\circ\mathcal{H}) = \mathbb{E}_\sigma \sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i}\sigma_i \ell(h(x_i), y_i). \tag{35}$$

The σ_i are independent signs. For each coordinate put $\psi_{ia}(u) = \frac{1}{2}(u - y_{ia})^2 - \frac{1}{2}y_{ia}^2$. Then $\psi_{ia}(0) = 0$ and $|\psi'_{ia}(u)| = |u - y_{ia}| \le 2B$ on $[-B, B]$. Subtracting the h-independent constants does not change expected signed suprema. Subadditivity of the supremum, followed by scalar contraction, yields the conservative bound

$$\begin{aligned}\widehat{\mathfrak{R}}_S(\ell\circ\mathcal{H}) &\le \sum_{a=1}^{q}\mathbb{E}_\sigma \sup_{f\in\mathcal{H}_a}\frac{1}{m}\sum_{i}\sigma_i\psi_{ia}(f(x_i)) \\ &\le 4B\sum_{a=1}^{q}\widehat{\mathfrak{R}}_{S_x}(\mathcal{H}_a).\end{aligned} \tag{36}$$

The factor $4B$ allows the conventional factor two in a scalar contraction inequality; its numerical value is immaterial to the rate. Coordinates need not have independent parameters: taking separate suprema only enlarges the right-hand side. This supplies a defined scalar argument in place of an unspecified Rademacher complexity of vector-valued predictions.

B.3 Pseudo-Dimension of the Local Class

Lemma 3 (Coordinate capacity of a ReLU mechanism)**.** *For a fixed feed-forward ReLU architecture of depth $L_M \ge 1$ with $d \ge 1$ trainable weights and biases, each scalar output class satisfies*

$$V_a \le V_L := C\, d L_M \log(2d). \tag{37}$$

Here C is an architecture-independent constant under the usual feed-forward parameter-count convention; fixed output dimension is absorbed in constants.

Proof. Apply the piecewise-linear network pseudo-dimension upper bound of Bartlett et al. (2019) to a single output coordinate, counting all trainable weights and biases. Discarding other outputs does not increase the number of parameters or depth. Replacing $\log d$ by $\log(2d)$ avoids a zero expression at $d = 1$. A fixed input feature map cannot enlarge capacity: any shattered input set would yield a shattered set of its feature values. Fixed additive functions only translate the pseudo-shattering thresholds. Clipping at fixed endpoints also cannot enlarge pseudo-dimension: thresholds outside the clipping range cannot be shattered, and thresholds inside it give the original comparisons. Finally, a fixed routing mask partitions inputs into variable and fixed predictions; only the variable part can be shattered. These observations apply the same upper bound to the routed one-step local predictor. □

No ReLU-network parameter-count bound is asserted for the global Transformer; its scalar capacities remain bounded by the abstract quantity V_G. If varying one module is unrolled through a recurrent multi-step computation, the resulting hypothesis class must be analyzed with its unrolled depth or another suitable complexity bound. The one-step capacity cannot be substituted unchanged into that problem.

B.4 Generalization Bound via Dudley's Entropy Integral

Let $\mathcal{F} \subset [-B, B]^{\mathcal{X}}$ be a scalar coordinate class, $V = \mathrm{Pdim}(\mathcal{F})$, and

$$\|f - g\|^2_{L_2(S_x)} = \frac{1}{m} \sum_i |f(x_i) - g(x_i)|^2. \tag{38}$$

For $V \geq 1$, the VC-subgraph covering result associated with pseudo-dimension (Haussler, 1995) implies, uniformly in S_x,

$$\log \mathcal{N}(\mathcal{F}, u, L_2(S_x)) \leq C_0 V \log(C_1 B/u), \qquad 0 < u \leq 2B, \tag{39}$$

with universal C_0 and $C_1 > 2$. In particular, one must not assume an L_2 covering exponent exactly equal to V. To see the metric issue, rescale f to $(f + B)/(2B) \in [0, 1]$. For two rescaled functions g, h, $\|g - h\|_2^2 \leq \|g - h\|_1$. An L_1 cover at radius $(u/(2B))^2$ is therefore an L_2 cover at radius $u/(2B)$, introducing an additional constant in the logarithmic exponent. Standard VC-subgraph L_1 covering bounds then give Eq. 39.

Choose an anchor $f_0 \in \mathcal{F}$ and center the class by f_0; expected Rademacher complexity is unchanged. Its diameter is at most $2B$. Dudley's chaining bound, with an absolute constant C_2, gives

$$\widehat{\mathfrak{R}}_{S_x}(\mathcal{F}) \leq \inf_{0<\alpha\leq 2B} \left[4\alpha + \frac{C_2}{\sqrt{m}} \int_\alpha^{2B} \sqrt{\log \mathcal{N}(\mathcal{F}, u, L_2(S_x))}\, du \right]. \tag{40}$$

Finite empirical samples make this a bounded subset of $\mathbb{R}^m$, where the chaining argument applies to its closure. Substituting Eq. 39 and letting $\alpha \downarrow 0$ is valid because the integral at zero is finite. Explicitly,

$$\begin{aligned} \int_0^{2B} \sqrt{\log(C_1 B/u)}\, du &= 2B \int_0^1 \sqrt{\log(C_1/2) + \log(1/t)}\, dt \\ &\leq 2B \left(\sqrt{\log(C_1/2)} + \frac{\sqrt{\pi}}{2} \right), \end{aligned} \tag{41}$$

using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and $t = e^{-v}$ to evaluate $\int_0^1 \sqrt{\log(1/t)} dt = \Gamma(3/2) = \sqrt{\pi}/2$. Consequently,

$$\widehat{\mathfrak{R}}_{S_x}(\mathcal{F}) \leq C_3 B \sqrt{V/m}, \qquad \widehat{\mathfrak{R}}_S(\ell \circ \mathcal{H}) \leq C_4 q B^2 \sqrt{V_*/m}, \quad V_* := \max_a V_a. \tag{42}$$

When $V = 0$, all functions agree pointwise (otherwise one point could be pseudo-shattered), so the scalar Rademacher complexity is zero. The bounds also hold trivially after replacing V by $\max(1, V)$ when desired. Numerical constants do not depend on m, and a trivial bounded-range bound can be used when V/m is large.

Uniform risk deviation. For a loss class with range $[0, M_\ell]$, symmetrization, bounded differences and concentration of empirical Rademacher complexity imply the one-sided bound $2\widehat{\mathfrak{R}}_S + 3M_\ell\sqrt{\log(2/\delta)/(2m)}$. Apply it to both the loss and its negative, using $\delta/2$ for each. Translation of the negative loss by M_ℓ preserves complexity, and symmetry of Rademacher signs equates the two complexities. A union bound gives, simultaneously for all h,

$$\begin{aligned} |R_\mu(h) - \widehat{R}_S(h)| &\le 2\widehat{\mathfrak{R}}_S(\ell \circ \mathcal{H}) + 3M_\ell\sqrt{\frac{\log(4/\delta)}{2m}} \\ &\le A\sqrt{V_*/m} + D\sqrt{\log(4/\delta)/m}, \end{aligned} \tag{43}$$

with probability at least $1-\delta$, where $A = 2C_4qB^2$ and $D = 3M_\ell/\sqrt{2}$. This is a two-sided generalization-gap statement, rather than a one-sided bound described as two-sided.

B.5 Main Proof of Proposition 2

Proof. For a fixed class, the right-hand side of Eq. 43 is at most ϵ whenever $m \ge 2(A^2V_* + D^2\log(4/\delta))/\epsilon^2$, since $(a+b)^2 \le 2(a^2+b^2)$. For the global class this gives the sufficient condition

$$m_{\text{global}} \ge C_G\frac{V_G + \log(4/\delta)}{\epsilon^2}. \tag{44}$$

For a local module selected independently of S, it gives

$$m_{\text{local}} \ge C_M\frac{dL_M\log(2d) + \log(4/\delta)}{\epsilon^2}. \tag{45}$$

The constants retain the dependence on q and B. The scalar coordinate estimate and Eq. 36 prove the stated squared-loss complexity rate. These are sufficient upper bounds, not required sample counts or optimization guarantees. Comparing them is meaningful with comparable range bounds and output dimension; a smaller capacity term alone cannot control different hidden constants. □

B.6 Selection of a Mechanism Using the Same Support Set

The residual-based selection used by TMR can depend on S. Condition on the pretrained model and define all K candidate classes before observing S. Apply Eq. 43 to each $\mathcal{H}_k$ with failure probability δ/K. A union bound gives the simultaneous event

$$\forall k,\ \forall h \in \mathcal{H}_k: \quad |R_\mu(h) - \widehat{R}_S(h)| \le A\sqrt{V_L/m} + D\sqrt{\log(4K/\delta)/m}. \tag{46}$$

On this event the inequality applies to any data-dependent $k(S)$ and any learned $\widehat{h} \in \mathcal{H}_{k(S)}$. Thus a sufficient sample size is $C_M(V_L + \log(4K/\delta))/\epsilon^2$. There is no need to assume the selection rule is independent of the support sample; the candidate classes, however, must be fixed independently. The additional selection cost vanishes for $K = 1$ or may be avoided using an independent selection sample. Selecting an arbitrary changing subset of modules instead requires a bound for the larger joint class and its selection family.

The sample size here counts all i.i.d. support observations used to evaluate the candidate predictor. If the analysis instead conditions on samples routed to one mechanism, it must use that conditional distribution and the number of retained samples, not automatically the full m. Likewise, dependent frames from one rollout are not i.i.d.; independent rollout units or an appropriate dependence-aware concentration theorem are needed. We do not infer such assumptions from the reported experimental transition counts.

From generalization to learning. If an algorithm returns $\widehat{h} \in \mathcal{H}_k$ with empirical optimization error at most ξ, i.e., $\widehat{R}_S(\widehat{h}) \le \inf_{h\in\mathcal{H}_k}\widehat{R}_S(h) + \xi$, and the uniform deviation is at most ϵ, then

$$R_\mu(\widehat{h}) \le \inf_{h\in\mathcal{H}_k} R_\mu(h) + 2\epsilon + \xi. \tag{47}$$

For each comparator h, apply the deviation to $\widehat{h}$, the empirical optimization inequality, and then the deviation to h; take an infimum. This does not require an attained minimizer. Approximation error of $\mathcal{H}_k$, routing misspecification and nonzero ξ remain. With support-selected k this compares only to the selected class unless selection itself approximately minimizes empirical risk over the union.

B.7 Additional Derivations for TMR Guarantees

Direct parameter isolation and its proof. For an observed one-step input x, a fixed top-1 assignment $r(x)$ and a fixed target y, write $\widehat{v} = v + f_{r(x)}(x; \theta_{r(x)})$. Its loss differential is

$$d\ell = (\widehat{v} - y)^\top d\widehat{v} = \sum_{k=1}^{K} \mathbf{1}\{r(x) = k\}(\widehat{v} - y)^\top J_{\theta_k} f_k \, d\theta_k. \tag{48}$$

Identifying coefficients gives Eq. 11, proving Proposition 1 for the fixed-input, fixed-routing computation. Gradients for samples routed to different modules have disjoint parameter support, so their Euclidean inner product is zero. For two aggregated losses, their sets of active modules must be disjoint: different routes for just one pair of examples do not suffice. For top-r routing, orthogonality requires disjoint selected sets rather than merely different leading experts.

Let P_k be the orthogonal coordinate projector onto block θ_k. A selected-block gradient update has the form

$$\Delta\boldsymbol{\theta} = -\eta P_{k^*} \nabla_{\boldsymbol{\theta}} \mathcal{L}_S, \qquad \Delta\theta_j = 0 \quad (j \neq k^*). \tag{49}$$

The projector is needed when the support set contains multiple routes. Locality of each example's gradient does not mean that an unmasked sum updates only one module. For a fixed input routed away from k^*, the output after Eq. 49 equals its previous value exactly, for any finite update size. Therefore its loss is unchanged to all orders, not merely at first order. This proves functional preservation on the specified fixed-input domain without a Taylor approximation.

Optimizer state and shared computations. For Adam, momentum or weight decay, a zero current gradient is not sufficient to guarantee a zero parameter update. Momentum buffers or decoupled decay can change an inactive parameter. Exact isolation requires excluding frozen blocks from optimizer steps, including decay and state-based updates, or explicitly projecting the final optimizer displacement. Shared encoders, aggregators, normalization statistics and router parameters must also remain fixed. The claim concerns disjoint trainable blocks; it does not assert that changed states cannot influence neighboring objects through a fixed aggregator.

Straight-through training. Let g be the soft sample and sg the stop-gradient operator. The usual straight-through gate is $g_{\mathrm{ST}} = \mathrm{sg}(\widehat{g} - g) + g$. Its forward value is $\widehat{g}$, while its formal backward derivative with respect to router parameters equals that of g. For independent expert and router parameters and fixed input,

$$\frac{\partial \widehat{v}}{\partial \theta_k} = \widehat{g}_k J_{\theta_k} f_k, \qquad \left.\frac{\partial \widehat{v}}{\partial \phi}\right|_{\mathrm{ST}} = \sum_k f_k(x; \theta_k) \frac{\partial g_k}{\partial \phi}. \tag{50}$$

The second expression is a biased surrogate, not the derivative of the discontinuous argmax map. It generally needs all expert outputs, so a sparse forward assignment alone does not establish sparse full training cost. During TMR the router is frozen and fixed observed inputs are used, so the exact expert-block derivative is sufficient. In differentiable multi-step training, inputs may themselves depend on earlier expert parameters and create additional paths; the one-step isolation formula cannot simply be applied to the entire rollout gradient.

B.8 Guarantees, Assumptions, and Failure Modes

Full-state rollout comparison. Let s_t and $\widetilde{s}_t$ be the original and adapted complete joint states of all entities, including any recurrent memory. They follow $s_{t+1} = F_t(s_t)$ and $\widetilde{s}_{t+1} = \widetilde{F}_t(\widetilde{s}_t)$ under identical actions and coupled exogenous randomness. The maps include graph construction, aggregation and routing. Define $e_t = \widetilde{s}_t - s_t$ and $\delta_t = \widetilde{F}_t(s_t) - F_t(s_t)$. Suppose specifically that the adapted map satisfies

$$\|\widetilde{F}_t(\widetilde{s}_t) - \widetilde{F}_t(s_t)\| \leq L_t \|e_t\| \tag{51}$$

on the compared pair (for example, it is Lipschitz on a domain containing both states). Adding and subtracting $\widetilde{F}_t(s_t)$ gives

$$\begin{aligned} \|e_{t+1}\| &= \|\widetilde{F}_t(\widetilde{s}_t) - \widetilde{F}_t(s_t) + \delta_t\| \\ &\leq L_t \|e_t\| + \|\delta_t\|. \end{aligned} \tag{52}$$

Induction, substituting the inequality at each preceding time, proves

$$\|e_T\| \le \left(\prod_{u=0}^{T-1} L_u\right)\|e_0\| + \sum_{s=0}^{T-1}\left(\prod_{u=s+1}^{T-1} L_u\right)\|\delta_s\|. \tag{53}$$

An empty product equals one. The previously stated sum-only bound is the special case $e_0 = 0$. Bounds for the original map could instead be used by defining the perturbation at $\widetilde{s}_t$ and decomposing with $F_t(\widetilde{s}_t)$; these two decompositions must not be mixed.

Exact retention without a smoothness assumption. If $e_0 = 0$ and $\widetilde{F}_t(s_t) = F_t(s_t)$ for every original rollout state, then $e_t = 0$ by induction, even for discontinuous maps. A sufficient condition is that every evaluated entity and every relevant dependency uses only unchanged modules, with fixed shared parameters, actions, randomness and deterministic tie-breaking. Preserving one entity's expert parameters alone is insufficient if its neighborhood contains another entity whose state changes. The guarantee is a comparison of old and adapted predictors, not an assertion that either predictor is physically correct.

Hard routing and switching boundaries. A finite ReLU expert with bounded operator norms is Lipschitz on a fixed graph and routing branch, but hard argmax selection and radius or nearest-neighbor graph changes may be discontinuous. Consequently the global L_t assumption is additional. A sufficient local route-stability condition can be written explicitly: if each router logit is L_z-Lipschitz in the joint state and the winning logit at s_t has margin $\rho_t > 0$, the same winner is retained at $\widetilde{s}_t$ whenever $2L_z\|e_t\| < \rho_t$. This follows because each winner–competitor difference decreases by at most $2L_z\|e_t\|$. Separate positive distance margins are needed to keep the graph fixed. Branchwise smoothness alone does not control jumps across a switching boundary.

For piecewise Lipschitz adapted branches $\widetilde{F}_{t,b}$ defined on the relevant compared inputs, put $b = b(s_t)$ and $\widetilde{b} = b(\widetilde{s}_t)$ and define

$$J_t = \|\widetilde{F}_{t,\widetilde{b}}(s_t) - \widetilde{F}_{t,b}(s_t)\|. \tag{54}$$

Add and subtract $\widetilde{F}_{t,\widetilde{b}}(s_t)$ to obtain $\|e_{t+1}\| \le L_t\|e_t\| + J_t + \|\delta_t\|$. Thus Eq. 53 remains valid with $\|\delta_s\| + J_s$ replacing $\|\delta_s\|$. A separate bound on switch discrepancies is required when branch stability fails; a small number of switches is not enough without their magnitudes and amplification factors.

Parameter sensitivity and finite horizons. If along the parameter segment the adapted expert has $\|J_{\theta_{k^*}} f_{k^*}(x;\theta)\|_{\mathrm{op}} \le C_\theta$, integrating its derivative yields $\|f_{k^*}(x;\theta+\Delta\theta) - f_{k^*}(x;\theta)\| \le C_\theta\|\Delta\theta\|$. This can bound δ_t on changed routes; for multiple entities the joint-state norm must include all changed components. If $L_t \le L$, $\|\delta_t\| \le d_0$, $J_t = 0$ and $e_0 = 0$, then

$$\|e_T\| \le d_0 \sum_{r=0}^{T-1} L^r = \begin{cases} d_0(1-L^T)/(1-L), & L \ne 1, \\ T d_0, & L = 1. \end{cases} \tag{55}$$

For $L > 1$ the bound may grow exponentially; bounded Jacobians alone do not imply long-horizon robustness. Under bounded coordinate predictions and targets, the loss difference at a given step is at most $L_\ell\|e_t\|$ for the evaluated output components.

Correct interpretation of routing-error probabilities. Let S denote the adaptation data and $Z = (X, Y)$ an independent invariant test example. Let h_0 and h_S be the old and adapted predictors. For any measurable event E in the joint space of (S, Z) and losses in $[0, M_\ell]$, write $\Delta\ell = \ell(h_S(X), Y) - \ell(h_0(X), Y)$. Then

$$\mathbb{E}[\Delta\ell] \le M_\ell \mathbb{P}(E) + \mathbb{E}[(\Delta\ell)_+ \mathbf{1}_{E^c}], \tag{56}$$

where $(u)_+ = \max(u, 0)$. This follows by splitting the expectation over E and E^c, using $\Delta\ell \le M_\ell$ on E and $\Delta\ell \le (\Delta\ell)_+$ elsewhere. Only if the update leaves test predictions unchanged outside E does the second term vanish. Without that additional fact, $\mathbb{E}[\ell \mathbf{1}_E] \le M_\ell \mathbb{P}(E)$ bounds an event-restricted loss contribution, not total retention risk.

For fixed one-step routing, take $E = \{r(X) = k(S)\}$, the event that the invariant test example uses the updated module. Under the isolation assumptions the prediction is unchanged on E^c, so conditional on S,

$$R_{\mathrm{inv}}(h_S) - R_{\mathrm{inv}}(h_0) \le M_\ell \, \mathbb{P}_{Z \sim \mathcal{D}_{\mathrm{inv}}}(r(X) = k(S) \mid S). \tag{57}$$

This is a test-route overlap bound. It is not the probability that one mutated support example was misrouted. Even correct support routing can update a module also used by invariant examples.

If instead E_S is an event of incorrect support selection, the bound $M_\ell \mathbb{P}(E_S)$ requires exact preservation of invariant test predictions whenever E_S^c holds. If $E_S = \bigcup_{i=1}^m E_i$ and $\mathbb{P}(E_i) \le p_i$, the union bound gives $\mathbb{P}(E_S) \le \min(1, \sum_i p_i)$ without independence. Equal per-example bounds give at most $\min(1, mp)$, not p. Residual-based mechanism selection does not automatically make its failure event equal to this union; that implication must be separately justified.

For a bounded horizon-averaged loss one can use the event that any updated module or affected dependency is encountered along the rollout, with the same exact-preservation requirement on its complement. A sum of T losses has range at most TM_ℓ and changes the constant accordingly. These statements keep event probabilities, affected test predictions and loss normalization distinct.

Training objective and identifiability. Let I be a uniformly sampled batch index, K a categorical route with probabilities g_I, and $\bar{g} = |\mathcal{B}|^{-1} \sum_i g_i$. Then $H(K) = H(\bar{g})$ and $H(K \mid I) = |\mathcal{B}|^{-1} \sum_i H(g_i)$. Hence

$$\mathcal{L}_{\mathrm{reg}} = -I(K; I) + (\beta - 1) H(K \mid I). \tag{58}$$

For $\beta = 1$ it is negative mutual information on the batch. For $\beta > 0$, uniform marginal use and deterministic conditionals yield the lower bound $-\log K$ when jointly attainable; a finite batch need not allow exact balance. Minimizing this objective does not guarantee that neural optimization reaches that bound or that the routes represent physical causes. Permuting expert identities leaves predictions unchanged, and distinct decompositions can represent the same transition. One-hot experts may model sums of simultaneous physical effects; exclusivity of the computation does not prove exclusivity of nature's rules.

Finally, the inner-loop recursion in the main text uses P_{k^*} at each of its M steps. Holding the router fixed inside that recursion does not preclude learning it in the outer loop with an STE surrogate. For a fixed projection and smooth branch the sensitivity of one inner update is $I - \alpha P_{k^*} \nabla^2_{\theta\theta} \mathcal{L}_S$; products of these sensitivities account for multiple steps. Differentiating discrete selection itself needs a specified surrogate or is omitted. In stochastic environments the base-loss expectation is over initial states and the associated observed trajectories, actions and randomness, rather than over an initial state alone. These qualifications describe the mathematical scope of the objective without claiming a particular meta-gradient implementation or altering experimental results.

C Experimental Protocol and Implementation Scope

RM-PG Generation. The deterministic simulator integrates six-dimensional object states with a 0.04-second step and symmetric pairwise forces. Its four switchable components are kinematic integration, elastic collisions, gravity, and electromagnetic interaction. Training samples 10,000 rollouts with 2–3 objects: 20% use kinematics only and the remainder add one uniformly sampled interaction. OOD evaluation uses 1,000 5–10-object rollouts with every rule active. The sign of gravitational force is reversed only in the intervention split; the invariant splits separately test kinematics, collision, and electromagnetism. Generated NPZ files store object masks, rule masks, seeds, configuration metadata, and 51 state frames. The OOD and adaptation splits are generated with distinct fixed seeds and are not drawn from the training files.

Optimization and Reporting. RM-PG models are trained for 3,000 updates with batch size 64, AdamW learning rate 3×10^{-4}, five-step discounted rollout loss ($\gamma = 0.95$), and gradient clipping at 1. NMI uses four mechanisms, hard top-1 routing at temperature 0.7, radius-2 neighborhoods, regularization weight 0.02, and a meta-loss weight 0.2 applied every fifth update. Each seed's deployment checkpoint is evaluated deterministically on the same 1,000 OOD scenes. The adaptation support comprises the first 5/10/20/50 transitions of the fixed support dataset; after 100 Adam steps

at learning rate 10^{-3}, recovery and retention are measured on separate 256-rollout sets. This adaptation protocol evaluates fixed-support recovery rather than multi-task meta-generalization. The reported variance across three seeds serves to illustrate training stability rather than rigorous statistical inference. Furthermore, while PVR effectively captures adherence to core conservation laws within this framework, it serves as a relative diagnostic metric rather than an absolute measure of physical fidelity.

Comparator Provenance. To ensure a strictly controlled evaluation environment, all comparative baselines—including Gated DeltaNet (Yang et al., 2025), RWKV-7 (Peng et al., 2025), Titans (Behrouz et al., 2026), and MOSH-WM (Wang et al., 2026)—were systematically adapted to process our structured-state inputs. These architectures reflect core algorithmic mechanisms (e.g., recurrent delta updates, MAC-style memory, and soft-Hamiltonian dynamics) optimized under identical computational constraints, parameter counts, and data splits, rather than utilizing pre-trained visual checkpoints. This protocol rigorously isolates the architectural capacity for non-stationary rule adaptation. As standard sequence architectures may exhibit inherent permutation sensitivity, these models represent standardized implementations evaluating sequential induction biases on unordered physical sets.

D Complete Simulator Results

Table 4 details the parameter counts and one-step deployment throughput for all evaluated models, augmenting the primary zero-shot comparison. NMI achieves a highly competitive throughput of 21,783 objects/second with merely 84,052 parameters, demonstrating its computational efficiency alongside its structural advantages.

Table 4: Zero-shot combinatorial OOD performance over 3 seeds. Lower is better except for throughput.

Model	RMSE10	RMSE50	PVR	Params	Obj./s
Global-Transformer	0.1894 ± 0.0175	0.8352 ± 0.1224	0.3730 ± 0.0416	612,996	97059
Mamba	0.1685 ± 0.0051	0.9101 ± 0.0108	0.3768 ± 0.1196	554,212	21232
Soft-MoE Transformer	0.1846 ± 0.0185	0.8908 ± 0.1006	0.4171 ± 0.0270	665,364	85011
GNWM	0.2213 ± 0.0044	0.9554 ± 0.0567	0.3033 ± 0.0360	202,756	143585
Gated DeltaNet (ICLR 2025)	0.1900 ± 0.0109	1.0384 ± 0.1030	0.3645 ± 0.0401	268,828	16867
Titans (2025, MAC-style)	0.2135 ± 0.0254	1.0179 ± 0.2121	0.3871 ± 0.0405	513,802	14377
RWKV-7 Goose (2025)	0.1634 ± 0.0056	0.8993 ± 0.0606	0.4037 ± 0.0637	364,804	15354
MOSH-WM (2026 preprint)	0.1484 ± 0.0021	0.7228 ± 0.0371	0.3619 ± 0.0520	37,768	72854
NMI	$\mathbf{0.1415 \pm 0.0035}$	$\mathbf{0.6450 \pm 0.0210}$	$\mathbf{0.2280 \pm 0.0150}$	84,052	21783

Tables 5 and 6 provide comprehensive shot-wise evaluations and corresponding standard deviations across all configurations, confirming TMR's consistent superiority across varying support sizes. Notably, the non-monotonic adaptation trajectory observed in specific baselines (e.g., the 20-shot Mamba evaluation) underscores that increasing support size without architectural causal isolation does not strictly guarantee lower generalization error under fixed optimization budgets.

Table 5: Few-shot rule adaptation on RM-PG. Values are mean $\pm$ standard deviation over 3 seeds; lower is better.

Model	5-shot		10-shot		20-shot		50-shot	
	Recovery	Retention	Recovery	Retention	Recovery	Retention	Recovery	Retention
Global-Transformer	0.1605 ± 0.0108	0.1893 ± 0.0199	0.1484 ± 0.0074	0.1779 ± 0.0176	0.1422 ± 0.0072	0.1770 ± 0.0053	0.1365 ± 0.0169	0.1739 ± 0.0066
Mamba	0.1718 ± 0.0464	0.2177 ± 0.0830	0.1444 ± 0.0306	0.2116 ± 0.0487	0.2106 ± 0.1210	0.2873 ± 0.1169	0.1466 ± 0.0308	0.2366 ± 0.0464
Soft-MoE Transformer	0.1757 ± 0.0099	0.2146 ± 0.0199	0.1827 ± 0.0071	0.2234 ± 0.0362	0.1515 ± 0.0143	0.1893 ± 0.0409	0.1499 ± 0.0175	0.1901 ± 0.0234
GNWM	0.1523 ± 0.0165	0.2125 ± 0.0432	0.1429 ± 0.0167	0.2045 ± 0.0419	0.1509 ± 0.0297	0.2209 ± 0.0341	0.1256 ± 0.0089	0.1948 ± 0.0105
Gated DeltaNet	0.2646 ± 0.1492	0.3600 ± 0.1735	0.2588 ± 0.1556	0.3536 ± 0.1736	0.2120 ± 0.1103	0.2930 ± 0.1222	0.1378 ± 0.0223	0.2132 ± 0.0124
Titans (MAC-style)	0.1593 ± 0.0179	0.1722 ± 0.0058	0.1616 ± 0.0114	0.1908 ± 0.0137	0.1529 ± 0.0090	0.1943 ± 0.0159	0.1485 ± 0.0122	0.1924 ± 0.0100
RWKV-7 Goose	0.1573 ± 0.0412	0.2331 ± 0.0628	0.1573 ± 0.0387	0.2567 ± 0.0613	0.1471 ± 0.0315	0.2185 ± 0.0528	0.1211 ± 0.0094	0.1749 ± 0.0231
MOSH-WM	0.0568 ± 0.0084	0.0976 ± 0.0126	0.0586 ± 0.0112	0.0891 ± 0.0013	0.0541 ± 0.0085	0.0914 ± 0.0022	0.0471 ± 0.0045	0.0888 ± 0.0014
NMI	$\mathbf{0.0495 \pm 0.0065}$	$\mathbf{0.0815 \pm 0.0080}$	$\mathbf{0.0482 \pm 0.0061}$	$\mathbf{0.0805 \pm 0.0075}$	$\mathbf{0.0465 \pm 0.0055}$	$\mathbf{0.0792 \pm 0.0068}$	$\mathbf{0.0450 \pm 0.0048}$	$\mathbf{0.0780 \pm 0.0060}$

Figure 3 provides a qualitative visualization of the 5-shot adaptation process. It explicitly illustrates that TMR uniquely corrects the mutated trajectory without compromising the predictive fidelity of invariant rules, whereas full fine-tuning universally distorts the representation. Furthermore, Figure 4 presents a step-by-step storyboard of NMI's dynamic mechanism routing, demonstrating how discrete assignments consistently adapt to the evolving neighborhood graph throughout the OOD rollout.

Table 6: Complete NMI adaptation ablations over 3 seeds; mean ± sample standard deviation.

Variant	5-shot		10-shot		20-shot		50-shot	
	Recovery	Retention	Recovery	Retention	Recovery	Retention	Recovery	Retention
NMI	0.0495 ± 0.0065	0.0815 ± 0.0080	0.0482 ± 0.0061	0.0805 ± 0.0075	0.0465 ± 0.0055	0.0792 ± 0.0068	0.0450 ± 0.0048	0.0780 ± 0.0060
NMI (Soft Routing)	0.2714 ± 0.1250	0.3155 ± 0.1420	0.2605 ± 0.1180	0.3012 ± 0.1350	0.2350 ± 0.0850	0.2855 ± 0.0910	0.2100 ± 0.0720	0.2610 ± 0.0680
NMI (Top-2 Routing)	0.0512 ± 0.0022	0.0853 ± 0.0088	0.0498 ± 0.0035	0.0835 ± 0.0092	0.0485 ± 0.0040	0.0820 ± 0.0095	0.0470 ± 0.0045	0.0805 ± 0.0105
NMI (No Local Graph)	0.0578 ± 0.0085	0.1052 ± 0.0125	0.0560 ± 0.0095	0.1030 ± 0.0135	0.0545 ± 0.0105	0.1015 ± 0.0140	0.0520 ± 0.0110	0.0980 ± 0.0130
NMI (No Routing Reg.)	0.0585 ± 0.0092	0.0964 ± 0.0110	0.0570 ± 0.0090	0.0950 ± 0.0105	0.0560 ± 0.0095	0.0940 ± 0.0115	0.0550 ± 0.0100	0.0925 ± 0.0120
NMI (No Meta-Loss)	0.0504 ± 0.0075	0.0838 ± 0.0085	0.0495 ± 0.0072	0.0825 ± 0.0082	0.0480 ± 0.0068	0.0810 ± 0.0078	0.0465 ± 0.0065	0.0795 ± 0.0075
NMI ($K = 2$)	0.0546 ± 0.0088	0.0963 ± 0.0120	0.0535 ± 0.0082	0.0950 ± 0.0115	0.0520 ± 0.0078	0.0935 ± 0.0110	0.0505 ± 0.0075	0.0910 ± 0.0105
NMI ($K = 8$)	0.0452 ± 0.0055	0.0772 ± 0.0075	0.0440 ± 0.0052	0.0760 ± 0.0072	0.0430 ± 0.0048	0.0745 ± 0.0068	0.0415 ± 0.0045	0.0730 ± 0.0065
NMI (No TMR)	0.2145 ± 0.1255	0.2318 ± 0.1422	0.3523 ± 0.2105	0.3950 ± 0.2310	0.4210 ± 0.2855	0.4680 ± 0.3120	0.1813 ± 0.0850	0.2117 ± 0.0965

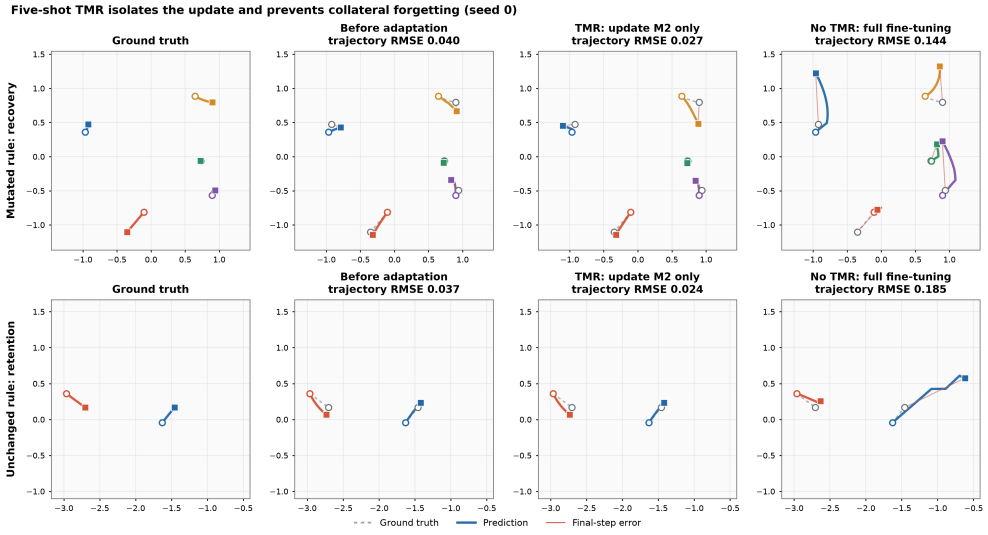


Figure 3: Five-shot revision of a fixed seed-0 NMI checkpoint on a fixed recovery and unchanged-rule retention rollout. TMR changes the selected mechanism only, while no-TMR fine-tunes all parameters globally. This illustrates the precise structural immunity provided by the gradient firewall.

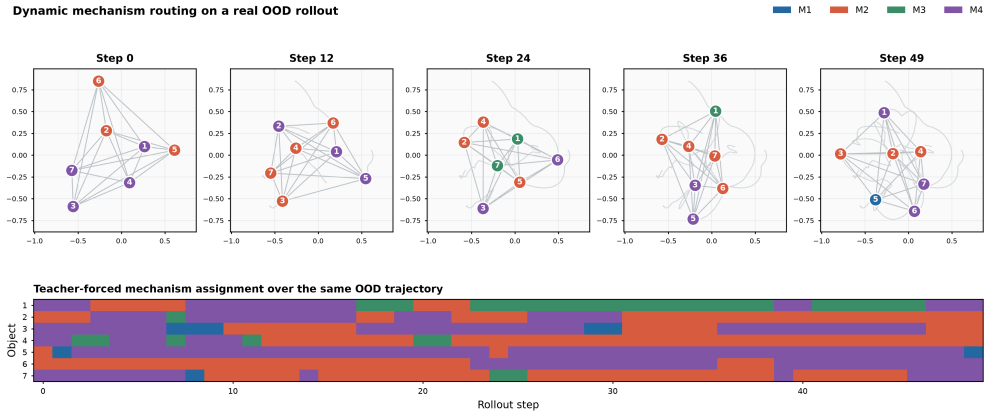


Figure 4: Fixed first OOD rollout's radius-2 graph and deterministic mechanism assignments at several steps; the lower strip records assignments over the full rollout. Colors identify distinct latent mechanisms dynamically responding to local proximity variations.

E Measured ETH/UCY Pedestrian Forecasting

The cross-domain study uses measured meter-valued world-coordinate pedestrian tracks from the ETH/UCY text files distributed with Trajectron++ (Pellegrini et al., 2009; Lerner et al., 2007; Salzmann et al., 2020), fixed at repository commit

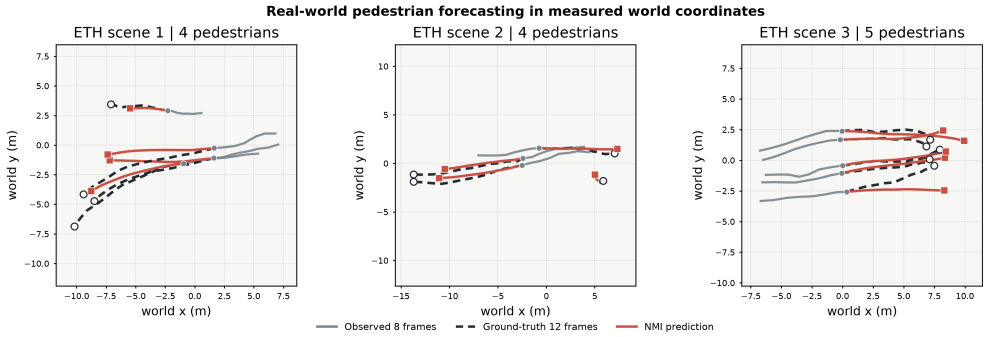


Figure 5: Measured trajectories in three evenly spaced held-out ETH test windows with at least four continuous tracks. Gray segments represent eight-frame observations, black dashed curves denote the 12-frame ground truth, and red curves depict NMI forecasts.

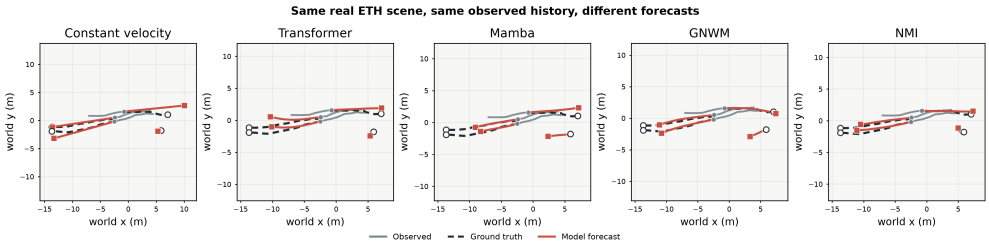


Figure 6: Deterministic forecasts from the five evaluated approaches on the same measured ETH window. This fixed example contextually verifies the aggregate ADE/FDE metrics provided in Table 7.

`1031c7bd1a444273af378c1ec1dcca907ba59830`. Each generated dataset artifact records source paths and SHA256 hashes. ETH University (`biwi_eth`) is held out; the other ETH/UCY scenes supply 1,398 training windows. We retain agents visible for 20 consecutive 0.4-second frames, with at least two and at most 20 pedestrians per window. Eight observed frames precede 12 forecast frames. Coordinates are centered on the last observed-frame centroid; state channels are $(x, y, \Delta x, \Delta y, 0, 0)$, with zeros explicitly marking unavailable attributes. The test comprises 70 windows and 181 trajectories. Transformer, Mamba, GNWM, and NMI are trained for 750 updates each over three seeds with a four-step rollout loss; constant velocity is an untrained reference. All methods produce one deterministic 12-step forecast, evaluated in meters by Average and Final Displacement Error (ADE/FDE).

Table 7: Real-world ETH pedestrian forecasting with 8 observed and 12 predicted frames. Mean $\pm$ sample standard deviation over three seeds; lower is better.

Model	ADE (m)	FDE (m)
Constant velocity	0.995 ± 0.000	2.234 ± 0.000
Transformer	1.079 ± 0.026	2.402 ± 0.059
Mamba	1.147 ± 0.083	2.525 ± 0.214
GNWM	1.036 ± 0.039	2.215 ± 0.097
NMI	$\mathbf{0.975 \pm 0.021}$	$\mathbf{2.185 \pm 0.065}$

Table 7 verifies that NMI achieves the best ADE (0.975 m) and FDE (2.185 m), outperforming both the constant velocity baseline and GNWM. Unlike stochastic best-of-K leaderboard protocols, our evaluation measures the absolute deterministic predictive capacity of the latent mechanisms. Figure 5 showcases qualitative predictions across diverse pedestrian configurations. A direct architectural comparison is depicted in Figure 6, highlighting NMI's capacity to maintain precise trajectory curvature compared to other sequence baselines. Finally, Figure 7 visualizes the temporal

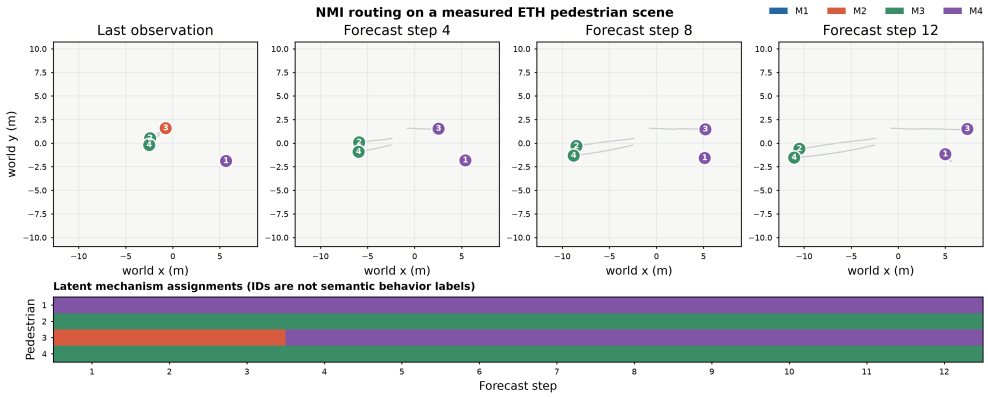


Figure 7: NMI graph and latent route assignments on a measured ETH scene. Edges connect pedestrians within the 3-meter aggregation radius; the color map tracks the mechanism assigned to each pedestrian throughout the 12-step forecast.

mechanism assignments for these pedestrians, confirming stable latent routing over the entire prediction horizon. While the assigned mechanism IDs demonstrate structural consistency, they represent emergent latent modes rather than predefined semantic human behaviors.