FUZZY-WORLD: INTERPRETABLE WORLD MODELS VIA GROUNDED IF-THEN DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

World models support dynamical prediction by learning state transitions. However, aggregate prediction accuracy neither explains the basis of local state changes nor establishes where the learned relationships apply. Here, we introduce Fuzzy-World, an intrinsically interpretable world model combining named physical states with sparse fuzzy rules. Explicit local increment equations and Top-2 activation decompose each predicted physical increment into the weighted contributions of at most two executed rules, directly linking physical semantics, rule explanations, and transition computations. Rules are fitted using inference-time activation weights, with shared dynamics and support-dependent residual shrinkage controlling local corrections. An optional residual flow supplements observation information while keeping physical rules fixed during subsequent learning. Experiments on four dynamical systems show that, among eight dynamics models with observation reconstruction, Fuzzy-World achieves the lowest mean out-of-distribution 50-step observation prediction error and the highest mean physical-trajectory R^2 on all four systems. Complete rule audits characterize empirical validity within actual activation regions through sample coverage, effective sample size, and local prediction accuracy. These results demonstrate competitive dynamical prediction with sparse, inspectable local rules.

1 INTRODUCTION

World models learn action-conditioned dynamics for multistep prediction and planning (Hafner et al., 2019). In physical systems, an interpretable model should also expose which variables drive a predicted change and where the learned relationships are supported by data. Aggregate rollout accuracy alone cannot establish the reliability of an individual local equation, particularly under distribution shift.

Semantic representations and structured models address complementary parts of this problem. Concept bottlenecks organize predictions through named variables (Koh et al., 2020); physically interpretable world models align representations with physical meaning (Peper et al., 2025); and causal dynamics models encode dependencies among states and actions (Wang et al., 2022). Self-explaining networks additionally incorporate faithfulness and stability into model design (Alvarez Melis & Jaakkola, 2018). Yet meaningful coordinates do not by themselves make a complex transition mapping inspectable. The remaining requirement is to connect semantic states to explicit local computations and assess their empirical validity.

Takagi-Sugeno (TSK) models represent nonlinear dynamics as activation-weighted mixtures of local functional consequents (Takagi & Sugeno, 1985). This structure provides readable local equations, but standard dense TSK inference combines all K rules for every prediction. Consequently, the number of terms that must be inspected for a single transition grows with the rule-library size, even when most rules receive only small activation weights. A readable rule library therefore does not necessarily provide a concise explanation of an executed transition. Hard local selection reduces the prediction to one rule, but removes graded cooperation between overlapping local regions.

This exposes a gap between fuzzy local modeling and transition-level auditability: a model should retain a sufficiently expressive rule library while bounding the number of local equations involved in

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.

each prediction. It should also distinguish the sparsity of one executed transition from the empirical validity of the complete rule library. In particular, aggregate prediction accuracy cannot determine whether an individual consequent is supported by data or accurately describes the region in which it is activated (Yen et al., 1998; Johansen et al., 2000).

Fuzzy-World addresses this gap by decoupling rule-library capacity from per-transition explanation complexity. The complete library contains K local dynamics rules, while Top-2 routing restricts each predicted physical increment to the normalized contributions of at most two executed rules. For state p_t and action a_t, let $\mathcal{S}(p_t)$ denote the selected rule indices. The increment is

$$\Delta\hat{p}_t = \sum_{k\in\mathcal{S}(p_t)} w_k(p_t) f_k(p_t, a_t), \qquad |\mathcal{S}(p_t)| \le 2, \qquad \sum_{k\in\mathcal{S}(p_t)} w_k(p_t) = 1, \tag{1}$$

where w_k is the normalized executed weight and f_k maps the current state and action to the increment proposed by rule k. Unlike a post-hoc explanation, the selected equations and weights constitute the transition computation itself. Local equations are fitted using inference-time activation weights, with shared dynamics and support-dependent residual shrinkage stabilizing rule-specific corrections; subsequent observation learning leaves the induced physical rules fixed. Rule validity is then evaluated within actual activation regions using coverage, effective sample support, and local prediction accuracy. The main contributions are:

- **An intrinsically interpretable world model over named physical states.** Fuzzy-World is proposed as an intrinsically interpretable world model over named physical states. It represents action-conditioned state increments through explicit local equations, stabilizes rule-specific corrections through shared dynamics and support-dependent residual shrinkage, and leaves the induced physical rules unchanged during subsequent observation learning.
- **Sparse executable fuzzy routing.** Fuzzy-World decouples rule-library capacity from per-transition explanation complexity: the complete library retains K local rules, whereas each physical-state prediction is computed from at most two normalized rule contributions. This makes the executed rule identities, weights, local increments, and resulting transition directly auditable while retaining graded cooperation between local regions.
- **Joint evaluation of predictive performance and rule validity.** Fuzzy-World is evaluated on four dynamical systems under in-distribution and out-of-distribution conditions. Comparisons with hard local selection and dense TSK aggregation examine the trade-off among local cooperation, predictive accuracy, and execution sparsity, while complete-library audits quantify activation coverage, effective sample support, and local predictive accuracy.

2 Related Work

Intrinsic interpretability and semantic representations. Intrinsically interpretable models expose understandable predictive structure (Rudin, 2019); self-explaining networks also constrain explanation faithfulness and stability. Disentanglement separates latent factors (Higgins et al., 2017), but identification requires inductive biases (Locatello et al., 2019). Beyond concept bottlenecks, concept whitening and embeddings connect predictions to named concepts through axis alignment and richer representations (Chen et al., 2020; Espinosa Zarlenga et al., 2022). Physically interpretable world models further emphasize physical semantics and verifiability. These designs organize meaningful states without necessarily exposing their transition equations. Fuzzy-World connects named coordinates to executable local equations. Complete-library audits assess their support and local predictive accuracy, while individual prediction decompositions expose the contributions of the selected rules.

Fuzzy rules and local system identification. Takagi-Sugeno models combine local functional consequents, and ANFIS learns their premise and consequent parameters (Jang, 1993). Yet conventional grid partitioning forms rules from combinations of input memberships, causing the rule library to grow multiplicatively with input dimensionality (Yong et al., 2026). Deep spatio-temporal predictors and evolving TS systems extend forecasting or update rule structures online (Su et al., 2025; Angelov & Filev, 2004); KANFIS uses additive decomposition and sparse masking to reduce global rule complexity. These advances do not ensure sparse transition explanations, because dense TSK inference can still assign nonzero weight to every retained rule. Aggregate accuracy also need not imply accurate local submodels, so readable coefficients require local support and error checks.

Fuzzy-World instead restricts each transition to at most two normalized contributions and audits the complete rule library within its actual activation regions.

Structured dynamics and world models. Latent world models support prediction and decision-making (Hafner et al., 2025), while object and graph representations organize components and interactions (Battaglia et al., 2016; Sanchez-Gonzalez et al., 2020; Kipf et al., 2020). Sparse identification and equation-learning networks extract dynamical expressions (Brunton et al., 2016; Sahoo et al., 2018), including jointly learned coordinates and equations (Champion et al., 2019). Such coordinates need not have predetermined physical names. Koopman and Hamiltonian models constrain evolution through linear representations and physical structure (Lusch et al., 2018; Greydanus et al., 2019); neurosymbolic programming combines neural learning with symbolic programs (Chaudhuri et al., 2021). Causal representation learning addresses dependencies and generalization (Schölkopf et al., 2021); OOCDM shares causal structure and parameters across objects of the same class (Yu et al., 2024). Links between robust decisions and causal models hold under specific assumptions (Richens & Everitt, 2024). These approaches expose interactions, equations, or structural constraints at different levels. Fuzzy-World instead emphasizes sparse local equations in named coordinates and their empirical validity under actual activation. Its rule audits assess predictive support and accuracy, without establishing causal identification or intervention validity.

3 METHOD

Given transitions $\mathcal{D} = \{(o_t, a_t, o_{t+1})\}$, Fuzzy-World predicts state evolution through explicit local increment equations. Here $o_t \in \mathbb{R}^M$ is the observation, $a_t \in \mathbb{R}^{d_a}$ is the action, and t indexes time. A known or supervised concept map G produces the named physical state $p_t = G(o_t) \in \mathbb{R}^P$, where M, d_a, and P are the observation, action, and physical-state dimensions. The selected state is assumed to be approximately Markov. Training-set statistics standardize the computational coordinates while preserving their names and physical-unit mappings. Figure 1 follows the prediction from observations to named states, sparse local increments, and observation reconstruction. Its Pendulum example illustrates how the concept map recovers angle and angular velocity; the lower branch supplies an optional recurrent residual representation for observations not fully specified by the named state.

3.1 STATE-GROUNDED FUZZY ROUTING

Each rule $k \in \{1, \dots, K\}$ has a Gaussian antecedent with center $c_k \in \mathbb{R}^P$ and positive coordinate-wise widths $\sigma_k \in (\mathbb{R}_{>0})^P$. Its score ℓ_k and dense normalized activation $\bar{w}_k$ are

$$\ell_k(p_t) = -\frac{1}{2P} \sum_{d=1}^{P} \left(\frac{p_{t,d} - c_{k,d}}{\sigma_{k,d}} \right)^2, \qquad \bar{w}_k(p_t) = \frac{\exp(\ell_k(p_t)/T)}{\sum_{j=1}^{K} \exp(\ell_j(p_t)/T)}, \tag{2}$$

where d indexes physical coordinates, j is a rule summation index, and the temperature $T > 0$ controls activation concentration. Routing depends only on the physical state; actions enter the consequent equations. Let $\mathcal{S}(p_t) = \mathrm{Top2}(\bar{w}(p_t))$ contain the indices of the two largest dense activations. The sparse executed weights are

$$w_k(p_t) = \frac{\mathbf{1}[k \in \mathcal{S}(p_t)]\bar{w}_k(p_t)}{\sum_{j \in \mathcal{S}(p_t)} \bar{w}_j(p_t)}, \tag{3}$$

where $\mathbf{1}[\cdot]$ is the indicator function. Thus $w_k \geq 0$, $\sum_{k=1}^{K} w_k = 1$, and at most two rules contribute to each prediction. The Gaussian regions may overlap, and different states can activate different pairs from the complete library.

The antecedent parameters specify a graded region rather than a hard interval: $c_{k,d}$ identifies its center and $\sigma_{k,d}$ sets its scale along physical coordinate d. A rule's executed region also depends on competing activations because Top-2 selection is relative to the entire library. Consequently, a displayed center and width alone do not establish where a rule has empirical support. Routing weights must be recorded on the evaluated transitions. The rule count K controls library capacity, whereas Top-2 controls the number of contributions inspected for one prediction.

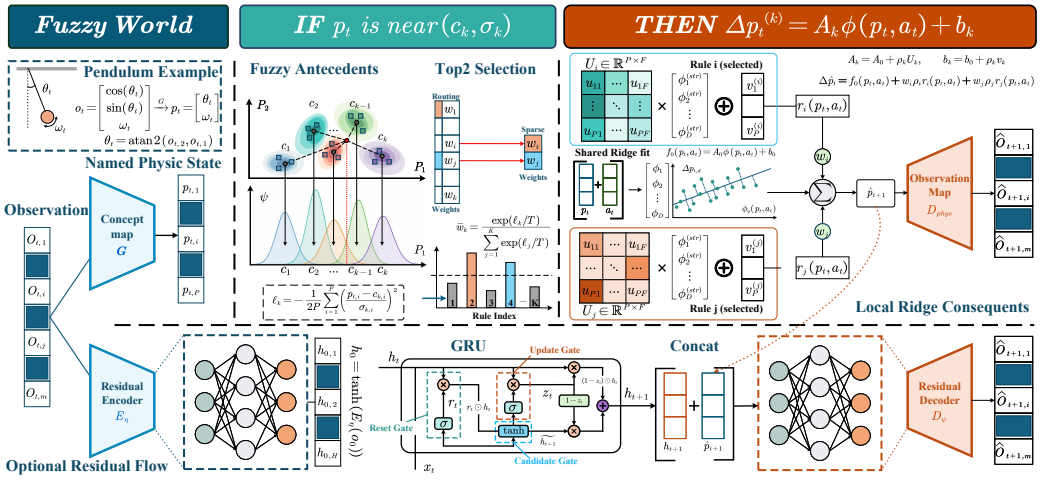


Figure 1: Fuzzy-World architecture. A concept map extracts named physical states, and Top-2 fuzzy routing combines at most two local state-action equations to predict physical increments. An optional GRU residual flow supplements observation reconstruction while the physical rules remain fixed during observation learning.

3.2 Explicit Local Dynamics and State Updates

Following Takagi-Sugeno functional consequents, each rule associates a physical state region with an explicit increment equation:

$$\mathcal{R}_k: \quad \textbf{IF } p_t \text{ is near } (c_k, \sigma_k), \quad \textbf{THEN } f_k(p_t, a_t) \equiv \Delta p_t^{(k)} = A_k \phi(p_t, a_t) + b_k. \tag{4}$$

Here $\phi(p_t, a_t) \in \mathbb{R}^F$ is the feature vector, F is its dimension, and $A_k \in \mathbb{R}^{P \times F}$ and $b_k \in \mathbb{R}^P$ map those features to changes in named state coordinates. Affine features use $\phi(p_t, a_t) = [p_t; a_t]$, where $[\cdot\,;\cdot]$ denotes vector concatenation. For a fixed coordinate set $I \subseteq \{1, \dots, P\}$, a structured alternative adds quadratic and cubic terms in $p_{t,I}$ and state-action interactions:

$$\phi_{\text{str}}(p_t, a_t) = [p_t;\ a_t;\ p_{t,I}^{\odot 2};\ p_{t,I}^{\odot 3};\ \text{vec}(p_{t,I} a_t^\top)], \tag{5}$$

where $p_{t,I}^{\odot m}$ denotes the element-wise mth power and vec vectorizes the state-action interaction matrix. The feature library is selected during development and fixed before final evaluation. Both configurations use the same routing and state update:

$$\Delta \hat{p}_t = \sum_{k \in \mathcal{S}(p_t)} w_k(p_t) f_k(p_t, a_t), \qquad \hat{p}_{t+1} = C(p_t + \Delta \hat{p}_t). \tag{6}$$

The canonicalization map $C : \mathbb{R}^P \to \mathbb{R}^P$ wraps periodic coordinates when needed and otherwise acts as the identity. Each weighted increment is an actual computational contribution. This additive interpretation holds before canonicalization and does not require an additive observation decoder.

For inspection, executed coefficients are converted to physical units. In this conversion, tildes explicitly mark the standardized quantities used above, and the superscript raw marks physical-unit quantities. Let $s = [p^{\text{raw}}; a^{\text{raw}}]$ be the raw state-action vector, μ_s its training mean, and D_s the diagonal matrix of training scales. The standardized input is $\tilde{s} = D_s^{-1}(s - \mu_s)$. If D_p is the diagonal output-scale matrix, $\Delta p^{\text{raw}} = D_p \Delta \tilde{p}$, and the standardized affine equation is $\Delta \tilde{p} = \tilde{A}_k \tilde{s} + \tilde{b}_k$, then

$$A_k^{\text{raw}} = D_p \tilde{A}_k D_s^{-1}, \qquad b_k^{\text{raw}} = D_p \tilde{b}_k - A_k^{\text{raw}} \mu_s. \tag{7}$$

Structured features undergo the same variable substitution. Export preserves the executed coefficients, including applicable numerical constraints, rather than fitting a separate explanation to model outputs. For polynomial features, centering changes several coefficients simultaneously: expanding a standardized cubic term produces cubic, quadratic, linear, and constant terms in physical units. Export therefore expands the entire expression before collecting coefficients by named variable. Rounding or suppressing small terms is used only for presentation; prediction and rule auditing use the full executed equations.

3.3 Activation-Aligned Rule Identification

Fuzzy-World uses shared-residual identification on all four tasks: a global Ridge fit captures shared dynamics, and activation-weighted local fits correct its residuals. K-means on standardized training states initializes antecedent centers. Widths use coordinate-wise cluster standard deviations with a positive floor, falling back to training-state scales when a cluster is too small. The resulting antecedents determine training weights $w_{tk} = w_k(p_t)$ using the same temperature and Top-2 selection as inference. For N_{tr} training transitions and true physical increments $\Delta p_t = p_{t+1} - p_t$, with wrapped differences for periodic coordinates, we first fit the shared Ridge model

$$(A_0, b_0) = \arg\min_{A,b} \sum_{t=1}^{N_{\mathrm{tr}}} \|\Delta p_t - A\phi(p_t, a_t) - b\|_2^2 + \lambda \|A\|_F^2. \tag{8}$$

This gives $f_0(p_t, a_t) = A_0\phi(p_t, a_t) + b_0$. Here $\lambda \geq 0$ is the Ridge coefficient and $\|\cdot\|_F$ is the Frobenius norm; the intercept is unpenalized.

We then form residual targets $e_t = \Delta p_t - f_0(p_t, a_t)$ and fit a local correction under each rule's executed activation weights:

$$(U_k, v_k) = \arg\min_{U,v} \sum_{t=1}^{N_{\mathrm{tr}}} w_{tk} \|e_t - U\phi(p_t, a_t) - v\|_2^2 + \lambda \|U\|_F^2. \tag{9}$$

The correction matrix $U_k \in \mathbb{R}^{P \times F}$ and vector $v_k \in \mathbb{R}^P$ define $r_k(p_t, a_t) = U_k\phi(p_t, a_t) + v_k$. Weighting by executed activation aligns local fitting with the regions where each correction contributes to prediction. Both stages remain linear in the coefficients even when ϕ contains polynomial features.

Finally, each local correction is shrunk and combined with the shared model, giving the executed rule $f_k = f_0 + \rho_k r_k$. The executed coefficients and support-dependent shrinkage are

$$\begin{aligned} A_k &= A_0 + \rho_k U_k, \qquad b_k = b_0 + \rho_k v_k, \\ \rho_k &= \frac{n_{\mathrm{eff},k}^{\mathrm{train}}}{n_{\mathrm{eff},k}^{\mathrm{train}} + \tau}, \quad n_{\mathrm{eff},k}^{\mathrm{train}} = \frac{(\sum_{t=1}^{N_{\mathrm{tr}}} w_{tk})^2}{\sum_{t=1}^{N_{\mathrm{tr}}} w_{tk}^2 + \epsilon_n}. \end{aligned} \tag{10}$$

Here $n_{\mathrm{eff},k}^{\mathrm{train}}$ is the training effective sample size of rule k, $\rho_k \in [0, 1]$ is its shrinkage factor, $\tau \geq 0$ controls shrinkage strength, and $\epsilon_n > 0$ stabilizes the denominator. Setting $\tau = 0$ disables additional shrinkage. Insufficient non-negligibly activated samples trigger a global-map fallback. The fitted antecedents and final consequents are subsequently fixed and shared by prediction and rule export. The shrinkage therefore limits weakly supported local corrections while preserving the shared dynamics.

3.4 Optional Residual Flow and Staged Training

When named states do not fully determine observations, an optional residual flow carries complementary information in $h_t \in \mathbb{R}^H$, where H is the residual-state dimension. The encoder E_η, recurrent map GRU_ξ, and decoder D_ψ have learnable parameters η, ξ, and ψ:

$$\begin{aligned} h_0 &= \tanh(E_\eta(o_0)), \qquad h_{t+1} = \mathrm{GRU}_\xi([p_t; a_t], h_t), \\ \hat{o}_{t+1} &= D_\psi([\hat{p}_{t+1}; h_{t+1}]). \end{aligned} \tag{11}$$

In Figure 1, r_t, z_t, and $\tilde{h}_{t+1}$ denote the GRU reset gate, update gate, and candidate state; $\sigma(\cdot)$ is the sigmoid function and $\odot$ denotes element-wise multiplication. When the physical state determines the observation, the known observation map $D_{\mathrm{phys}} : \mathbb{R}^P \to \mathbb{R}^M$ gives $\hat{o}_{t+1} = D_{\mathrm{phys}}(\hat{p}_{t+1})$ in place of learned decoding. During rollouts, both flows evolve from their predicted states without ground-truth reinitialization. At each step, antecedents are reevaluated at the predicted physical state, so the active pair can change along a trajectory. An explanation trace records the active rule indices, their normalized weights, and their individual increment vectors. Their weighted sum reconstructs the physical increment before C is applied. The same trace does not by itself decompose the decoded observation, whose dependence on the optional residual state is generally nonlinear.

Training first identifies the physical rules, then holds their parameters fixed while optimizing the active observation modules. For N_{obs} observation-training transitions, their loss is $\mathcal{L}_{\mathrm{obs}} =$

$N_{\text{obs}}^{-1} \sum_{t=1}^{N_{\text{obs}}} \|\hat{o}_{t+1} - o_{t+1}\|_2^2$. This stage adapts observation prediction without changing the induced physical equations. Fixing both antecedents and consequents preserves the routing function as well as the local responses for a given physical input. The experiments assess the resulting model at complementary levels: dynamical prediction measures aggregate accuracy, and the rule audit in Section 4.2 evaluates support and local predictive accuracy on held-out transitions. Appendix F illustrates how selected rule contributions combine into a physical increment. The observation residual flow supplements state information; the local corrections in Section 3.3 instead determine physical consequent coefficients. The audit retains the complete library, including rarely activated rules, and complements the explanation of an individual transition with an assessment of local validity.

4 EXPERIMENTS

The experiments examine prediction under distribution shift, the validity and execution of local rules, and the contributions of local correction and sparse aggregation.

Tasks and evaluation protocol. Pendulum-v1, HalfCheetah-v4, and Hopper-v4 are simulated using Gymnasium and MuJoCo (Towers et al., 2026; Todorov et al., 2012). A controlled periodic 1D advection task complements these environments with field dynamics, a setting also studied in PDE benchmarks (Takamoto et al., 2022); its trajectories are generated by Fourier phase shifts on a 64-point grid. For Pendulum, θ and ω denote angle and angular velocity, and the scalar action u denotes applied torque. The other named states comprise the 17 HalfCheetah coordinates, the 11 Hopper coordinates, and four Fourier statistics for advection. Training uses 64, 32, and 96 trajectories of length 80 for Pendulum, HalfCheetah, and advection, respectively, and 48 trajectories of at most 60 steps for Hopper. Out-of-distribution (OOD) tests expand the action range from $[-1, 1]$ to $[-2, 2]$ for Pendulum and from $[-0.5, 0.5]$ to $[-1, 1]$ for both MuJoCo tasks; Hopper additionally uses threefold reset perturbations. Advection velocity changes from $[0.4, 0.8]$ to $[1.0, 1.4]$. In-distribution (ID) and OOD splits use separate trajectories and training-set normalization. Results use five seeds; hyperparameter development uses a separate seed and an additional seed for HalfCheetah, including their OOD splits. Rule counts are fixed at $48/16/32/18$ for Pendulum/HalfCheetah/advection/Hopper.

4.1 DYNAMICAL PREDICTION UNDER DISTRIBUTION SHIFT

The dynamical prediction experiments evaluate whether explicit sparse rules achieve competitive one-step and multistep prediction accuracy. The seven comparators are hard local Ridge, dense Gaussian TSK, sparse identification of nonlinear dynamics (SINDy), global Ridge (Hoerl & Kennard, 1970), random-Fourier-feature (RFF) Ridge (Rahimi & Recht, 2007), Extra Trees (Geurts et al., 2006), and a two-hidden-layer physical-state multilayer perceptron (MLP). All receive the same named states, actions, and data splits. Baseline implementation details are provided in Appendix E. One-step and autoregressive rollout mean squared error (MSE) evaluate standardized observations over up to 50 available steps; variance-weighted R^2 separately evaluates physical trajectories. Pendulum uses analytic observation reconstruction, and the MuJoCo tasks use the identity map. Advection uses a 16-dimensional residual state and the same encoder-GRU-decoder architecture for every comparator, trained for 50 epochs with AdamW, batch size 512, and learning rate 0.003. All physical predictors remain fixed during observation learning. Rollouts use predicted physical and residual states, with the same physical-state numerical bound of $[-20, 20]$ for all methods. Performance tables report mean $\pm$ sample standard deviation across five seeds.

Table 1 shows competitive ID and OOD one-step prediction, with a more consistent advantage in long-horizon OOD prediction. Across the eight ID task-metric combinations, Fuzzy-World ranked first or second in six, indicating that its OOD rollout advantage coexists with strong relative performance on most ID metrics. Among the eight methods compared, Fuzzy-World achieved the lowest mean OOD 50-step observation MSE on all four systems. Relative to the best comparator mean, errors decreased by 88.4%, 0.2%, 6.1%, and 28.2% on Pendulum, HalfCheetah, advection, and Hopper, respectively. Gains were largest on Pendulum and Hopper, while HalfCheetah remained close to the strongest comparator. Figure 2 shows the same ranking for physical trajectories: Fuzzy-World attained the highest mean OOD R^2 in the main comparison on all four tasks (0.971, 0.556, 0.995, and 0.522). The long-horizon advantage therefore appeared in both observation reconstruction and physical-state prediction.

Table 1: Observation prediction errors for eight dynamics models with observation reconstruction over five seeds. Both one-step and rollout MSE evaluate standardized observations; rollouts cover up to 50 available steps. Entries are mean ± sample standard deviation. Bold and underline mark the lowest and second-lowest unrounded means within each dataset and metric, respectively.

Dataset	Model	ID		OOD	
		One-step MSE ↓	Rollout-50 MSE ↓	One-step MSE ↓	Rollout-50 MSE ↓
Pendulum	Fuzzy-World	$(2.119 \pm 0.910) \times 10^{-6}$	$(7.127 \pm 9.204) \times 10^{-3}$	$\mathbf{(8.622 \pm 6.215) \times 10^{-6}}$	$\mathbf{0.0104 \pm 0.0088}$
	Hard local	$(3.738 \pm 0.481) \times 10^{-4}$	0.1527 ± 0.0581	$(5.452 \pm 1.433) \times 10^{-4}$	0.2153 ± 0.0980
	Dense TSK	$(2.268 \pm 0.268) \times 10^{-4}$	0.1656 ± 0.0490	$(3.147 \pm 0.535) \times 10^{-4}$	0.2040 ± 0.0552
	SINDy	$\mathbf{(1.657 \pm 1.158) \times 10^{-6}}$	$\mathbf{(7.497 \pm 7.579) \times 10^{-4}}$	$(2.736 \pm 2.149) \times 10^{-5}$	0.0897 ± 0.1073
	Global Ridge	$(3.127 \pm 0.121) \times 10^{-3}$	1.1754 ± 0.0549	$(3.484 \pm 0.076) \times 10^{-3}$	1.3442 ± 0.0937
	RFF Ridge	$(4.377 \pm 0.795) \times 10^{-4}$	0.2784 ± 0.0528	$(2.027 \pm 0.315) \times 10^{-3}$	0.7438 ± 0.1020
	Extra Trees	$(5.880 \pm 2.349) \times 10^{-5}$	0.0864 ± 0.0205	$(3.020 \pm 0.474) \times 10^{-4}$	0.2037 ± 0.0884
	State MLP	$(2.011 \pm 2.003) \times 10^{-5}$	0.0491 ± 0.0511	$(5.906 \pm 1.562) \times 10^{-5}$	0.1221 ± 0.0789
HalfCheetah	Fuzzy-World	0.0294 ± 0.0014	0.0868 ± 0.0030	$\mathbf{1.7793 \pm 0.9867}$	$\mathbf{2.1700 \pm 0.6954}$
	Hard local	0.0309 ± 0.0016	0.0906 ± 0.0056	2.2929 ± 2.4982	2.2346 ± 0.6871
	Dense TSK	0.0318 ± 0.0013	0.0941 ± 0.0036	2.2650 ± 2.4123	2.1793 ± 0.6949
	SINDy	0.0174 ± 0.0008	0.0498 ± 0.0027	$(7.919 \pm 11.970) \times 10^{4}$	350.1077 ± 7.8879
	Global Ridge	0.0348 ± 0.0015	0.0988 ± 0.0029	2.1697 ± 1.3079	2.1739 ± 0.6898
	RFF Ridge	0.9278 ± 0.0209	8.4134 ± 2.1312	6.0170 ± 0.0455	16.6930 ± 1.1713
	Extra Trees	0.5062 ± 0.0075	0.9718 ± 0.0314	3.5747 ± 0.0323	4.3544 ± 0.5964
	State MLP	$\mathbf{0.0161 \pm 0.0006}$	$\mathbf{0.0440 \pm 0.0027}$	1.9038 ± 1.2068	2.2835 ± 0.7257
1D Advection	Fuzzy-World	$\mathbf{(7.607 \pm 0.416) \times 10^{-4}}$	0.1548 ± 0.0298	$(1.501 \pm 0.218) \times 10^{-3}$	$\mathbf{0.1797 \pm 0.0273}$
	Hard local	$(7.790 \pm 0.526) \times 10^{-4}$	0.1622 ± 0.0321	$(3.049 \pm 0.694) \times 10^{-3}$	0.7132 ± 0.2543
	Dense TSK	$(7.705 \pm 0.483) \times 10^{-4}$	0.1626 ± 0.0325	$(2.717 \pm 0.699) \times 10^{-3}$	0.7246 ± 0.2611
	SINDy	$(7.607 \pm 0.418) \times 10^{-4}$	0.1551 ± 0.0297	$\mathbf{(1.495 \pm 0.201) \times 10^{-3}}$	0.1914 ± 0.0305
	Global Ridge	$(7.667 \pm 0.399) \times 10^{-4}$	0.2853 ± 0.0483	$(3.337 \pm 0.593) \times 10^{-3}$	1.9942 ± 0.0232
	RFF Ridge	$(7.704 \pm 0.456) \times 10^{-4}$	0.1667 ± 0.0284	0.0134 ± 0.0016	2.0606 ± 0.1327
	Extra Trees	$(7.621 \pm 0.415) \times 10^{-4}$	0.1563 ± 0.0298	$(3.751 \pm 0.292) \times 10^{-3}$	1.5650 ± 0.0482
	State MLP	$(7.608 \pm 0.419) \times 10^{-4}$	$\mathbf{0.1543 \pm 0.0292}$	$(1.866 \pm 0.187) \times 10^{-3}$	0.4140 ± 0.0518
Hopper	Fuzzy-World	0.0101 ± 0.0020	0.1782 ± 0.0349	0.0593 ± 0.0196	$\mathbf{0.8750 \pm 0.2073}$
	Hard local	0.0175 ± 0.0088	0.2728 ± 0.0994	0.1123 ± 0.0793	1.2632 ± 0.4653
	Dense TSK	0.0152 ± 0.0078	0.2767 ± 0.0281	0.0878 ± 0.0423	1.2182 ± 0.3130
	SINDy	0.0145 ± 0.0130	53.0956 ± 33.6236	3.0365 ± 3.7770	233.4438 ± 41.6991
	Global Ridge	0.0137 ± 0.0023	0.3795 ± 0.0878	0.0579 ± 0.0080	1.3786 ± 0.3739
	RFF Ridge	0.0208 ± 0.0050	0.3848 ± 0.2166	0.2440 ± 0.0130	1.7005 ± 0.2348
	Extra Trees	0.0169 ± 0.0018	0.5528 ± 0.1506	0.1134 ± 0.0097	1.4875 ± 0.2123
	State MLP	$\mathbf{(5.604 \pm 1.372) \times 10^{-3}}$	$\mathbf{0.1403 \pm 0.0714}$	$\mathbf{0.0431 \pm 0.0227}$	2.4181 ± 3.1743

The ID and OOD comparisons separate fitting accuracy from recursive prediction under distribution shift. SINDy had the lowest ID rollout error on Pendulum, and the state MLP led on the other three tasks. On Pendulum, SINDy's ID rollout MSE was 7.497×10^{-4}, below Fuzzy-World's 7.127×10^{-3}, but its OOD error rose to 0.0897, compared with 0.0104 for Fuzzy-World. Hopper showed a related reversal across prediction horizons: the state MLP had lower OOD one-step MSE (0.0431 versus 0.0593), yet higher rollout MSE (2.4181 versus 0.8750). Thus, the main benefit of Fuzzy-World emerged during recursive OOD prediction, rather than from having the lowest error on every single-step or ID metric.

The architecture separates state-region selection from action response. Antecedents depend only on named physical states, while both states and actions enter explicit consequents. At a fixed state, a change in action magnitude therefore changes the predicted increment without directly changing the rule partition. HalfCheetah and Hopper use affine consequents; Pendulum and advection additionally use higher-order state terms and state-action interactions. This separation provides a structural rationale for prediction under the action-range and velocity shifts tested here: actions are evaluated through explicit local response functions within state-defined regions.

Shared dynamics provide a common fitted map, with local residuals representing region-specific departures. HalfCheetah and Hopper additionally shrink weakly supported corrections according to effective sample size. The matched ablations in Section 4.3 show that local corrections improve OOD rollout prediction particularly on Pendulum and Hopper, while a small active-rule budget retains competitive performance. Fuzzy-World also achieved lower mean OOD rollout MSE than Hard local and Dense TSK on all four tasks.

The CEM-MPC experiment in Appendix C provides complementary evidence for action selection. With the same planning budget, Fuzzy-World achieved the highest mean environment returns on HalfCheetah and Hopper (24.9 and 173.3), and performed comparably to the other local-rule con-

trollers on Pendulum. These short-horizon control results show that the learned dynamics can support receding-horizon action selection as well as offline trajectory prediction.

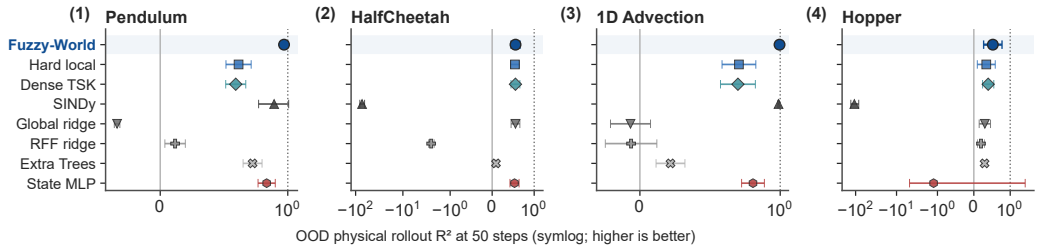


Figure 2: OOD 50-step physical-trajectory R^2 for the eight dynamics models with observation reconstruction over five seeds. Points and whiskers show mean $\pm$ sample standard deviation. Higher is better; a symmetric logarithmic horizontal scale retains the large negative outliers.

4.2 Local Rule Validity and Prediction Decomposition

To examine the local equations underlying aggregate OOD performance, we audited every rule using its actual Top-2 weights on held-out transitions. A rule was considered supported when its held-out effective sample size satisfied $n_{\text{eff},k}^{\text{eval}} \geq F + 2$ and its mean activation mass satisfied $m_k \geq 10^{-4}$, where F is the consequent feature dimension. Table 2 pools the complete libraries across five seeds and reports local R_k^2 for supported rules. Appendix F gives the metric definitions, physical-unit equations, and an executed prediction.

Table 2: Complete-library rule audit, pooling five seeds within each task and split. Supported rules satisfy both support thresholds. "Mass" is their fraction of total activation weight. Median, weighted mean, and negative fraction summarize local R^2 among supported rules; these are rule-level summaries, not five-seed means or whole-model R^2.

		Supported (%)		Local R^2 on supported rules		
Dataset	Split	Rules	Mass	Median	Weighted mean	Negative (%)
Pendulum	ID	100.00	100.00	0.9990	0.9977	0.00
	OOD	100.00	100.00	0.9994	0.9952	0.00
HalfCheetah	ID	100.00	100.00	0.9731	0.9717	0.00
	OOD	100.00	100.00	0.8897	0.6383	5.00
1D Advection	ID	91.88	99.31	1.0000	0.9999	0.00
	OOD	93.12	99.80	0.9951	0.9949	0.00
Hopper	ID	83.33	97.48	0.8749	0.8596	0.00
	OOD	77.78	95.53	0.8510	0.7856	0.00

Supported rules carried 95.53%-100% of OOD activation mass, with median local R_k^2 between 0.8510 and 0.9994. On Hopper, 77.78% of rules passed the support criteria but carried 95.53% of activation mass. The audit also revealed regional failures: four of 80 HalfCheetah OOD rule records had negative local R^2, and the activation-weighted mean (0.6383) was below the median (0.8897). Good aggregate OOD performance thus accompanied accurate predictions from most supported equations, while support and local accuracy remained distinct checks.

Sparse execution connects these equations to individual predictions. For Pendulum, the consequents expose how angle, angular velocity, and torque enter each physical increment; the example term $(0.116 - 0.012\theta)u$ in Appendix F.3 expresses an angle-dependent action response. The selected rules, their normalized weights, and their proposed increments reconstruct the physical prediction directly. The audit tests where the equations predict accurately, while execution decomposition identifies which equations produce a particular increment.

4.3 Ablation Studies

Table 3 compares Top-1, Top-2 (Fuzzy-World), and dense aggregation under the main protocol, with identical antecedents, rule counts, feature libraries, regularization, and shrinkage. Local residuals were refitted with each variant's execution weights; physical predictors remained fixed during observation learning. The w/o local correction variant retains the feature library but fits only the shared Ridge map. Appendix G reports the complete Top-1-8 sweep.

Table 3: Matched ablations of activation and local corrections. Entries are mean ± sample standard deviation over five seeds; all metrics are standardized observation MSE. Top-2 was rerun with the same Fuzzy-World configuration as Table 1. Dense retains all rules; w/o local correction retains the feature library but removes local residuals. Bold marks the lowest unrounded mean within each task and metric.

Dataset	Variant	ID		OOD	
		One-step MSE	Rollout-50 MSE	One-step MSE	Rollout-50 MSE
Pendulum	Top-1	$(2.450 \pm 0.860) \times 10^{-6}$	$(6.855 \pm 9.669) \times 10^{-3}$	$(8.956 \pm 6.206) \times 10^{-6}$	$\mathbf{0.0103 \pm 0.0083}$
	Top-2 (Fuzzy-World)	$(2.119 \pm 0.910) \times 10^{-6}$	$(7.127 \pm 9.204) \times 10^{-3}$	$\mathbf{(8.622 \pm 6.215) \times 10^{-6}}$	0.0104 ± 0.0088
	Dense	$\mathbf{(2.043 \pm 0.898) \times 10^{-6}}$	$\mathbf{(5.337 \pm 4.923) \times 10^{-3}}$	$(8.638 \pm 6.225) \times 10^{-6}$	0.0108 ± 0.0090
	w/o local correction	$(8.471 \pm 0.265) \times 10^{-5}$	0.0695 ± 0.0109	$(9.293 \pm 0.870) \times 10^{-5}$	0.0859 ± 0.0268
HalfCheetah	Top-1	0.0306 ± 0.0015	0.0889 ± 0.0027	1.8490 ± 1.0564	2.1725 ± 0.6870
	Top-2 (Fuzzy-World)	$\mathbf{0.0294 \pm 0.0014}$	$\mathbf{0.0868 \pm 0.0030}$	1.7793 ± 0.9867	$\mathbf{2.1700 \pm 0.6954}$
	Dense	0.0313 ± 0.0014	0.0910 ± 0.0030	$\mathbf{1.5027 \pm 0.7110}$	2.1741 ± 0.6935
	w/o local correction	0.0348 ± 0.0015	0.0988 ± 0.0029	2.1697 ± 1.3079	2.1739 ± 0.6898
1D Advection	Top-1	$(7.607 \pm 0.416) \times 10^{-4}$	0.1548 ± 0.0298	$(1.501 \pm 0.216) \times 10^{-3}$	0.1792 ± 0.0270
	Top-2 (Fuzzy-World)	$(7.606 \pm 0.416) \times 10^{-4}$	0.1548 ± 0.0298	$(1.501 \pm 0.219) \times 10^{-3}$	0.1797 ± 0.0273
	Dense	$\mathbf{(7.606 \pm 0.416) \times 10^{-4}}$	$\mathbf{0.1548 \pm 0.0298}$	$\mathbf{(1.501 \pm 0.219) \times 10^{-3}}$	0.1798 ± 0.0272
	w/o local correction	$(7.606 \pm 0.417) \times 10^{-4}$	0.1548 ± 0.0297	$(1.502 \pm 0.217) \times 10^{-3}$	$\mathbf{0.1792 \pm 0.0272}$
Hopper	Top-1	0.0117 ± 0.0022	0.2086 ± 0.0424	0.0653 ± 0.0208	1.0307 ± 0.2202
	Top-2 (Fuzzy-World)	$\mathbf{0.0101 \pm 0.0020}$	$\mathbf{0.1782 \pm 0.0349}$	0.0593 ± 0.0196	$\mathbf{0.8750 \pm 0.2073}$
	Dense	0.0117 ± 0.0022	0.2260 ± 0.0469	$\mathbf{0.0566 \pm 0.0123}$	0.9317 ± 0.1505
	w/o local correction	0.0136 ± 0.0023	0.3804 ± 0.0852	0.0591 ± 0.0084	1.3811 ± 0.3798

Removing local corrections raised mean OOD rollout MSE from 0.0104 to 0.0859 on Pendulum and from 0.8750 to 1.3811 on Hopper. HalfCheetah changed little (2.1700 versus 2.1739), while advection's shared feature map already matched the Full model closely (0.1792 versus 0.1797). Local corrections therefore provided their largest benefits on Pendulum and Hopper under the retained feature libraries.

Top-2 reduced mean OOD rollout MSE on Hopper from 1.0307 with Top-1 to 0.8750, a 15.1% reduction, and also improved ID and OOD one-step prediction. HalfCheetah showed a smaller rollout improvement, while Top-1 had slightly lower OOD rollout error on Pendulum and advection. Relative to dense aggregation, Top-2 had lower mean OOD rollout error on all four tasks, with reductions of 3.6%, 0.2%, 0.05%, and 6.1%, respectively.

Top-2 is the smallest budget permitting cooperation between overlapping local rules, with each prediction inspectable through at most two weighted contributions. It balances predictive flexibility and concise explanation, although Top-3/4 further improve Hopper in the full sweep.

5 Conclusion

Fuzzy-World combines a shared physical map with sparse local corrections in named coordinates, making predicted transitions traceable to active rules. Across the evaluated systems, it maintained competitive ID and OOD one-step prediction and achieved the strongest mean OOD long-horizon performance among the main comparators. Matched ablations support local correction and show that a small activation budget can preserve predictive performance, while rule audits assess local predictive accuracy. Together, these results support sparse local dynamics for prediction under distribution shift with directly inspectable transitions.

REFERENCES

David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf.

Plamen P. Angelov and Dimitar P. Filev. An approach to online identification of Takagi–Sugeno fuzzy models. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1): 484–498, 2004. doi: 10.1109/TSMCB.2003.817053.

Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, and Koray Kavukcuoglu. Interaction networks for learning about objects, relations and physics. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/3147da8ab4a0437c15ef51a5cc7f2dc4-Paper.pdf.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016. doi: 10.1073/pnas.1517384113.

Kathleen Champion, Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences*, 116 (45):22445–22451, 2019. doi: 10.1073/pnas.1906995116.

Swarat Chaudhuri, Kevin Ellis, Oleksandr Polozov, Rishabh Singh, Armando Solar-Lezama, and Yisong Yue. Neurosymbolic programming. *Foundations and Trends in Programming Languages*, 7(3):158–243, 2021. doi: 10.1561/2500000049.

Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2:772–782, 2020. doi: 10.1038/s42256-020-00265-z.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*, 2018.

Pieter-Tjerk de Boer, Dirk P. Kroese, Shie Mannor, and Reuven Y. Rubinstein. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134(1):19–67, 2005. doi: 10.1007/s10479-005-5724-z.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lió, and Mateja Jamnik. Concept embedding models: Beyond the accuracy-explainability trade-off. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/867c06823281e506e8059f5c13a57f75-Abstract-Conference.html.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63:3–42, 2006. doi: 10.1007/s10994-006-6226-1.

Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/26cd8ecadce0d4efd6cc8a8725cbd1f8-Abstract.html.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 2025. doi: 10.1038/s41586-025-08744-2.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

J.-S. R. Jang. ANFIS: Adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man, and Cybernetics*, 23(3):665–685, 1993. doi: 10.1109/21.256541.

T.A. Johansen, R. Shorten, and R. Murray-Smith. On the interpretation and identification of dynamic takagi-sugeno fuzzy models. *IEEE Transactions on Fuzzy Systems*, 8(3):297–313, 2000. doi: 10.1109/91.855918.

Thomas Kipf, Elise van der Pol, and Max Welling. Contrastive learning of structured world models. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1gax6VtDB.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124, 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9(1):4950, 2018. doi: 10.1038/s41467-018-07210-0.

Jordan Peper, Zhenjiang Mao, Yuang Geng, Siyuan Pan, and Ivan Ruchkin. Four principles for physically interpretable world models. In *Proceedings of the International Conference on Neurosymbolic Systems*, volume 288 of *Proceedings of Machine Learning Research*, pp. 66–89, 2025. URL https://proceedings.mlr.press/v288/peper25a.html.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007. URL https://proceedings.neurips.cc/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html.

Jonathan Richens and Tom Everitt. Robust agents learn causal world models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/44a2b9f7bf9aec3f1fa333ad875b0ee0-Abstract-Conference.html.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2019. doi: 10.1038/s42256-019-0048-x.

Subham Sahoo, Christoph Lampert, and Georg Martius. Learning equations for extrapolation and control. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4442–4450, 2018. URL https://proceedings.mlr.press/v80/sahoo18a.html.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8459–8468, 2020. URL https://proceedings.mlr.press/v119/sanchez-gonzalez20a.html.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

Zhao Su, Jun Shen, Yu Sun, Rizhen Hu, Qingguo Zhou, and Binbin Yong. Deep spatio-temporal fuzzy model for NDVI forecasting. *IEEE Transactions on Fuzzy Systems*, 33(1):290–301, 2025. doi: 10.1109/TFUZZ.2024.3409894.

Tomohiro Takagi and Michio Sugeno. Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-15(1):116–132, 1985. doi: 10.1109/TSMC.1985.6313399.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An extensive benchmark for scientific machine learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-0117. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/0a9747136d411fb83f0cf81820d44afb-Abstract-Datasets_and_Benchmarks.html.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.

Mark Towers, Ariel Kwiatkowski, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Kallinteris Andreas, Markus Krimmel, Arjun KG, Rodrigo De Lazcano Perez-Vicente, J K Terry, Andrea Pierré, Sander V Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=qPMLvJxtPK.

Zizhao Wang, Xuesu Xiao, Zifan Xu, Yuke Zhu, and Peter Stone. Causal dynamics learning for task-independent state abstraction. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23151–23180. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wang22ae.html.

J. Yen, Liang Wang, and C.W. Gillespie. Improving the interpretability of tsk fuzzy models by combining global learning and local learning. *IEEE Transactions on Fuzzy Systems*, 6(4):530–537, 1998. doi: 10.1109/91.728447.

Binbin Yong, Haoran Pei, Jun Shen, Haoran Li, Qingguo Zhou, and Zhao Su. KANFIS: A neuro-symbolic framework for interpretable and uncertainty-aware learning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=MKvy0hgAK4.

Zhongwei Yu, Jingqing Ruan, and Dengpeng Xing. Learning causal dynamics models in object-oriented environments. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57597–57638, 2024. URL https://proceedings.mlr.press/v235/yu24j.html.

LLM USE STATEMENT

We used large language models to assist with drafting and polishing the manuscript, retrieving and identifying relevant literature, and developing and checking proofs of mathematical claims. All AI-assisted content, references, and derivations were reviewed and verified by the authors, who take full responsibility for the accuracy and integrity of the final paper.

A THEORETICAL ANALYSIS OF FUZZY-WORLD

This appendix establishes properties of rule identification and aggregation in Sections 3.3 and 3.2. Antecedents and features are fixed during consequent estimation. Prediction identities concern physical increments before canonicalization by C or observation decoding. Features are finite, and aggregation weights are nonnegative and sum to one. The results characterize the model structure; they do not assume access to true dynamics outside the observed distribution.

A.1 WEIGHTED LOCAL IDENTIFICATION: SOLUTION AND UNIQUENESS

Let $X \in \mathbb{R}^{N_{\mathrm{tr}} \times (F+1)}$ have rows $[\phi(p_t, a_t)^\top, 1]$, and let $Y \in \mathbb{R}^{N_{\mathrm{tr}} \times P}$ have rows $\Delta p_t^\top$. Define $W_k = \mathrm{diag}(w_{1k}, \ldots, w_{N_{\mathrm{tr}}k})$ and $\Lambda = \mathrm{diag}(1, \ldots, 1, 0)$, which leaves the intercept unpenalized. The shared Ridge fit in Section 3.3 gives

$$B_0 = (X^\top X + \lambda\Lambda)^{-1} X^\top Y, \qquad B_0 = [A_0\ b_0]^\top, \tag{12}$$

for $\lambda > 0$ and a nonempty training set. Holding B_0 fixed, define residual targets $E = Y - XB_0$. For a correction matrix $Q \in \mathbb{R}^{(F+1)\times P}$, local fitting minimizes

$$\mathcal{J}_k(Q) = \|W_k^{1/2}(E - XQ)\|_F^2 + \lambda\,\mathrm{tr}(Q^\top \Lambda Q), \tag{13}$$

where tr denotes matrix trace.

Proposition A.1 (Unique weighted Ridge correction). If $\lambda > 0$ and $\sum_{t=1}^{N_{\mathrm{tr}}} w_{tk} > 0$, the unique minimizer is

$$Q_k = (X^\top W_k X + \lambda\Lambda)^{-1} X^\top W_k E, \qquad Q_k = [U_k\ v_k]^\top. \tag{14}$$

Proof. Differentiation gives $\nabla_Q \mathcal{J}_k = 2X^\top W_k(XQ - E) + 2\lambda\Lambda Q$. Let $q = [v^\top, c]^\top \neq 0$, with $v \in \mathbb{R}^F$, $c \in \mathbb{R}$, and abbreviate $\phi_t = \phi(p_t, a_t)$. Then

$$q^\top (X^\top W_k X + \lambda\Lambda) q = \sum_t w_{tk}(\phi_t^\top v + c)^2 + \lambda\|v\|_2^2 > 0. \tag{15}$$

If $v \neq 0$, the regularizer is positive. Otherwise $c \neq 0$, and the sum equals $c^2 \sum_t w_{tk} > 0$. The normal matrix is positive definite, so the objective is strictly convex. Setting its gradient to zero proves Equation (14). □

For $\lambda = 0$, uniqueness instead requires $W_k^{1/2} X$ to have full column rank. Zero total activation leaves the unpenalized intercept unidentified. The implementation's minimum-support fallback is more conservative than this algebraic existence condition.

The final execution parameters are $B_k = [A_k\ b_k]^\top = B_0 + \rho_k Q_k$. Proposition A.1 establishes uniqueness of the residual correction Q_k for fixed B_0. Support-dependent scaling and combination with the shared map follow fitting; Section A.4 analyzes this scaling.

A.2 ERROR PROPERTIES OF FUZZY AGGREGATION

Fix (p, a) and a target increment $y \in \mathbb{R}^P$. Write $f_k = f_k(p, a)$, $w_k = w_k(p)$, and $\bar{f} = \sum_k w_k f_k$. Zero weights allow sums over the complete library.

Proposition A.2 (Local-error decomposition). For nonnegative normalized weights,

$$\sum_k w_k \|y - f_k\|_2^2 = \|y - \bar{f}\|_2^2 + \sum_k w_k \|f_k - \bar{f}\|_2^2. \tag{16}$$

Consequently,

$$\|y - \bar{f}\|_2^2 \le \sum_k w_k \|y - f_k\|_2^2. \tag{17}$$

Proof. Expand $y - f_k = (y - \bar{f}) - (f_k - \bar{f})$. The first squared terms sum to $\|y - \bar{f}\|_2^2$ because $\sum_k w_k = 1$. The cross terms vanish since $\sum_k w_k(f_k - \bar{f}) = 0$. The remaining squared terms give Equation (16); their nonnegativity gives the bound. □

For Top-2 weights w_i and $w_j = 1 - w_i$, the disagreement term is

$$\sum_{k \in \{i,j\}} w_k \|f_k - \bar{f}\|_2^2 = w_i w_j \|f_i - f_j\|_2^2. \tag{18}$$

Accurate active consequents therefore bound aggregate increment error. The converse does not hold: local errors can compensate. Summing Equation (17) over transitions connects the unregularized local fitting errors to aggregate squared error, explaining why individual equations merit separate evaluation. The identity does not directly bound aggregate error using median local R^2, which also involves rule-specific target variances and summary operations.

A.3 Approximation Error Induced by Top-2 Truncation

Fix the input and use identical fitted consequents for both aggregation schemes. Let $\bar{w}_k$ denote dense Softmax weights and $S = \mathcal{S}(p)$ the Top-2 set, with fixed tie breaking. Define discarded activation mass and retained prediction by

$$\varepsilon = \sum_{k \notin S} \bar{w}_k, \qquad f_S = \sum_{k \in S} \frac{\bar{w}_k}{1 - \varepsilon} f_k. \tag{19}$$

Finite Softmax scores ensure $1 - \varepsilon > 0$. For $\varepsilon > 0$, define $f_{\bar{S}} = \sum_{k \notin S} (\bar{w}_k / \varepsilon) f_k$, the discarded rules' normalized prediction.

Proposition A.3 (Truncation bound). Let $f_{\text{dense}} = \sum_k \bar{w}_k f_k$ and $f_{\text{Top2}} = f_S$. For $0 < \varepsilon < 1$,

$$\begin{aligned} f_{\text{Top2}} - f_{\text{dense}} &= \varepsilon (f_S - f_{\bar{S}}), \\ \|f_{\text{Top2}} - f_{\text{dense}}\|_2 &\le \varepsilon D(p, a), \end{aligned} \tag{20}$$

where $D(p, a) = \max_{r,s} \|f_r(p, a) - f_s(p, a)\|_2$ is the output diameter of the finite rule library. If $\varepsilon = 0$, both predictions coincide.

Proof. Partition the dense sum to obtain $f_{\text{dense}} = (1 - \varepsilon) f_S + \varepsilon f_{\bar{S}}$. This gives the first identity. Set $\alpha_i = \bar{w}_i / (1 - \varepsilon)$ for $i \in S$ and $\beta_j = \bar{w}_j / \varepsilon$ for $j \notin S$. Each set sums to one, hence

$$f_S - f_{\bar{S}} = \sum_{i \in S} \sum_{j \notin S} \alpha_i \beta_j (f_i - f_j). \tag{21}$$

By the triangle inequality, its norm is at most $\sum_{i,j} \alpha_i \beta_j D(p, a) = D(p, a)$. Multiplication by ε proves the bound. □

For any target increment y, it follows that

$$\|y - f_{\text{Top2}}\|_2 \le \|y - f_{\text{dense}}\|_2 + \varepsilon D(p, a). \tag{22}$$

Thus truncation is a controlled approximation when discarded mass and consequent disagreement have a small product. Low discarded mass alone is insufficient if discarded outputs differ greatly. This is a comparison of aggregation on fixed fitted rules, not a bound on the difference between separately trained sparse and dense models.

A.4 Support-Dependent Bounds on Local Corrections

Following Equation (10), define $r_k(p, a) = U_k \phi(p, a) + v_k$ and write $f_k = f_0 + \rho_k r_k$. Here r_k is a residual function, not an activation weight. Effective training sample size determines ρ_k; the following comparison holds fitted residual functions fixed.

Proposition A.4 (Support-dependent correction control). For $\tau > 0$, the function $\rho(n) = n/(n+\tau)$ for $n \geq 0$ is increasing, lies in $[0, 1)$, and satisfies

$$\rho'(n) = \frac{\tau}{(n+\tau)^2} > 0, \qquad \rho(0) = 0, \qquad \lim_{n\to\infty} \rho(n) = 1. \tag{23}$$

For normalized nonnegative execution weights,

$$\begin{aligned} \|\Delta\hat{p} - f_0\|_2 &\leq \sum_k w_k \rho_k \|r_k\|_2 \\ &\leq \sum_k w_k \rho_k \big(\|U_k\|_F \|\phi(p,a)\|_2 + \|v_k\|_2\big). \end{aligned} \tag{24}$$

Proof. Differentiation and taking limits give the stated properties of ρ. Since the weights sum to one,

$$\Delta\hat{p} - f_0 = \sum_k w_k (f_0 + \rho_k r_k) - f_0 = \sum_k w_k \rho_k r_k. \tag{25}$$

The triangle inequality gives the first bound, and $\|U_k\phi\|_2 \leq \|U_k\|_F \|\phi\|_2$ gives the second. □

For example, if every active rule satisfies $n_{\mathrm{eff},k}^{\mathrm{train}} \leq n_\star$ and $\|r_k(p,a)\|_2 \leq B_{\mathrm{res}}$, where $n_\star$ bounds support and B_{res} bounds residual outputs at the input, then

$$\|\Delta\hat{p} - f_0\|_2 \leq \frac{n_\star}{n_\star + \tau} B_{\mathrm{res}}. \tag{26}$$

This quantifies attenuation of weakly supported corrections relative to the shared map. It does not imply that the fitted residuals stay unchanged when training data change, or that shrinkage toward an inaccurate shared map must reduce prediction error. Setting $\tau = 0$ gives $\rho_k = 1$ for positive support and disables additional attenuation. Zero-support rules use the shared-map fallback rather than the undefined ratio $0/0$.

B Limitations

Fuzzy-World assumes that named physical states are available and sufficiently Markovian for local one-step equations to support autoregressive rollouts. When named states must be estimated from observations, estimation errors may alter rule routing and accumulate during rollout. Structured feature libraries can also lengthen individual consequents, while support-aware shrinkage does not prevent every poorly fitting rule.

The main model uses a fixed Top-2 activation budget. Although the sensitivity analysis compares Top-1 through Top-8 and dense routing, the model does not automatically select the number of active rules for each task or input. A fixed budget can limit adaptation to regions requiring different levels of local cooperation; learning an adaptive budget is a direction for future work. Evaluation currently covers four dynamical systems and short-horizon control, leaving broader tasks and longer-horizon control to be assessed. The audited equations capture predictive regularities in observed dynamics rather than causal mechanisms.

C Short-Horizon Planning and Control

This experiment tests whether the learned local dynamics support action selection in a real simulator. Each dynamics model is fitted to the task's random-data training split and then held fixed during evaluation. At each environment step, a receding-horizon model predictive controller optimizes candidate action sequences using the cross-entropy method (CEM) (de Boer et al., 2005), executes the first action, and replans from the newly observed physical state. Hard local, Dense TSK, and SINDy use the same task-specific planning budget as Fuzzy-World; random and zero actions are non-planning references.

Pendulum uses a planning horizon of 10 steps and 128 candidates per iteration; HalfCheetah and Hopper use 6 steps and 96 candidates. All controllers perform three CEM iterations, retain the top 10% of candidates rounded down with a minimum of four elites, and score predicted rewards with

discount factor 0.99. Action sequences are clipped to environment bounds. The search distribution is updated from elite means and standard deviations, with an additional standard-deviation floor of 5% of the action range. Planning uses task-specific reward proxies computed from predicted physical states, including an unhealthy-state penalty for Hopper; reported returns are undiscounted rewards from actual environment execution. The implementation also clips predicted standardized states to $[-8, 8]$ after canonicalization and replaces non-finite values. These safeguards are shared across planning models. Evaluation uses five seeds and one environment rollout per seed, capped at 100 steps for Pendulum and 150 for the MuJoCo tasks, with early termination when the environment ends.

Table 4: True-environment return under short-horizon CEM-MPC, reported as mean $\pm$ sample standard deviation over five seeds. Higher is better. Random and zero actions do not use a dynamics model.

Controller	HalfCheetah	Hopper	Pendulum
Random	-43.7 ± 21.5	18.5 ± 9.9	-592.8 ± 149.1
Zero	-0.1 ± 0.8	123.9 ± 11.5	-597.0 ± 124.1
Fuzzy-World	24.9 ± 23.7	173.3 ± 2.9	-182.3 ± 71.7
Hard local	11.6 ± 43.0	111.4 ± 59.9	-181.9 ± 71.1
Dense TSK	7.9 ± 9.3	146.5 ± 50.5	-181.6 ± 71.0
SINDy	4.6 ± 16.5	45.8 ± 67.2	-157.3 ± 58.3

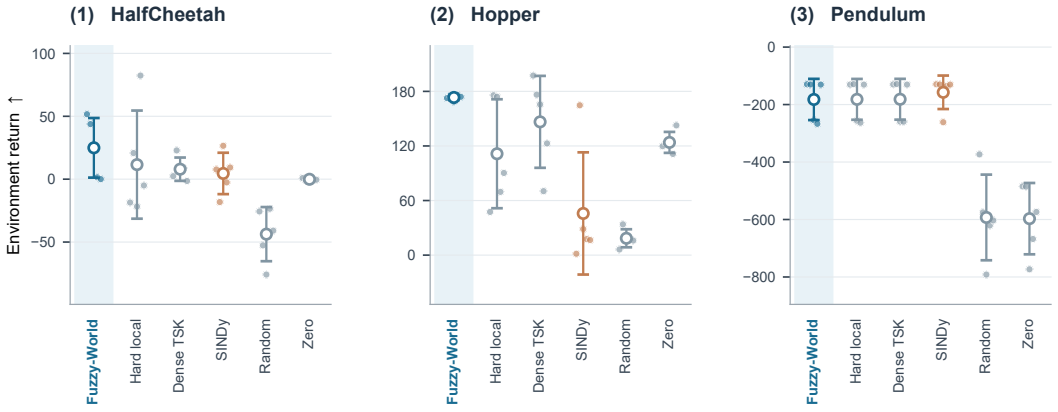


Figure 3: Short-horizon CEM-MPC returns. Small filled points show five individual seeds; open circles and error bars indicate the mean and sample standard deviation. Higher is better.

Table 4 and Figure 3 give Fuzzy-World the highest mean return on HalfCheetah and Hopper, improving on zero actions by 25.0 and 49.4 points, respectively. On Pendulum, its mean return improves by 414.7 points over zero actions and is close to the other local-rule controllers, while SINDy obtains the highest mean return. These results support the use of the learned dynamics for action selection under a common short-horizon planning protocol.

D Additional Five-Member Probabilistic Ensemble Comparison

An additional neural baseline uses a five-member probabilistic dynamics ensemble inspired by PETS (Chua et al., 2018). Each member receives the same standardized physical states and actions as the direct-transition models in Section 4.1. It has three hidden layers of width 256 with SiLU activations and predicts an increment mean and diagonal log variance. Members receive independently sampled training batches with replacement. The baseline is trained for 300 epochs with batch size 512 using AdamW, learning rate 10^{-3}, weight decay 10^{-5}, and gradient-norm clipping at 20.

The implemented objective combines mean-squared increment error with Gaussian negative log likelihood weighted by 0.01, together with a 0.01-weighted regularizer on the learned upper and lower log-variance bounds. At prediction time, the five member means are averaged to obtain the

physical increment. This PETS-style baseline evaluates deterministic ensemble-mean dynamics, without uncertainty-aware trajectory sampling or PETS planning. Training splits, normalization, and physical-state evaluation follow the main direct-transition comparison; results use five seeds.

Table 5: Additional probabilistic-ensemble comparison over five seeds. Entries are mean ± sample standard deviation. All metrics use physical-state rollouts over up to 50 available steps. Trajectory R^2 is variance weighted and higher is better; standardized-state MSE is lower is better. Fuzzy-World values are from the saved direct physical-state evaluation used for this supplementary comparison.

Dataset	Model	ID R^2	OOD R^2	OOD MSE
Pendulum	Fuzzy-World	0.977 ± 0.009	0.971 ± 0.012	0.030 ± 0.012
Pendulum	Probabilistic ensemble	0.781 ± 0.082	0.779 ± 0.117	0.228 ± 0.123
HalfCheetah	Fuzzy-World	0.913 ± 0.003	0.556 ± 0.106	2.170 ± 0.695
HalfCheetah	Probabilistic ensemble	0.973 ± 0.001	-9.074 ± 2.873	48.089 ± 15.214
1D Advection	Fuzzy-World	1.000 ± 0.000	0.995 ± 0.001	$(5.000 \pm 1.000) \times 10^{-3}$
1D Advection	Probabilistic ensemble	1.000 ± 0.000	0.973 ± 0.004	0.028 ± 0.004
Hopper	Fuzzy-World	0.809 ± 0.039	0.522 ± 0.253	0.875 ± 0.207
Hopper	Probabilistic ensemble	0.029 ± 1.954	-1.409 ± 2.309	3.255 ± 2.365

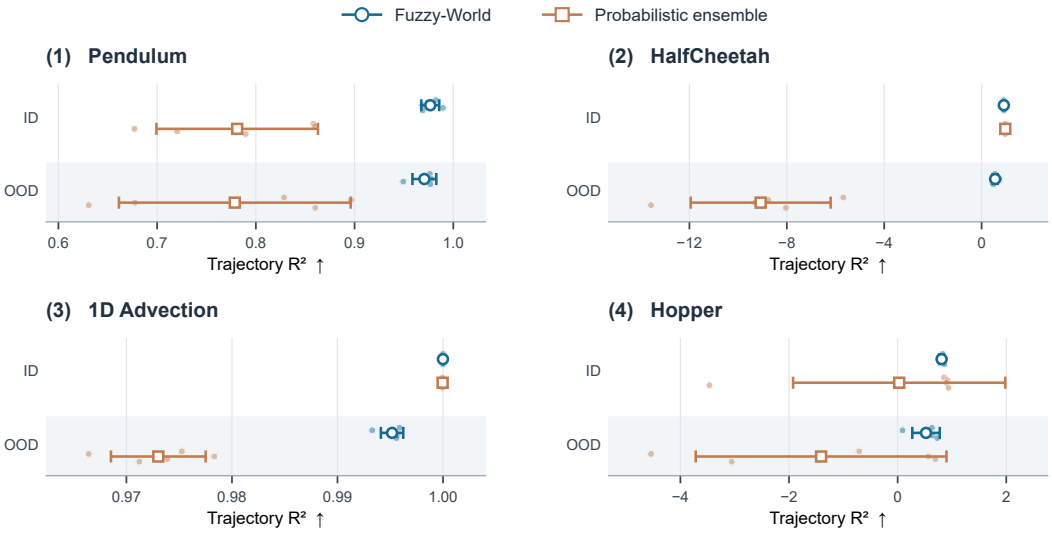


Figure 4: ID and OOD trajectory R^2 for Fuzzy-World and the five-member probabilistic ensemble, using rollouts of up to 50 steps. Small points show five individual seeds; open markers and error bars indicate the mean and sample standard deviation. Shaded rows denote OOD evaluation. Linear axis ranges differ across tasks.

Fuzzy-World achieves higher mean OOD trajectory R^2 on all four systems in this additional comparison (Table 5; Figure 4). The ensemble retains a high OOD score on advection (0.973), compared with 0.995 for Fuzzy-World, but obtains lower scores on Pendulum and negative mean scores on both MuJoCo tasks. These results extend the observed OOD prediction advantage to a five-member probabilistic neural ensemble.

E Local Baseline Configurations

Hard local and Dense TSK in Table 1 both construct antecedents by K-means clustering of concatenated standardized physical states and actions. Each cluster fits an affine Ridge consequent directly to physical increments, with a global fit used when cluster support is insufficient. Hard local selects the rule with the smallest width-normalized distance; Dense TSK combines all rules using normalized Gaussian activations. Thus, dense aggregation is used at inference, while its consequents are fitted within hard clusters.

These baselines differ from Fuzzy-World in their antecedents, feature libraries, rule counts, and fitting configurations. Fuzzy-World routes on physical states and fits weighted local corrections to a shared map, using structured features for Pendulum and advection and support-aware shrinkage for the MuJoCo tasks. The main comparison therefore evaluates complete model configurations. In contrast, the matched Top-1 and Dense ablations in Section 4.3 preserve the Fuzzy-World antecedents, feature library, and fitting hyperparameters, and refit the local corrections under the modified activation budget. These matched variants isolate activation-budget effects.

F Local Rule Interpretation and Empirical Validity

F.1 Semantic Regions and Local Equations

A rule combines a state region with an increment equation. Gaussian centers c_k and widths σ_k define the antecedent in named physical coordinates, and $f_k(p_t, a_t) = A_k\phi(p_t, a_t) + b_k$ maps state–action features to an increment. In Pendulum, $p_t = (\theta_t, \omega_t)$ contains angle and angular velocity, and the scalar action u_t is torque. Converting coefficients to physical units makes their contributions directly inspectable. Gaussian widths describe fuzzy regions, not hard boundaries or confidence intervals; rule execution also depends on the competing activations.

F.2 Coverage, Support, and Local Accuracy

For N_{eval} held-out transitions, let $i \in \{1, \dots, N_{\text{eval}}\}$ index an evaluated transition and $w_{ik} = w_k(p_i)$ be the executed weight of rule k. Its mean activation mass m_k and Top-1 sample coverage γ_k are

$$m_k = \frac{1}{N_{\text{eval}}} \sum_{i=1}^{N_{\text{eval}}} w_{ik}, \qquad \gamma_k = \frac{1}{N_{\text{eval}}} \sum_{i=1}^{N_{\text{eval}}} \mathbf{1}[k = \arg\max_j w_{ij}]. \tag{27}$$

The held-out effective sample size is recomputed on these transitions as $n_{\text{eff},k}^{\text{eval}} = (\sum_i w_{ik})^2/(\sum_i w_{ik}^2 + \epsilon_n)$, following Equation 10. For an activated rule, define the across-sample normalized weight $\tilde{w}_{ik} = w_{ik}/\sum_{i=1}^{N_{\text{eval}}} w_{ik}$ and weighted target mean $\bar{\Delta p}_k = \sum_{i=1}^{N_{\text{eval}}} \tilde{w}_{ik}\Delta p_i$. The local coefficient of determination is

$$R_k^2 = 1 - \frac{\sum_{i=1}^{N_{\text{eval}}} \tilde{w}_{ik}\|\Delta p_i - f_k(p_i, a_i)\|_2^2}{\max(\sum_{i=1}^{N_{\text{eval}}} \tilde{w}_{ik}\|\Delta p_i - \bar{\Delta p}_k\|_2^2, \epsilon_R)}, \tag{28}$$

where $\epsilon_R > 0$ stabilizes the variance denominator. This compares each equation with true increments under its actual activation weights. Supported activation mass is the sum of m_k over supported rules. Support thresholds establish data sufficiency for the audit, not correctness in unobserved regions.

Low Top-1 coverage for an individual rule is compatible with specialization to a limited state region. Coverage should therefore be considered together with activation mass and effective sample size. The supported-rule fraction measures the share of the library meeting both thresholds, whereas supported activation mass measures the share of prediction weight carried by those rules. Support is computed on held-out transitions and describes the sample basis of the audit, rather than proving that training data cover all OOD regions.

Local R_k^2 evaluates one complete consequent against observed increments where that rule is activated. It differs from the aggregate model's accuracy and from the 50-step physical-trajectory R^2. A negative value indicates greater weighted squared error than the corresponding constant-mean predictor. Retaining medians, activation-weighted means, and negative fractions distinguishes typical rule accuracy from frequently used rules and local failures. Five seeds contribute 240, 80, 160, and 90 rule records per split for Pendulum, HalfCheetah, advection, and Hopper, respectively, totaling 1,140 across ID and OOD. These are pooled records, not the number of rules in a single model.

F.3 Representative IF–THEN Rules

The following equations and execution example use the same saved Pendulum run from the five-seed evaluation. The rule interpretation analysis illustrates how named physical variables and executed local equations contribute to individual predictions. Table 6 presents the three rules with the largest

OOD activation mass in this Pendulum run. Their OOD effective sample sizes are 150.18, 94.48, and 81.21 for P47, P45, and P27, respectively, all above the threshold of 11. Full-rule local R^2 values are 0.99862, 0.99937, and 0.99899. Gaussian centers and widths identify each antecedent, while polynomial consequents expose how angle, angular velocity, and action enter the predicted increment. These usage-selected examples illustrate individual equations, while Section 4.2 evaluates the complete rule library.

Table 6: Pendulum rules from one of the five runs, selected by OOD activation mass. Parentheses give Top-1 sample coverage and full-rule local R^2. The IF entries denote Gaussian centers and widths, not hard intervals. Consequents are shown in physical units with coefficients below 0.005 omitted and remaining coefficients rounded; displayed equations are approximations.

Rule and IF state region	THEN physical increment
P47 $(6.2\%, 0.999)$ $\theta = -2.851 \pm 0.427$ $\omega = 0.092 \pm 0.784$	$\Delta\theta \approx -0.020 + 0.020\theta + 0.050\omega + 0.006u$ $\Delta\omega \approx -0.405 + 0.402\theta + 0.098\theta^2 - 0.023\theta^3 + (0.116 - 0.012\theta)u$
P45 $(3.9\%, 0.999)$ $\theta = 2.853 \pm 0.428$ $\omega = -0.168 \pm 0.778$	$\Delta\theta \approx 0.020 + 0.020\theta - 0.005\theta^2 + 0.050\omega + 0.008u$ $\Delta\omega \approx 0.402 + 0.409\theta - 0.102\theta^2 - 0.022\theta^3 + 0.152u$
P27 $(3.4\%, 0.999)$ $\theta = -2.912 \pm 0.428$ $\omega = -1.176 \pm 0.843$	$\Delta\theta \approx -0.013 + 0.022\theta + 0.050\omega + 0.007u$ $\Delta\omega \approx -0.263 + 0.444\theta + 0.079\theta^2 - 0.029\theta^3 + 0.143u$

A held-out prediction illustrates how the equations enter execution. For the saved first OOD transition of the same Pendulum run, $(\theta_t, \omega_t, u_t) = (-2.70687, 0.58935, -1.72325)$, rules P2 and P47 receive weights 0.31979 and 0.68021. Evaluating their full, unrounded equations gives

$$\begin{aligned} \Delta\hat{p}_t &\approx 0.31979 \begin{bmatrix} 0.00094454 \\ -0.570456 \end{bmatrix} + 0.68021 \begin{bmatrix} 0.00078125 \\ -0.573722 \end{bmatrix} \\ &\approx \begin{bmatrix} 0.00083346 \\ -0.572678 \end{bmatrix}, \qquad \Delta p_t = \begin{bmatrix} 0.00074959 \\ -0.574357 \end{bmatrix}. \end{aligned} \tag{29}$$

Each displayed contribution is part of the physical increment actually computed by the model. The example was selected by its position in the held-out sequence, without filtering by prediction error. It illustrates computational traceability; the audit in Section 4.2 supplies the broader evidence of local predictive validity. Truncated equations in Table 6 are for presentation only and are not used to compute the reported scores or this decomposition.

The term $(0.116 - 0.012\theta)u$ in P47 makes its fitted action response depend on angle, while 0.050ω exposes the angular-velocity contribution to the angle increment. These coefficients describe learned local predictive relationships; they are not identified physical constants or intervention-based causal effects. Likewise, the two-rule decomposition explains the physical increment. An optional neural residual decoder can also affect reconstructed observations, so its computations are outside this symbolic decomposition.

G Sensitivity to the Active-Rule Budget

G.1 Matched Top-q Protocol

Let q denote the maximum number of active rules, distinct from the rule index k. We varied q from one to eight and included dense aggregation, which uses all stored rules. Each setting used the same antecedent centers and widths as the corresponding Fuzzy-World checkpoint, without reclustering. The stored rule counts were 48, 16, 32, and 18 for Pendulum, HalfCheetah, advection, and Hopper. Feature libraries, Ridge penalties, shared dynamics, and support-dependent shrinkage settings were held fixed. For each activation budget, the highest-weight rules were retained and renormalized, and local residuals were refitted with those weights. Thus, training and inference used the same routing scheme.

All settings followed the main data splits, five seeds, observation reconstruction, and rollout evaluation. Advection retained its 50-epoch residual observation-learning stage with fixed physical dynamics; the other tasks retained their analytic or identity observation maps. The comparison contains 180 task–setting–seed records. The activation-budget experiments were run separately from the main comparison, including a Top-2 rerun under the same configuration. Dataset hashes matched the reference runs. This sweep characterizes sensitivity on the existing evaluation splits; it does not replace the original Top-2 configuration with a test-selected setting.

G.2 Prediction with Few Active Rules

Tables 7 and 8 report the complete ID and OOD results, and Figure 5 shows how OOD rollout error changes with the activation budget. To expose within-seed differences, the figure expresses each Top-q error as a percentage change from its matched dense control before averaging across seeds. Absolute errors and sample standard deviations remain available in the table.

A small number of active rules retained competitive prediction across all four systems, and using more rules did not consistently improve performance. Pendulum reached its lowest mean OOD observation rollout MSE with Top-1 (0.010283), while its highest mean physical-trajectory R^2 occurred with Top-3 (0.971422). Beyond Top-3, observation error approached the dense result with little further change. HalfCheetah achieved its lowest mean OOD rollout MSE with Top-2 (2.17002); all nine settings lay between 2.17002 and 2.17517. Advection was similarly insensitive to larger activation budgets, with Top-1 yielding the lowest mean OOD rollout MSE (0.179209) and the highest physical-trajectory R^2 (0.995493).

Hopper showed the clearest benefit from combining a few local equations. Mean OOD rollout MSE decreased from 1.03068 with Top-1 to 0.874999 with Top-2 and 0.790688 with Top-3. It then increased across Top-4–8 toward the dense value of 0.931748. Top-3 reduced mean error by 15.1% relative to dense aggregation; Top-4 achieved the highest mean physical-trajectory R^2 (0.608929). Thus, neither a single rule nor the full library gave the best mean prediction on this task.

G.3 Why Retain Top-2?

Top-2 permits graded cooperation between two local equations while keeping the execution trace to at most two contributions. Larger budgets improve some tasks, as Top-3/4 do on Hopper, but offer no consistent advantage across systems. Top-2 therefore balances local predictive flexibility and concise explanation; it is not a universally optimal rule count. Here interpretability is characterized by explicit contributions and their number, rather than a human-user evaluation.

Table 7: Complete matched activation-budget sweep: ID results. Values are mean ± sample standard deviation over five seeds. MSE evaluates standardized observations (lower is better); trajectory R^2 evaluates physical states (higher is better). Bold denotes the best unrounded mean per task and metric. Dense uses all stored rules.

Activation	One-step MSE	Rollout-50 MSE	Trajectory R^2
Pendulum			
Top-1	$(2.450 \pm 0.860) \times 10^{-6}$	$(6.855 \pm 9.669) \times 10^{-3}$	0.9795 ± 0.0073
Top-2	$(2.119 \pm 0.910) \times 10^{-6}$	$(7.127 \pm 9.204) \times 10^{-3}$	0.9767 ± 0.0091
Top-3	$\mathbf{(2.041 \pm 0.900) \times 10^{-6}}$	$\mathbf{(5.223 \pm 4.795) \times 10^{-3}}$	0.9800 ± 0.0078
Top-4	$(2.041 \pm 0.898) \times 10^{-6}$	$(5.312 \pm 4.895) \times 10^{-3}$	$\mathbf{0.9806 \pm 0.0071}$
Top-5	$(2.043 \pm 0.898) \times 10^{-6}$	$(5.331 \pm 4.916) \times 10^{-3}$	0.9806 ± 0.0071
Top-6	$(2.043 \pm 0.898) \times 10^{-6}$	$(5.335 \pm 4.921) \times 10^{-3}$	0.9806 ± 0.0071
Top-7	$(2.043 \pm 0.898) \times 10^{-6}$	$(5.336 \pm 4.923) \times 10^{-3}$	0.9806 ± 0.0071
Top-8	$(2.043 \pm 0.898) \times 10^{-6}$	$(5.336 \pm 4.923) \times 10^{-3}$	0.9806 ± 0.0071
Dense	$(2.043 \pm 0.898) \times 10^{-6}$	$(5.337 \pm 4.923) \times 10^{-3}$	0.9806 ± 0.0071
HalfCheetah			
Top-1	0.0306 ± 0.0015	0.0889 ± 0.0027	0.9111 ± 0.0028
Top-2	0.0294 ± 0.0014	0.0868 ± 0.0030	0.9132 ± 0.0028
Top-3	$\mathbf{0.0292 \pm 0.0013}$	$\mathbf{0.0864 \pm 0.0031}$	$\mathbf{0.9136 \pm 0.0028}$
Top-4	0.0293 ± 0.0013	0.0869 ± 0.0029	0.9131 ± 0.0027
Top-5	0.0295 ± 0.0013	0.0871 ± 0.0030	0.9129 ± 0.0027
Top-6	0.0297 ± 0.0013	0.0874 ± 0.0027	0.9126 ± 0.0025
Top-7	0.0300 ± 0.0013	0.0880 ± 0.0028	0.9120 ± 0.0025
Top-8	0.0302 ± 0.0013	0.0885 ± 0.0029	0.9115 ± 0.0026
Dense	0.0313 ± 0.0014	0.0910 ± 0.0030	0.9091 ± 0.0028
1D Advection			
Top-1	$(7.607 \pm 0.416) \times 10^{-4}$	0.1548 ± 0.0298	$\mathbf{1.0000 \pm 0.0000}$
Top-2	$(7.606 \pm 0.416) \times 10^{-4}$	0.1548 ± 0.0298	1.0000 ± 0.0000
Top-3	$(7.607 \pm 0.417) \times 10^{-4}$	$\mathbf{0.1548 \pm 0.0298}$	1.0000 ± 0.0000
Top-4	$(7.606 \pm 0.417) \times 10^{-4}$	0.1548 ± 0.0298	1.0000 ± 0.0000
Top-5	$(7.606 \pm 0.416) \times 10^{-4}$	0.1548 ± 0.0298	1.0000 ± 0.0000
Top-6	$(7.606 \pm 0.416) \times 10^{-4}$	0.1548 ± 0.0298	1.0000 ± 0.0000
Top-7	$\mathbf{(7.606 \pm 0.416) \times 10^{-4}}$	0.1548 ± 0.0298	1.0000 ± 0.0000
Top-8	$(7.606 \pm 0.416) \times 10^{-4}$	0.1548 ± 0.0298	1.0000 ± 0.0000
Dense	$(7.606 \pm 0.416) \times 10^{-4}$	0.1548 ± 0.0298	1.0000 ± 0.0000
Hopper			
Top-1	0.0117 ± 0.0022	0.2086 ± 0.0424	0.7721 ± 0.0420
Top-2	$\mathbf{0.0101 \pm 0.0020}$	0.1782 ± 0.0349	0.8090 ± 0.0392
Top-3	0.0102 ± 0.0020	$\mathbf{0.1718 \pm 0.0309}$	0.8221 ± 0.0303
Top-4	0.0103 ± 0.0021	0.1727 ± 0.0350	$\mathbf{0.8241 \pm 0.0272}$
Top-5	0.0105 ± 0.0021	0.1807 ± 0.0389	0.8157 ± 0.0280
Top-6	0.0107 ± 0.0022	0.1858 ± 0.0401	0.8089 ± 0.0301
Top-7	0.0108 ± 0.0022	0.1898 ± 0.0437	0.8051 ± 0.0311
Top-8	0.0110 ± 0.0022	0.1964 ± 0.0455	0.7974 ± 0.0314
Dense	0.0117 ± 0.0022	0.2260 ± 0.0469	0.7627 ± 0.0372

Table 8: Complete matched activation-budget sweep: OOD results. Values are mean ± sample standard deviation over five seeds. MSE evaluates standardized observations (lower is better); trajectory R^2 evaluates physical states (higher is better). Bold denotes the best unrounded mean per task and metric. Dense uses all stored rules.

Activation	One-step MSE	Rollout-50 MSE	Trajectory R^2
Pendulum			
Top-1	$(8.956 \pm 6.206) \times 10^{-6}$	$\mathbf{0.0103 \pm 0.0083}$	0.9702 ± 0.0131
Top-2	$(8.622 \pm 6.215) \times 10^{-6}$	0.0104 ± 0.0088	0.9707 ± 0.0121
Top-3	$\mathbf{(8.621 \pm 6.216) \times 10^{-6}}$	0.0107 ± 0.0089	$\mathbf{0.9714 \pm 0.0132}$
Top-4	$(8.633 \pm 6.222) \times 10^{-6}$	0.0108 ± 0.0090	0.9714 ± 0.0132
Top-5	$(8.637 \pm 6.225) \times 10^{-6}$	0.0108 ± 0.0090	0.9714 ± 0.0132
Top-6	$(8.637 \pm 6.225) \times 10^{-6}$	0.0108 ± 0.0090	0.9714 ± 0.0132
Top-7	$(8.638 \pm 6.225) \times 10^{-6}$	0.0108 ± 0.0090	0.9714 ± 0.0132
Top-8	$(8.638 \pm 6.225) \times 10^{-6}$	0.0108 ± 0.0090	0.9714 ± 0.0132
Dense	$(8.638 \pm 6.225) \times 10^{-6}$	0.0108 ± 0.0090	0.9714 ± 0.0132
HalfCheetah			
Top-1	1.8490 ± 1.0564	2.1725 ± 0.6870	0.5551 ± 0.1037
Top-2	1.7793 ± 0.9867	$\mathbf{2.1700 \pm 0.6954}$	$\mathbf{0.5558 \pm 0.1060}$
Top-3	1.7479 ± 0.9581	2.1725 ± 0.6966	0.5553 ± 0.1063
Top-4	1.7174 ± 0.9315	2.1730 ± 0.6960	0.5552 ± 0.1061
Top-5	1.6700 ± 0.8833	2.1738 ± 0.6950	0.5550 ± 0.1058
Top-6	1.6438 ± 0.8600	2.1752 ± 0.6949	0.5547 ± 0.1058
Top-7	1.6176 ± 0.8403	2.1749 ± 0.6950	0.5547 ± 0.1058
Top-8	1.6020 ± 0.8267	2.1749 ± 0.6945	0.5547 ± 0.1057
Dense	$\mathbf{1.5027 \pm 0.7110}$	2.1741 ± 0.6935	0.5549 ± 0.1054
1D Advection			
Top-1	$(1.501 \pm 0.216) \times 10^{-3}$	$\mathbf{0.1792 \pm 0.0270}$	$\mathbf{0.9955 \pm 0.0007}$
Top-2	$(1.501 \pm 0.219) \times 10^{-3}$	0.1797 ± 0.0273	0.9952 ± 0.0010
Top-3	$\mathbf{(1.500 \pm 0.217) \times 10^{-3}}$	0.1797 ± 0.0272	0.9952 ± 0.0013
Top-4	$(1.500 \pm 0.217) \times 10^{-3}$	0.1797 ± 0.0272	0.9952 ± 0.0013
Top-5	$(1.501 \pm 0.218) \times 10^{-3}$	0.1798 ± 0.0272	0.9952 ± 0.0013
Top-6	$(1.501 \pm 0.217) \times 10^{-3}$	0.1798 ± 0.0272	0.9952 ± 0.0014
Top-7	$(1.501 \pm 0.217) \times 10^{-3}$	0.1798 ± 0.0272	0.9952 ± 0.0014
Top-8	$(1.501 \pm 0.218) \times 10^{-3}$	0.1798 ± 0.0272	0.9952 ± 0.0014
Dense	$(1.501 \pm 0.219) \times 10^{-3}$	0.1798 ± 0.0272	0.9952 ± 0.0013
Hopper			
Top-1	0.0653 ± 0.0208	1.0307 ± 0.2202	0.4211 ± 0.2502
Top-2	0.0593 ± 0.0196	0.8750 ± 0.2073	0.5222 ± 0.2529
Top-3	0.0586 ± 0.0209	$\mathbf{0.7907 \pm 0.1212}$	0.6065 ± 0.0886
Top-4	0.0567 ± 0.0181	0.7929 ± 0.1140	$\mathbf{0.6089 \pm 0.0662}$
Top-5	0.0574 ± 0.0179	0.7987 ± 0.1078	0.6071 ± 0.0479
Top-6	0.0571 ± 0.0167	0.8117 ± 0.1082	0.6003 ± 0.0499
Top-7	0.0570 ± 0.0163	0.8239 ± 0.1210	0.5931 ± 0.0364
Top-8	0.0570 ± 0.0157	0.8381 ± 0.1262	0.5858 ± 0.0338
Dense	$\mathbf{0.0566 \pm 0.0123}$	0.9317 ± 0.1505	0.5388 ± 0.0290

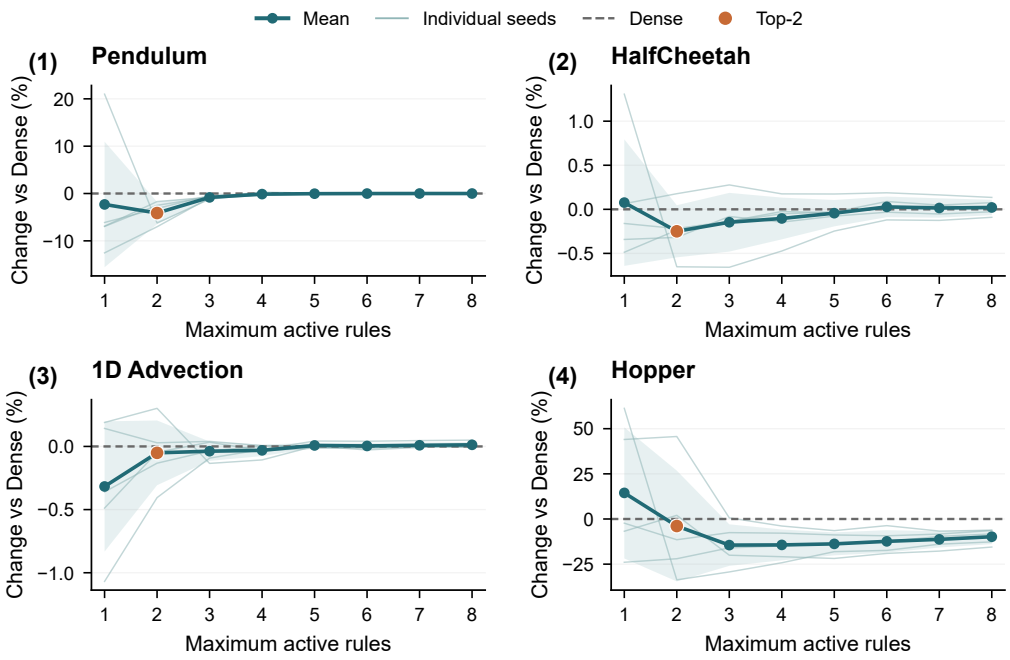


Figure 5: OOD rollout error versus active-rule budget. Each seed's percentage change is $100(E_{q,s}/E_{\text{dense},s} - 1)$, where E is observation MSE, q the activation budget, and s the seed. Thin lines show five seeds; the thick line and shading show their mean and sample standard deviation. The dashed line is Dense; orange marks Top-2. Lower is better; vertical scales differ. Plotted means of ratios differ from ratios of table means.