ALLOCATION ENTROPY FOR SINGLE-PASS RISK TRIAGE IN NEURAL OPERATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

The same predicted field can receive a different risk decision when its allocation is read in another basis. Allocation Uncertainty Decomposition (AUD) makes this representation effect measurable in one pass: it reads threshold entropy from wavelet, Fourier, and branch–trunk components, and energy dispersion from selective updates, without changing the predicted field or training a risk head. For threshold adapters, entropy equals conditional uncertainty in component membership under fixed-threshold logistic perturbations; held-out data connect that structural quantity to physical error. On five frozen Haar-trained models, switching the readout to db2 changes 22.5% of accept/reject decisions at 80% coverage. Crossed training finds EM matching gains and a Darcy preference for Haar under either training basis. Four electromagnetic adapters rank in-domain error with five-seed mean correlations of 0.238–0.492; newly trained MWNO and FNO models retain correlations near 0.31–0.32 on advection–diffusion. AUD improves screening over random selection in an independent label-budget test, while inexpensive input controls screen more strongly. At batch 96 on one RTX 4090, prediction plus AUD takes 0.513 ms/sample versus 5.070 for ten-pass Monte Carlo dropout. These results identify representation choice as a measurable axis of single-pass risk triage.

1 INTRODUCTION

Neural operators (Li et al., 2021; Kovachki et al., 2023; Tripura & Chakraborty, 2023; Tran et al., 2023; Li et al., 2023) make repeated partial differential equation (PDE) solves cheap; deployment still must decide which predicted fields deserve expensive verification. In design optimization or inverse problems, an inaccurate field can steer the next query before a high-fidelity solver runs. Deep ensembles (Lakshminarayanan et al., 2017), Monte Carlo (MC) dropout (Gal & Ghahramani, 2016), Laplace approximations (Daxberger et al., 2021), and probabilistic operators (Garg & Chakraborty, 2025a; Bülte et al., 2025) assess risk at different computation and supervision costs. We examine what the operator's representation reveals during its prediction pass.

Structured components make that evidence accessible. Wavelet neural operators (WNOs) represent detail across locations and scales; Fourier neural operators (FNOs) use spectral modes; DeepONets combine branch and trunk basis components. A threshold separates large and small components in each representation. Near-threshold components can change membership under a small perturbation; soft-threshold entropy measures this ambiguity for a direct test against prediction error.

Allocation Uncertainty Decomposition (AUD) converts native operator components into a scalar readout with no trainable risk-prediction parameters. For threshold adapters, the score summarizes proximity to the allocation boundary; fixed-threshold logistic perturbations give that entropy an exact conditional-membership interpretation. For selective state-space updates, the readout instead measures update-energy dispersion. Held-out validation fixes the score direction and checks error ranking or detection of a named distribution shift.

The allocation adapter determines how the readout relates to the predictor. Masked wavelet neural operator (MWNO) masks and latent Mamba operator (LaMO) selective updates participate in prediction; we call these allocations prediction-coupled. Dense WNO, FNO, and Fourier-DeepONet expose prediction-preserving decompositions. Enabling the entropy readout preserves the predicted field in all five adapters. This separation allows controlled changes to the scoring representation on a frozen predictor.

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.

Readout representation changes which predictions pass a fixed-coverage screening rule, even when every predicted field stays fixed: a Haar-to-db2 swap changes $22.5 \pm 8.0\%$ of decisions at 80% coverage (Section 5). A crossed training- and readout-basis experiment separates matching with the trained representation from a general readout preference: electromagnetic (EM) scattering favors matching by 0.00066 ± 0.00020 in normalized retained error, while controlled Darcy favors Haar under either training basis. Matched EM adapters rank error across wavelet, Fourier, and branch–trunk readouts. In a separate fixed-prediction control, randomizing only the score extractor reduces ρ_s from 0.507 to 0.313. Darcy, Burgers, Navier–Stokes, and Maxwell-3D test readout behavior across physical regimes. Newly trained MWNO and FNO predictors on controlled advection–diffusion provide an independent prospective transfer test, followed by a capacity-matched FNO control. Comparisons with input statistics and learned heads locate the zero-head readout on a measured quality–cost landscape. The independent EM label-budget and advection–diffusion tests identify task-specific input statistics as stronger screening references. The fixed-prediction and crossed-training controls establish representation choice as a distinct variable in risk triage, and the input comparisons quantify its practical value.

Contributions. We present:

- **A single-pass allocation readout.** AUD converts native operator components into a risk score with no trainable head, preserves the predictor's output, and links threshold entropy to component-decision sensitivity.
- **Representation-dependent triage.** Fixed-prediction controls measure changes in acceptance decisions; crossed training shows that readout preference depends on the physical task, and signed perturbations distinguish instability from task error.
- **Five architecture adapters and new-task transfer.** MWNO, dense WNO, FNO, Fourier-DeepONet, and LaMO span wavelet, Fourier, branch–trunk, and state-space representations; new advection–diffusion models test transfer beyond the original physical tasks.
- **Controlled triage comparisons.** Matched timing and shared-label-budget tests compare allocation entropy with input statistics, learned heads, and multi-pass estimators for specific deployment targets.

2 Related Work

Neural operators and operator uncertainty quantification (UQ). FNO (Li et al., 2021) established spectral operator learning, later variants addressed geometry and factorization (Li et al., 2023; Tran et al., 2023), wavelet operators improved locality for interface-dominated solutions (Tripura & Chakraborty, 2023; Gupta et al., 2021), DeepONet introduced branch–trunk decomposition (Lu et al., 2021), and LaMO introduced a latent state-space operator for continuous long-range PDE dependencies (Tiwari et al., 2025). Operator-specific UQ methods include probabilistic neural data fusion (Mora et al., 2023), probabilistic neural operators (Bülte et al., 2025), randomized-prior WNOs (Garg & Chakraborty, 2025a), multi-fidelity WNOs (Thakur et al., 2022), and conformalized randomized-prior operators (Garg & Chakraborty, 2025b). DiverseNO obtains predictive disagreement through multiple final heads, while Operator-ProbConserv uses calibrated uncertainty to improve out-of-distribution (OOD) predictions and physical conservation (Mouli et al., 2024). LUNO linearizes a trained neural operator into a function-valued Gaussian-process belief for post-hoc Bayesian UQ (Magnani et al., 2025). Approximation and numerical analyses characterize solution-operator error through architecture, sampling, regularity, stability, and training effects (De Ryck & Mishra, 2022; 2024). AUD tests a different source of risk information: the operator's own basis allocation, read in one pass and evaluated against a named target.

Efficient and deterministic uncertainty estimation. Deep ensembles and MC-dropout remain strong baselines (Lakshminarayanan et al., 2017; Gal & Ghahramani, 2016), but their cost scales with the number of passes. Deterministic alternatives such as evidential learning, direct variance heads, posterior networks, energy scores, and spectral-normalized neural Gaussian processes (SNGP) (Sensoy et al., 2018; Van Amersfoort et al., 2020; Mukhoti et al., 2023; Charpentier et al., 2020; Liu et al., 2020b;a) obtain risk information through output statistics or fitted uncertainty modules. AUD extracts a rank statistic from a predictor's existing basis decomposition, then uses held-out validation to choose a selective policy. Calibration methods additionally map scores to predictive uncertainty (Ovadia et al., 2019; Kuleshov et al., 2018).

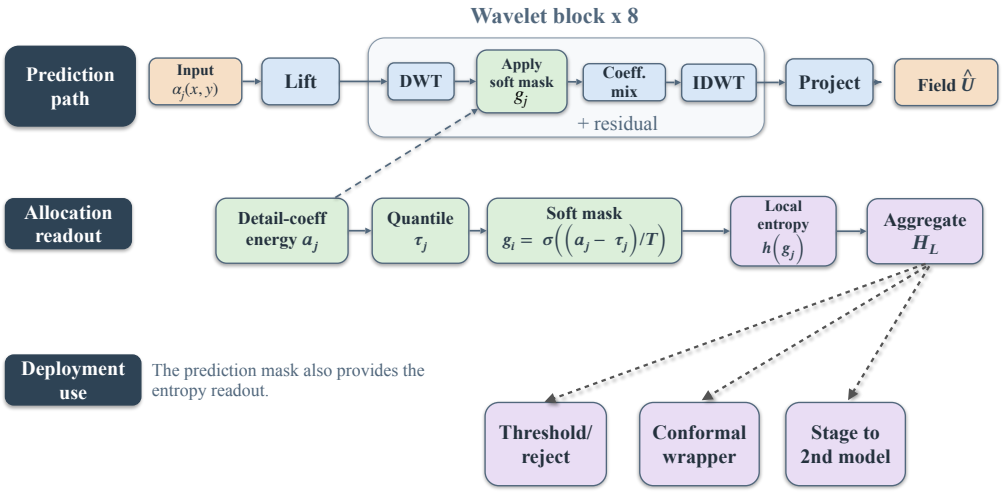

Figure 1: MWNO instantiates AUD's predictor–allocation–readout–policy pipeline. Its wavelet masks take part in prediction, while the entropy readout leaves the field fixed. Top: lift, repeated wavelet blocks, and field projection. Middle: detail energy induces a sample-wise threshold and soft mask; local entropy aggregates into H_L. Bottom: H_L supports thresholding, conformal wrapping, and staged escalation.

Routing and allocation uncertainty. Mixture-of-experts (MoE) models (Shazeer et al., 2017; Fedus et al., 2022; Lepikhin et al., 2021; Du et al., 2022) expose expert-routing signals; layer ensembles estimate uncertainty from internal predictive diversity (Kushibar et al., 2022), and learned MoE gate entropy has served as routing uncertainty for segmentation (Pavlitska et al., 2025). AUD reads energy allocation among indexed operator-basis components, including components that leave prediction unchanged. Fixed-prediction swaps isolate the readout representation's effect on risk ranking, while matched layer probes quantify ranking and inference cost.

3 Method

3.1 Problem Setup

Let $\mathcal{F}_\theta : \mathcal{A} \to \mathcal{U}$ be a neural operator mapping PDE coefficients $a \in \mathcal{A}$ to solutions $u \in \mathcal{U}$. For each sample, we require both a prediction and a scalar triage score from one forward pass:

$$\hat{u} = \mathcal{F}_\theta(a), \quad S(a) \in \mathbb{R}. \tag{1}$$

The score supplies an ordering for selective prediction. We evaluate error ranking and detection of each named distribution shift separately, then measure the inference cost of the resulting policy.

3.2 Allocation Uncertainty Decomposition

AUD uses allocation entropy to describe how an operator distributes magnitude across its components. For threshold-based adapters, this description measures proximity to a change in active-set membership. For block b, let $\mathcal{C}_b$ be a finite, architecture-defined set of components and let $z_{b,c}(a)$ be the representation assigned to component $c \in \mathcal{C}_b$. The component index has a physical or operator-basis interpretation: location–scale cells for WNOs, retained Fourier coefficients for FNOs, branch–trunk basis products for DeepONets, and direction–state update groups for selective state-space operators. Threshold-margin adapters use component energy $\to$ threshold margin $\to$ soft active probability $\to$ allocation entropy. We define a nonnegative component magnitude, sample-wise threshold, threshold

margin, and soft active probability as

$$\begin{aligned} E_{b,c}(a) &= \Phi_b(z_{b,c}(a)) \geq 0, \\ \tau_b(a) &= \mathrm{Quantile}_{1-r_g}\{E_{b,c}(a) : c \in \mathcal{C}_b\}, \end{aligned} \tag{2}$$

$$m_{b,c}(a) = E_{b,c}(a) - \tau_b(a), \qquad p_{b,c}(a) = \sigma(m_{b,c}(a)/T)\,. \tag{3}$$

Here $r_g \in (0,1)$ is the retained fraction, $T > 0$ is the soft-gate temperature, and σ is the logistic sigmoid. Fix the input and threshold, draw independent $\xi_{b,c} \sim \mathrm{Logistic}(0,T)$, and set $A^{\xi}_{b,c} = \mathbb{1}\{m_{b,c} + \xi_{b,c} \geq 0\}$. Logistic symmetry gives

$$\Pr_{\xi}(A^{\xi}_{b,c} = 1 \mid a) = \sigma(m_{b,c}/T) = p_{b,c}, \qquad H_{\xi}(A^{\xi}_{b,c} \mid a) = \mathsf{h}(p_{b,c}). \tag{4}$$

This interprets the deterministic gate: T is an architectural temperature, not a fitted noise scale, and the uncertainty concerns allocation rather than weights or physical fields. For fixed component energies, increasing r_g lowers the quantile threshold and expands the hard active set $\{c : m_{b,c} \geq 0\}$, up to ties. Changing T leaves that hard set fixed but changes soft-mask values and entropy; larger T moves probabilities toward $1/2$. The adapter fixes Φ_b: detail-subband norm for wavelets, channel-mean Fourier-coefficient magnitude for FNO, and mean absolute basis contribution for DeepONet. The block-level entropy readout is

$$\begin{aligned} H_b(a) &= \frac{1}{|\mathcal{C}_b|} \sum_{c \in \mathcal{C}_b} \mathsf{h}(p_{b,c}(a))\,, \\ \mathsf{h}(p) &= -p \log p - (1-p)\log(1-p), \end{aligned} \tag{5}$$

With independent perturbations, $H_{\xi}((A^{\xi}_{b,c})_c \mid a) = \sum_c \mathsf{h}(p_{b,c}) = |\mathcal{C}_b| H_b(a)$: AUD decomposes this conditional allocation entropy across components. Each term falls as its absolute margin grows relative to T. For a continuous allocation such as a selective state-space update, the adapter instead normalizes nonnegative component energy, with $\eta > 0$ a numerical stabilizer,

$$\begin{aligned} q_{b,c} &= \frac{E_{b,c}}{\sum_{c'} E_{b,c'} + \eta}, \\ H^{\mathrm{dist}}_b &= -\frac{\sum_c q_{b,c} \log(q_{b,c} + \eta)}{\log |\mathcal{C}_b|}. \end{aligned} \tag{6}$$

and validation selects among step-size, state-coupling, and combined update-energy readouts. Because $\sum_c q_{b,c} \leq 1$, this readout indexes energy dispersion rather than threshold-perturbation uncertainty. The readout computes these entropy statistics during the prediction pass.

For readout blocks $\mathcal{B}$, the raw entropy $H(a) = |\mathcal{B}|^{-1} \sum_b H_b(a)$ describes allocation geometry (the wavelet adapter pools all scale–location cells within each equal-sized block, as in Eq. (15)); a deployment score requires a declared target. For in-distribution (ID) error ranking and selective prediction (Geifman & El-Yaniv, 2017), validation fixes

$$S_{\mathrm{rank}}(a) = sH(a), \qquad s \in \{-1, +1\}. \tag{7}$$

For OOD detection, an independent ID/OOD validation split chooses between the oriented score and the standardized two-sided deviation

$$S_{\mathrm{dev}}(a) = \frac{|H(a) - \mathrm{median}_{\mathrm{val}}\, H|}{\mathrm{MAD}_{\mathrm{val}}(H) + \eta}. \tag{8}$$

Here MAD denotes the median absolute deviation on validation data. Validation fixes the score family, orientation, and threshold before test evaluation. AUD treats ranking and OOD detection as distinct targets because a feature may order errors within a regime while shifting in either direction across regimes.

Entropy-scaled conformal prediction uses disjoint scale-fitting and quantile-calibration samples (Appendix G).

3.3 Architecture Adapters and Predictor Factorization

An internal readout can change while the field prediction stays fixed. AUD makes this separation explicit through four objects:

$$\underbrace{\mathcal{F}_\theta}_{\text{predictor}} \to \underbrace{\{z_{b,c}\}}_{\text{allocation adapter}} \to \underbrace{H}_{\text{zero-head readout}} \to \underbrace{S_{\mathcal{T}}}_{\text{validation-fixed policy}}. \tag{9}$$

The predictor produces the physical field; the adapter exposes its components, the readout measures allocation ambiguity, and validation checks the intended risk target. In MWNO, the energy masks are part of $\mathcal{F}_\theta$, making the allocation prediction-coupled. The wavelet transforms, coefficient mixing, and residual path form the backbone; the deterministic masks modify its prediction, while the entropy readout adds a scalar diagnostic. Dense WNO applies a two-level discrete wavelet transform (DWT) to each block feature and reads the three detail subbands without applying the masks. FNO reads retained Fourier coefficients; Fourier-DeepONet reads branch–trunk contributions before summation. LaMO reads selective update energies; validation chooses among step-size, state-coupling, and combined distributions. Appendix F gives component dimensions and extraction details. Dense WNO, FNO, and Fourier-DeepONet are prediction-preserving; MWNO and LaMO read allocations that participate in prediction, while enabling any entropy readout leaves $\hat{u}$ unchanged. For the preserving adapters, the "active set" indexes large components in the readout decomposition while the predictor retains the full component set.

3.4 Wavelet Operator Instantiation

MWNO-AUD stacks eight wavelet blocks, each with a discrete wavelet transform, deterministic energy masks, coefficient mixing, inverse transform, and residual addition. The masks participate in prediction; averaging their binary entropy over blocks, scales, and cells defines H_L. The EM deployment track fixes $r_g = 0.15$ and $T = 0.2$ before testing; Appendix A gives the full construction, and Appendix I reports sensitivity across physical regimes.

3.5 Theory Separates Mechanism from Risk

Equation (4) identifies the random quantity measured by threshold entropy. For one additive hard-allocation block $v(A) = \sum_c A_c z_c$ followed by an L_D-Lipschitz decoder D, independent logistic margin perturbations induce $Y_\xi = D(v(A^\xi))$ with

$$\underbrace{\mathbb{E}_\xi \|Y_\xi - \mathbb{E}_\xi Y_\xi\|_2^2}_{V_\xi(a)} \le L_D^2 \sum_c \|z_c\|_2^2 p_c(1-p_c) \le \frac{L_D^2}{2\log 2} \sum_c \|z_c\|_2^2 \mathsf{h}(p_c). \tag{10}$$

The same assumptions connect the soft-gate output $Y_p = D(\sum_c p_c z_c)$ to this surrogate:

$$\|Y_p - \mathbb{E}_\xi Y_\xi\|_2 \le L_D \Big(\sum_c \|z_c\|_2^2 p_c(1-p_c) \Big)^{1/2}. \tag{11}$$

Appendix B proves both bounds. The identity $\mathbb{E}_\xi \|Y_\xi - u\|_2^2 = \|\mathbb{E}_\xi Y_\xi - u\|_2^2 + V_\xi(a)$ separates prediction-path sensitivity from the physical residual. The bounds weight components by $\|z_c\|_2^2$; AUD averages their entropies. Held-out data fix the error-ranking direction, and MWNO raw entropy tracks allocation flips positively but field error negatively (Section 5). Appendix B proves the active-set, path-deviation, concentration, held-out ordering, selective-failure, and split-conformal results under their respective assumptions. The threshold-margin theory describes the threshold adapters, while the state-space adapter uses its energy-dispersion index. Localized wavelet compressibility motivates the prediction-coupled specialization (DeVore et al., 1992). Appendix D specifies the empirical activation procedure.

3.6 Training and Complexity

The prediction loss is

$$\mathcal{L} = \mathcal{L}_{L^2} + \lambda_{H^1} \mathcal{L}_{H^1} + \lambda_{HF} \mathcal{L}_{HF}, \tag{12}$$

Table 1: Matched EM transfer separates predictive accuracy from allocation risk ranking: the readout ranks ID error across wavelet, Fourier, and branch–trunk adapters. Coupling indicates allocation's role in prediction. Five independent training seeds; mean±sample standard deviation (SD). Bold marks column extrema; †: validation-fixed two-sided OOD score.

AUD adapter		Predictor	AUD risk readout		
Backbone	**Coupling**	**Rel. L_2 ↓**	**ID ρ_s ↑**	**Geom. AUROC↑**	**Freq. AUROC↑**
MWNO	Coupled	$.709 \pm .004$	$.459 \pm .005$	$.459 \pm .001$	$.438 \pm .048$
Dense WNO	Preserved	$.708 \pm .004$	$\mathbf{.492 \pm .003}$	$.455 \pm .001$	$.488 \pm .082$
FNO	Preserved	$\mathbf{.599 \pm .002}$	$.238 \pm .025$	$.425 \pm .004$	$.618 \pm .105$
Fourier-DeepONet	Preserved	$.776 \pm .003$	$.478 \pm .040$	$\mathbf{.496 \pm .008}^{\dagger}$	$\mathbf{1.000 \pm .000}^{\dagger}$

where $\mathcal{L}_{HF}$ is the high-frequency (HF) loss. We set $(\lambda_{H^1}, \lambda_{HF}) = (0.1, 0.05)$ in the EM deployment and factorial tracks. Those tracks use AdamW, OneCycleLR, a 100-epoch budget, batch size 96, and mixed precision; validation data select all thresholds and calibration maps. The trajectory and volumetric tracks use their architecture-specific batch sizes and validation/early-stopping protocols detailed in Appendix F. Per block, the DWT, inverse DWT (IDWT), and coefficient mixing scale as $\mathcal{O}(CHW)$, while allocation construction adds energy aggregation and quantile selection at $\mathcal{O}(CHW + HW \log(HW))$. The entropy readout adds 4.8% to the batch-96 pass; prediction plus AUD is $9.88\times$ faster than ten-pass MC-dropout on one RTX 4090. Appendix F records timing and batch-one costs.

4 Risk-Triage Evaluation

4.1 Experimental Setup

We evaluate on an EM scattering case study, PDEBench Darcy flow and one-shot Burgers space-time prediction (Takamoto et al., 2022), standard 64^2 Navier–Stokes trajectories (Li et al., 2021), and native Maxwell-3D trajectories from PDEArena (Gupta & Brandstetter, 2023; Brandstetter et al., 2023). Baselines include input and internal statistics, supervised heads, ensembles, Laplace/KFAC, SNGP, and MC-dropout (Table 2). We use the matched factorial experiments for the main EM comparison; conformal calibration uses a separate five-seed protocol. The *factorial track* uses a common data split, loss, and training budget across MWNO, dense WNO, FNO, and Fourier-DeepONet, with matched wavelet backbones to separate mask and readout effects over five seeds. Validation selects score direction and operating parameters before test evaluation. EM backbone comparisons use the 100-epoch training budget and the method-specific validation grids in Appendix F. Appendix H retains descriptive summaries from the separate deployment track.

Metrics. For each sample, prediction error is relative field error $\epsilon_i = \|\hat{u}_i - u_i\|_2/(\|u_i\|_2 + \eta)$. We report ID ranking quality with Spearman ρ_s and OOD detection with the area under the receiver operating characteristic curve (AUROC). Sel.@80% is the fraction of high-error samples rejected when accepting 80% of samples. The area under the risk–coverage curve (AURC) integrates mean accepted-sample error across acceptance fractions; excess AURC (E-AURC) subtracts the oracle ordering. Normalized E-AURC divides this excess by the same model's mean test error; lower values are better. Conformal evaluation reports in-domain marginal coverage and mean relative-error threshold; Eq. (17) supplies the sample-wise entropy scale before the second validation split calibrates the conformal quantile. We report batch cost relative to WNO at fixed batch size 96.

The OOD columns in Table 1 report validation-oriented raw entropy, except the marked two-sided Fourier-DeepONet scores. The target-specific validation check can instead select the two-sided deviation in Eq. (8); for FNO the geometry signal passes validation and obtains test AUROC 0.571 ± 0.003, while frequency remains routed.

4.2 Ranking Quality and Inference Cost

Table 2 compares error ranking and per-sample latency for the fixed entropy readout, supervised heads, posterior estimators, and multi-pass methods. Layer probes and variance heads reach $\rho_s = 0.617$–

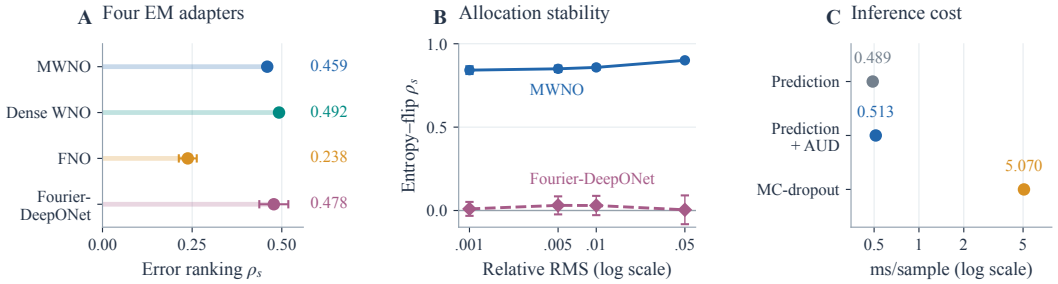

Figure 2: One-pass allocation readouts expose risk information across operator families at measured inference cost. **A:** ID error ranking. **B:** Entropy–allocation-flip correlation under Gaussian energy perturbations with recomputed quantiles. A–B show mean±sample SD over five seeds. **C:** Prediction and uncertainty latency on one RTX 4090 at batch 96 (mean±SD, 100 synchronized iterations); MC-dropout uses ten passes.

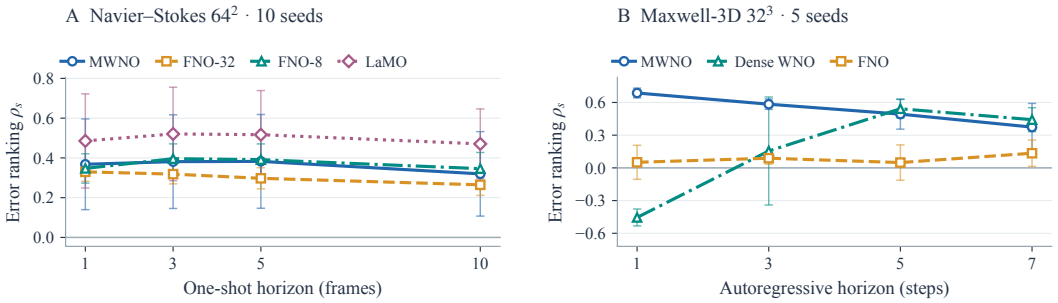

Figure 3: Allocation ranking persists in one-shot global dynamics and autoregressive volumetric fields. **A:** Navier–Stokes one-shot prediction with parameter-matched FNO-8 and larger FNO-32 controls ($n = 10$ seeds). **B:** Native Maxwell-3D autoregressive rollout ($n = 5$). Markers show means and error bars show sample SD. Validation fixes each score and orientation before test; the two panels use different vertical ranges.

0.672. The variance head takes 0.496 ms/sample, compared with 0.513 for AUD ($\rho_s = 0.459$) and 5.070 for ten-pass MC-dropout ($\rho_s = 0.613$). Optional escalation costs $\mathcal{C}_{\text{AUD}} + q\mathcal{C}_{\text{strong}}$: $\mathcal{C}_{\text{AUD}}$ is the first-pass prediction-plus-readout cost, q is the routed fraction, and $\mathcal{C}_{\text{strong}}$ counts additional computation. The learned readouts use auxiliary supervision. Entropy-scaled conformal prediction adds the fitted scale map in Eq. (17). The descriptive deployment summaries and internal-score ablations appear in Appendix H. Entropy-scaled conformal prediction reaches $95.76 \pm 0.17\%$ in-domain coverage with mean relative-error threshold 0.964 ± 0.003, versus $95.34 \pm 0.09\%$ and 0.959 ± 0.005 unscaled (Appendix G, Table 10).

5 Representation-Dependent Risk Across Operators and Physics

5.1 Allocation Readouts Transfer Beyond Wavelets

Prediction-preserving adapters test whether allocation entropy carries risk information in operators whose predictions use no AUD allocation masks. The matched five-seed EM protocol gives positive ID ranking for wavelet, Fourier, and branch–trunk readouts (Table 1). Dense WNO and MWNO have similar prediction errors (0.708 and 0.709), with entropy–error correlations of 0.492 and 0.459. Dense WNO thus supplies risk information from an unchanged predictor. Fourier-DeepONet reaches $\rho_s = 0.478 \pm 0.040$ and detects frequency shift in all five seeds; the validation-selected FNO score supports geometry-shift detection. The independently generated compound-shift validation set (1,000 ID and 1,000 OOD samples) gives MWNO/dense-WNO/FNO AUROC $0.852/0.856/0.430$. The lower 95% confidence bound passes the signal check for MWNO and dense WNO on that declared shift and routes FNO before any disjoint test evaluation.

Table 2: EM baselines on one RTX 4090 at batch 96. "Added state" counts uncertainty-specific inference state; RFFs are random Fourier features and KFAC denotes Kronecker-factored approximate curvature. Predictive errors expose backbone differences; Appendix Table 8 lists validation-only tuning.

Method	**Passes**	**Added state**	**ID rel.** $L_2 \downarrow$	**ID** $\rho_s \uparrow$	**ms/sample**$\downarrow$
Zero-head allocation readouts					
MWNO entropy	1	0	0.709	0.459 ± 0.005	0.513
FNO Fourier entropy	1	0	0.599	0.238 ± 0.025	0.377
Learned, posterior, and multi-pass references					
MWNO ensemble (3)	3	2 models	0.703	0.498	1.57
MWNO layer probes	1	12,870 params	0.709	0.672 ± 0.032	0.619
MWNO variance head	1	4,290 params	0.700 ± 0.002	0.617 ± 0.020	0.496 ± 0.001
MWNO Laplace/KFAC	1	KFAC factors	0.700 ± 0.002	0.484 ± 0.077	0.541 ± 0.001
SNGP operator	1	128 RFFs	0.777 ± 0.013	0.344 ± 0.038	0.889 ± 0.002
MWNO MC-dropout	10	0	0.705 ± 0.002	0.613 ± 0.111	5.070

Table 3: New controlled advection–diffusion task: five seeds per model family, mean±sample SD. We fixed the FNO-8 capacity follow-up after the ten-model primary experiment. Retained error is the accepted-set mean divided by the same model's full-test mean at 80% coverage; lower is better. All score signs use calibration labels only.

Predictor	Parameters	Rel. $L_2 \downarrow$	AUD $\rho_s \uparrow$	AUD retained$\downarrow$	Input ridge retained$\downarrow$
MWNO-32	156,073	$.277 \pm .006$	$.313 \pm .043$	$.955 \pm .005$	$\mathbf{.937 \pm .003}$
FNO-32	2,364,833	$.169 \pm .003$	$.320 \pm .025$	$.942 \pm .011$	$\mathbf{.931 \pm .002}$
FNO-8	147,881	$.342 \pm .021$	$.310 \pm .050$	$.965 \pm .012$	$\mathbf{.955 \pm .006}$

5.2 Within-Regime Ranking Transfers Across Physics

Six of seven Darcy/Burgers task–adapter pairs show positive within-regime ranking; Fourier-DeepONet reaches Darcy $\rho_s = 0.241 \pm 0.009$ (Appendix Table 4). Validation routes coefficient-shift cases whose detection AUROC is near chance.

New advection–diffusion models retain positive MWNO/FNO error ranking (Table 3). At 80% coverage, a two-feature input ridge screens better in every cohort; FNO-32/FNO-8 shift AUROC is $0.513/0.504$, versus 1.000 for known speed/diffusivity (Appendix O).

5.3 Ranking Persists in Global, 3D, and Long-Horizon Regimes

On 1,200 Navier–Stokes trajectories, LaMO reaches ten-frame $\rho_s = 0.470$ versus 0.346 for parameter-matched FNO-8 and 0.265 for FNO-32. The LaMO–FNO-32 difference has 95% CI $[0.082, 0.330]$; Holm adjustment over all six architecture pairs gives $p = 0.059$. LaMO has the lowest mean normalized E-AURC (0.058; versus FNO-32, $p_{\mathrm{Holm}} = 0.012$). The primary raw E-AURC comparison gives $p_{\mathrm{Holm}} = 0.742$. Native 32^3 Maxwell evaluation gives positive MWNO ranking through seven autoregressive steps ($0.687/0.584/0.493/0.374$). Figure 3 and Appendix Tables 6–7 distinguish one-shot prediction from autoregressive rollout.

Resolution and target choice change ranking: Darcy correlation rises from 0.022 ± 0.020 at 32^2 to 0.352 ± 0.057 at 128^2, while positive physical-target intervals vary by operator (Appendix Figure 5 and Tables 12–13).

5.4 Allocation Representation Changes Risk Ranking

Readout choice changes decisions on identical predictions: Haar-to-db2 replacement changes $22.5 \pm 8.0\%$ of test decisions at 80% coverage (Figure 4). The retained-error gap survives Haar-anchored scale correction, making readout basis an independent post-training choice (Appendix M).

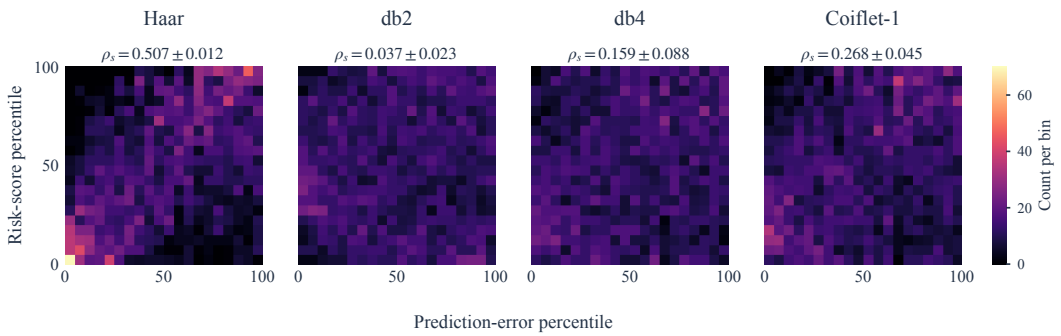

Figure 4: Changing only the readout basis changes risk ranking while field predictions stay fixed. Each panel pools unsmoothed 20×20 error-versus-oriented-score percentile counts from five frozen Haar-trained models on the same 1,000 EM test fields, with a shared color scale. Titles give model-level Spearman mean±sample SD.

A crossed cohort trains Haar/db2 MWNOs on EM and independently generated Darcy fields (Table 23). Across 20 paired-seed models with neutral readout temperature, matching slightly lowers retained error in EM; Haar wins in Darcy under both training bases. The mean matching effects are 0.00066 ± 0.00020 and 0.00133 ± 0.00805, respectively. The Darcy mean averages opposing training-arm effects (Appendix N).

Replacing only the score extractor with a random network reduces ρ_s from 0.507 to 0.313 and increases E-AURC by 0.0222 (95% CI $[0.0181, 0.0262]$). MWNO raw entropy correlates positively with allocation flips (0.842–0.901) and negatively with field error (-0.507); flip–error correlations range from -0.553 to -0.548 across the four perturbation scales. Validation selects the risk-score direction, while these signed correlations distinguish internal instability from task error (Appendix M). Independent and matched EM test quintiles expose raw-entropy counterexamples and adapter-specific error modes after orientation (Appendix Tables 20–21).

Input roughness and occupied area rank EM error at least as strongly as raw entropy in the factorial suite. Within joint strata of contrast, roughness, and occupied area, allocation–error ranking remains positive at every tested granularity ($\rho_s = 0.080$–0.161). Exploratory five-seed input-only fits reach $\rho_s = 0.576/0.559$ for ridge/gradient boosting. Adding entropy yields $0.576/0.554$ and increases mean E-AURC for both learners (Appendix K).

5.5 Calibration Budgets Quantify Retained Error

We compare 15 screening policies when a frozen predictor has 8–256 calibration labels, with 1,000 labels as a reference. Five MWNO models use shared calibration subsets and a new 1,000-example EM test set. At 80% test coverage, the budget-integrated retained error, normalized by each model's mean error, is 0.9693 for AUD. Selected input statistics, input ridge, and score-summary ridge reach 0.9684, 0.9665, and 0.9682, respectively, outperforming AUD in all five models (Appendix L). Across budgets, AUD's zero-head score reduces retained error by 3.07% relative to random screening.

The validation signal ledger identifies selected EM and trajectory targets and routes all 12 PDEBench coefficient-shift cases to the stronger backend (Appendix Table 5). For a new regime, calibration checks score–error ranking and selective risk against same-budget input statistics before choosing a screening policy; each named shift receives its own detection check.

6 Discussion and Conclusion

An operator's representation shapes the evidence available for triage. AUD measures that evidence in one pass across five architectures. Fixed-prediction basis swaps alter decisions, crossed training exposes task-specific preferences, and new MWNO/FNO models rank advection–diffusion error. Input controls remain strong comparators. Representation choice is a measurable design axis for neural-operator risk readouts.

Ethics Statement

Single-pass triage can concentrate solver, ensemble, or expert evaluation on high-risk queries. Safety-sensitive deployment requires validation for the intended physical regime and a fallback solver or model.

Reproducibility Statement

Appendix B states assumptions and proofs; Appendix F specifies datasets, splits, checkpoint pools, and validation-only tuning. Appendices G and I describe conformal calibration and the complete sensitivity surfaces. Appendix L specifies the independent EM label-budget experiment and its complete comparison groups. Appendix M defines the secondary fixed-prediction decision analysis and signed mechanism comparisons. Appendix N specifies the crossed training experiment, independent data generation, and complete primary and secondary comparisons. Appendix O records the prospective new-physics and later capacity-matched cohorts, all model-level results, and fixed input/representation controls.

AI Use Statement

Generative AI assisted with manuscript editing, review-response organization, clarification of theoretical assumptions and proofs, conversion to the ICLR format, figure preparation, and implementation and analysis of the label-budget, readout-characterization, and crossed-training studies.

References

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Johannes Brandstetter, Rianne van den Berg, Max Welling, and Jayesh K. Gupta. Clifford neural layers for PDE modeling. In *International Conference on Learning Representations*, 2023.

Christopher Bülte, Philipp Scholl, and Gitta Kutyniok. Probabilistic neural operators for functional uncertainty quantification. *Transactions on Machine Learning Research*, 2025.

Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. Posterior network: Uncertainty estimation without OOD samples via density-based pseudo-counts. *Advances in Neural Information Processing Systems*, 33:1356–1367, 2020.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux—effortless Bayesian deep learning. *Advances in Neural Information Processing Systems*, 34:20089–20103, 2021.

Tim De Ryck and Siddhartha Mishra. Generic bounds on the approximation error for physics-informed (and) operator learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 10945–10958, 2022.

Tim De Ryck and Siddhartha Mishra. Numerical analysis of physics-informed neural networks and related models in physics-informed machine learning. *Acta Numerica*, 33:633–713, 2024. doi: 10.1017/S0962492923000089.

Ronald A DeVore, Björn Jawerth, and Vasil Popov. Compression of wavelet decompositions. *American Journal of Mathematics*, 114(4):737–785, 1992.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. GLaM: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pp. 5547–5569. PMLR, 2022.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Shailesh Garg and Souvik Chakraborty. Randomized prior wavelet neural operator for uncertainty quantification. *Probabilistic Engineering Mechanics*, 81:103817, 2025a. doi: 10.1016/j.probengmech.2025.103817.

Shailesh Garg and Souvik Chakraborty. Distribution free uncertainty quantification for neuroscience-inspired deep neural operators. *Journal of Computational Physics*, 534:114012, 2025b. doi: 10.1016/j.jcp.2025.114012.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in Neural Information Processing Systems*, 30, 2017.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Gaurav Gupta, Xiongye Xiao, and Paul Bogdan. Multiwavelet-based operator learning for differential equations. *Advances in Neural Information Processing Systems*, 34:24048–24062, 2021.

Jayesh K. Gupta and Johannes Brandstetter. Towards multi-spatiotemporal-scale generalized PDE modeling. *Transactions on Machine Learning Research*, 2023.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

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

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *ICML*, pp. 2796–2804. PMLR, 2018.

Kaisar Kushibar, Victor Campello, Lidia Garrucho, Akis Linardos, Petia Radeva, and Karim Lekadir. Layer ensembles: A single-pass uncertainty estimation in deep learning for segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pp. 514–524. Springer, 2022.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

Zongyi Li, Daniel Zhengyu Huang, Burigede Liu, and Anima Anandkumar. Fourier neural operator with learned deformations for PDEs on general geometries. *Journal of Machine Learning Research*, 24(388):1–26, 2023.

Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, 33:7498–7512, 2020a.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020b.

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

Emilia Magnani, Marvin Pförtner, Tobias Weber, and Philipp Hennig. LUNO: Linearization turns neural operators into function-valued gaussian processes. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 42421–42452. PMLR, 2025. URL https://proceedings.mlr.press/v267/magnani25a.html.

Pascal Massart. The tight constant in the dvoretzky–kiefer–wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990. doi: 10.1214/aop/1176990746.

Carlos Mora, Jonathan Tammer Eweis-Labolle, Tyler Johnson, Likith Gadde, and Ramin Bostanabad. Probabilistic neural data fusion for learning from an arbitrary number of multi-fidelity data sets. *Computer Methods in Applied Mechanics and Engineering*, 415:116207, 2023.

S Chandra Mouli, Danielle C. Maddix, Shima Alizadeh, Gaurav Gupta, Andrew Stuart, Michael W. Mahoney, and Bernie Wang. Using uncertainty quantification to characterize and improve out-of-domain learning for PDEs. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36372–36418. PMLR, 2024. URL https://proceedings.mlr.press/v235/mouli24a.html.

Jishnu Mukhoti, Andreas Kirsch, Joost Van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24384–24394, 2023.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, 32, 2019.

Svetlana Pavlitska, Beyza Keskin, Alwin Faßbender, Christian Hubschneider, and J. Marius Zöllner. Extracting uncertainty estimates from mixtures of experts for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 311–320, 2025.

Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems*, 31, 2018.

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.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An extensive benchmark for scientific machine learning. *Advances in Neural Information Processing Systems*, 35:1596–1611, 2022.

Akshay Thakur, Tapas Tripura, and Souvik Chakraborty. Multi-fidelity wavelet neural operator with application to uncertainty quantification. *arXiv preprint arXiv:2208.05606*, 2022.

Karn Tiwari, Niladri Dutta, N. M. Anoop Krishnan, and Prathosh Ap. Latent mamba operator for partial differential equations. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59639–59667. PMLR, 2025. URL https://proceedings.mlr.press/v267/tiwari25a.html.

Alasdair Tran, Alexander Mathews, Lexing Xie, and Cheng Soon Ong. Factorized fourier neural operators. In *International Conference on Learning Representations*, 2023.

Tapas Tripura and Souvik Chakraborty. Wavelet neural operator for solving parametric partial differential equations in computational mechanics problems. *Computer Methods in Applied Mechanics and Engineering*, 404:115783, 2023.

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning*, pp. 9690–9700. PMLR, 2020.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

A Wavelet Implementation Details

A.1 Wavelet Operator Instantiation

For the wavelet instantiation, MWNO-AUD stacks $B{=}8$ wavelet blocks. Each block applies a discrete wavelet transform (DWT), deterministic energy allocation, coefficient mixing, an inverse discrete wavelet transform (IDWT), and residual addition. Wavelet coefficient energy determines the gate. The masks weight detail coefficients during prediction and supply the probabilities for the entropy readout. Figure 1 summarizes the prediction path, allocation readout, and deployment use.

At wavelet level $j \in \{1, 2\}$, LL denotes the low–low approximation subband, while LH, HL, and HH denote the low–high, high–low, and high–high detail subbands. We compute the channel mean of each channel's three-detail-subband L_2 norm:

$$a_j(x, y) = \frac{1}{C} \sum_{c=1}^{C} \sqrt{\sum_{b\in\{\mathrm{LH,HL,HH}\}} |v_j^b(c, x, y)|^2}. \tag{13}$$

Given retained fraction r_g, the threshold τ_j is the sample-wise $(1 - r_g)$ quantile of a_j. Gates are soft masks

$$g_j(x, y) = \sigma\left(\frac{a_j(x, y) - \tau_j}{T}\right), \tag{14}$$

where T controls sharpness. For the wavelet score we average entropy over blocks, scales, and spatial cells:

$$H_L(a) = \frac{1}{\sum_{b,j} H_j\, W_j} \sum_{b,j,x,y} \mathsf{h}\Big(g_j^{(b)}(x, y)\Big). \tag{15}$$

The primary EM experiment fixes $r_g = 0.15$ and $T = 0.2$ before test evaluation. The robustness study varies prediction-coupled masks on frozen checkpoints and reports the validation/test response surface (Appendix I). The score averages local ambiguity: a component's entropy decreases as its absolute threshold margin grows relative to T. Equation (7) converts this raw geometry into a target-specific risk score using held-out validation only.

B Complete Proofs of Theoretical Results

B.1 Statements of Theoretical Results

Proposition 1 (Clipped Allocation-Entropy Concentration)**.** *For $0 < \epsilon_c < 1/2$ and n clipped component probabilities $\tilde{p}_i = \mathrm{clip}(\sigma((E_i - \tau)/T), \epsilon_c, 1 - \epsilon_c)$, partition the component field into n_{eff} equal-size independent blocks and let $\hat{H}_c$ be the mean of their bounded entropy averages. Then*

$$|\hat{H}_c - \mathbb{E}\hat{H}_c| \le (\log 2)\sqrt{\ln(2/\delta)/(2n_{\mathrm{eff}})}$$

with probability at least $1 - \delta$. For two margin fields m, m', the clipped entropy map also obeys

$$|\Delta\hat{H}_c| \le \frac{\log((1 - \epsilon_c)/\epsilon_c)}{4Tn} \sum_i |m_i - m'_i|.$$

Proposition 2 (Holdout Pairwise-Ranking Diagnostic)**.** *Fix the score and minimum separation γ on data independent of an i.i.d. diagnostic sample. Randomly match diagnostic samples independently of their observed values, then retain the $K_\gamma \ge 1$ pairs with score gap at least γ. Let q_γ denote the conditional probability of a score–error sign violation given this gap. If $\hat{q}_\gamma$ is the empirical sign-violation rate between AUD scores and relative errors, then with probability at least $1 - \delta$,*

$$q_\gamma \le \hat{q}_\gamma + \sqrt{\log(1/\delta)/(2K_\gamma)}.$$

Proposition 3 (Allocation Entropy Controls Active-Set Instability)**.** *Conditional on fixed margins, let ξ_i be centered σ_i^2-sub-Gaussian noise, with $\sigma_i > 0$. For $A_i = \mathbb{1}\{m_i \ge 0\}$ and $A'_i = \mathbb{1}\{m_i + \xi_i \ge 0\}$, hard active-set flips satisfy*

$$\Pr(A'_i \ne A_i) \le \min\{1, 2\exp(-m_i^2/(2\sigma_i^2))\}.$$

Because $\mathsf{h}(\sigma(m_i/T))$ decreases with $|m_i|$, a monotone function of allocation entropy bounds expected active-set flips.

Proposition 4 (Allocation Entropy Bounds Prediction-Path Deviation). *Let an allocation block produce $v(A) = \sum_i A_i z_i$ and let its downstream decoder D be L_D-Lipschitz. Set $A_i = \mathbb{1}\{m_i \geq 0\}$ and perturb each margin by independent logistic noise ξ_i with scale T, giving $A'_i = \mathbb{1}\{m_i + \xi_i \geq 0\}$ and $p_i = \sigma(m_i/T)$. Write $\Delta_D = \|D(v(A')) - D(v(A))\|_2$. Then*

$$\begin{aligned}\mathbb{E}_\xi[\Delta_D] &\leq L_D \sum_i \|z_i\|_2 \min\{p_i, 1-p_i\} \\ &\leq \frac{L_D}{2\log 2} \sum_i \|z_i\|_2 \mathsf{h}(p_i).\end{aligned}$$

Thus weighted allocation entropy bounds the expected output change induced by near-boundary component decisions.

Proposition 5 (Allocation Entropy Bounds Perturbed-Path Variance). *Fix an input, component vectors z_i, and margins m_i. Let independent $\xi_i \sim \mathrm{Logistic}(0, T)$ define $A_i^\xi = \mathbb{1}\{m_i + \xi_i \geq 0\}$ and $p_i = \Pr_\xi(A_i^\xi = 1) = \sigma(m_i/T)$. For $Y_\xi = D(\sum_i A_i^\xi z_i)$ with an L_D-Lipschitz decoder,*

$$\mathbb{E}_\xi \|Y_\xi - \mathbb{E}_\xi Y_\xi\|_2^2 \leq L_D^2 \sum_i \|z_i\|_2^2 p_i(1-p_i) \leq \frac{L_D^2}{2\log 2} \sum_i \|z_i\|_2^2 \mathsf{h}(p_i).$$

If D is linear, the first bound is an equality after replacing $L_D^2\|z_i\|_2^2$ by $\|Dz_i\|_2^2$. The corresponding soft-allocation output $Y_p = D(\sum_i p_i z_i)$ also satisfies

$$\|Y_p - \mathbb{E}_\xi Y_\xi\|_2 \leq L_D \Big(\sum_i \|z_i\|_2^2 p_i(1-p_i) \Big)^{1/2}.$$

Proposition 6 (Uniform Bound for Held-Out Threshold Policies). *Fix a score S and a finite set $\mathcal{T}$ of J candidate thresholds independently of n i.i.d. policy-evaluation samples. For bounded error $\epsilon \in [0, M]$, incremental cost $c_1 \geq 0$, and fixed $\lambda \geq 0$, define $R_\tau = \mathbb{E}[\epsilon \mathbb{1}\{S \leq \tau\}]$ and $C_\tau = c_0 + c_1 \Pr(S > \tau)$. With probability at least $1 - \delta$, all $\tau \in \mathcal{T}$ satisfy*

$$\begin{aligned}&|(\widehat{R} + \lambda \widehat{C}) - (R + \lambda C)| \\ &\quad \leq (M + \lambda c_1)\sqrt{\log(2J/\delta)/(2n)}.\end{aligned}$$

Proposition 7 (Uniform Audit Bound for Accepted High-Error Rate). *Fix a predictor, an oriented score S, and an error tolerance $e_\star$ independently of n i.i.d. audit examples. For any threshold t, let*

$$\widehat{A}(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{S_i \leq t\}, \qquad \widehat{B}(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{S_i \leq t, \epsilon_i > e_\star\},$$

and define $A(t)$ and $B(t)$ as their population probabilities. For $r_n = \sqrt{\log(4/\delta)/(2n)}$, with probability at least $1 - \delta$, simultaneously for every t with $\widehat{A}(t) > r_n$,

$$\Pr(\epsilon > e_\star \mid S \leq t) \leq \min\left\{1, \frac{\widehat{B}(t) + r_n}{\widehat{A}(t) - r_n}\right\}, \qquad A(t) \geq \widehat{A}(t) - r_n.$$

The threshold may be selected using the audit labels because the bound holds for every threshold at once.

Corollary 1 (Exact Audit for a Validation-Fixed Threshold). *Under the sampling assumptions of Proposition 7, also fix t independently of the audit set. Let $K = \sum_i \mathbb{1}\{S_i \leq t\}$ and $F = \sum_i \mathbb{1}\{S_i \leq t, \epsilon_i > e_\star\}$. For $0 \leq F < K$, let $U_{\mathrm{bin}}(F, K, \delta)$ solve*

$$\sum_{j=0}^F \binom{K}{j} U_{\mathrm{bin}}^j (1 - U_{\mathrm{bin}})^{K-j} = \delta;$$

set $U_{\mathrm{bin}} = 1$ when $F = K$ or $K = 0$. Then $\Pr(\epsilon > e_\star \mid S \leq t) \leq U_{\mathrm{bin}}$ with probability at least $1 - \delta$. For $F = 0$ and $K > 0$, $U_{\mathrm{bin}} = 1 - \delta^{1/K}$.

Corollary 2 (Entropy-Scaled Split-Conformal Field Coverage). *Fit the positive scale $\sigma_\beta(a)$ in Eq. (17) on data disjoint from conformal calibration, and define calibration nonconformity $r_i = \epsilon_i / \sigma_\beta(a_i)$. Under exchangeability, the relative-error prediction set with scale $q_{1-\alpha}\sigma_\beta(a)$ has marginal coverage at least $1 - \alpha$.*

B.2 Proof of Proposition 1 (Clipped Entropy Concentration)

Full statement. Let $p_i = \sigma(m_i/T)$, $\tilde{p}_i = \mathrm{clip}(p_i, \epsilon_c, 1-\epsilon_c)$, and $\hat{H}_c = n^{-1}\sum_i \mathsf{h}(\tilde{p}_i)$, with $\mathsf{h}(p) = -p\log p - (1-p)\log(1-p)$. Partition the n components into n_{eff} equal-size blocks whose block entropy averages are independent; their mean equals $\hat{H}_c$. Independence here concerns the block entropy averages after thresholding. A shared sample-wise quantile can couple local components; a deployed-network concentration bound therefore requires a dependence analysis. We use this proposition for its conditional mechanism statement and assess deployed scores empirically, without assigning an n_{eff} to these networks. Then, with probability at least $1-\delta$,

$$\left|\hat{H}_c - \mathbb{E}[\hat{H}_c]\right| \le (\log 2)\sqrt{\frac{\ln(2/\delta)}{2n_{\mathrm{eff}}}}.$$

For perturbed threshold margins m_i and m_i',

$$\left|\Delta\hat{H}_c\right| \le \frac{L_{\epsilon_c}}{4Tn}\sum_i |m_i - m_i'|, \quad L_{\epsilon_c} = \log\frac{1-\epsilon_c}{\epsilon_c}.$$

Detailed proof. For every clipped probability, $\mathsf{h}(\tilde{p}_i) \in [0, \log 2]$. Let Y_k be the mean entropy within independent block k; then $Y_k \in [0, \log 2]$ and $\hat{H}_c = n_{\mathrm{eff}}^{-1}\sum_k Y_k$. Hoeffding's inequality gives

$$\Pr\left(\left|\hat{H}_c - \mathbb{E}[\hat{H}_c]\right| > t\right) \le 2\exp\left(-\frac{2n_{\mathrm{eff}}t^2}{(\log 2)^2}\right).$$

Solving for t gives the stated concentration inequality.

For the stability part, clipping restricts the entropy derivative:

$$|\mathsf{h}'(p)| = \left|\log\frac{1-p}{p}\right| \le L_{\epsilon_c} \quad \text{for } p \in [\epsilon_c, 1-\epsilon_c].$$

The sigmoid probability $p_i = \sigma(m_i/T)$ is $1/(4T)$-Lipschitz in the threshold margin. Thus

$$|\mathsf{h}(\tilde{p}_i) - \mathsf{h}(\tilde{p}_i')| \le L_{\epsilon_c}|p_i - p_i'| \le \frac{L_{\epsilon_c}}{4T}|m_i - m_i'|.$$

Averaging over components yields the claimed perturbation bound.

B.3 Proof of Proposition 2 (Holdout Pairwise-Ranking Diagnostic)

Full statement. After fixing S and γ on $\mathcal{D}_{\mathrm{cal}}$, draw independent identically distributed diagnostic samples from the deployment distribution. Randomly match their indices independently of the observations, then retain the $K_\gamma \ge 1$ disjoint pairs satisfying $|S_i - S_j| \ge \gamma$. For each retained pair, define

$$V_{ij} = \mathbb{1}[\mathrm{sign}(S_i - S_j) \ne \mathrm{sign}(\epsilon_i - \epsilon_j)].$$

Let $\hat{q}_\gamma = K_\gamma^{-1}\sum V_{ij}$. Then, with probability at least $1-\delta$,

$$q_\gamma \le \hat{q}_\gamma + \sqrt{\frac{\log(1/\delta)}{2K_\gamma}}.$$

Detailed proof. Conditional on the retained count, independent initial pairs and pairwise gap filtering give independent bounded violation indicators with common mean q_γ. Hoeffding's inequality gives

$$\Pr\left(q_\gamma - \hat{q}_\gamma > t\right) \le \exp(-2K_\gamma t^2).$$

Solving for t at failure probability δ yields the stated one-sided upper confidence bound. All-pairs diagnostics require a U-statistic concentration bound because pairs share observations.

B.4 Proof of Proposition 3 (Active-Set Instability)

A hard active-set decision flips only if the perturbation crosses zero:

$$A_i' \neq A_i \quad \Rightarrow \quad |\xi_i| \geq |m_i|.$$

The assumed sub-Gaussian tail therefore implies

$$\Pr(A_i' \neq A_i) \leq \Pr(|\xi_i| \geq |m_i|) \leq 2\exp\left(-\frac{m_i^2}{2\sigma_i^2}\right),$$

truncated at one for small margins. The function $\mathsf{h}(\sigma(x/T))$ is symmetric and strictly decreasing in $|x|$ for $|x| \geq 0$. Let $\eta_T(h)$ be its inverse in $|x|$ on $h \in [0, \log 2]$. Then one may set

$$\psi_i(h) = \min\left\{1, 2\exp\left(-\frac{\eta_T(h)^2}{2\sigma_i^2}\right)\right\},$$

which is monotone increasing in h. Summing the resulting component-wise flip probabilities yields the expected flip-count bound.

B.5 Proof of Proposition 4 (Prediction-Path Deviation)

The Lipschitz property and triangle inequality give

$$\begin{aligned}\|D(v(A')) - D(v(A))\|_2 &\leq L_D\|v(A') - v(A)\|_2 \\ &\leq L_D \sum_i \|z_i\|_2 \mathbb{1}\{A_i' \neq A_i\}.\end{aligned}$$

For logistic noise with scale T, a positive margin flips with probability $\sigma(-m_i/T) = 1 - p_i$, while a negative margin flips with probability $\sigma(m_i/T) = p_i$. Hence $\Pr(A_i' \neq A_i) = \min\{p_i, 1 - p_i\}$. Taking expectations proves the first inequality.

For $q \in [0, 1/2]$, concavity of binary entropy and its endpoint values imply $\mathsf{h}(q) \geq 2(\log 2)q$. Substituting $q = \min\{p_i, 1 - p_i\}$ proves the second inequality. The bound applies to threshold-margin adapters whose hard-allocation surrogate satisfies the additive-block and Lipschitz-decoder assumptions.

B.6 Proof of Proposition 5 (Perturbed-Path Variance)

The logistic cumulative distribution function and symmetry give $\Pr(\xi_i \geq -m_i) = \sigma(m_i/T)$. At the fixed input, A_i^ξ is therefore Bernoulli with conditional entropy $\mathsf{h}(p_i)$; independent perturbations make the vector entropy the sum of its component entropies. Let $M_i = \mathbb{E}_\xi[Y_\xi \mid A_1^\xi, \ldots, A_i^\xi]$, with $M_0 = \mathbb{E}_\xi Y_\xi$. Orthogonality of the martingale differences in Euclidean output space yields

$$\mathbb{E}_\xi\|Y_\xi - M_0\|_2^2 = \sum_i \mathbb{E}_\xi\|M_i - M_{i-1}\|_2^2.$$

Given the first $i - 1$ decisions, the two values of M_i differ by a vector of norm at most $L_D\|z_i\|_2$: couple the remaining decisions and change only A_i^ξ, then apply the decoder's Lipschitz bound and Jensen's inequality. The conditional variance of this Bernoulli choice is consequently at most $L_D^2\|z_i\|_2^2 p_i(1 - p_i)$. Summing proves the first bound. Concavity of binary entropy on $[0, 1/2]$ gives $\mathsf{h}(p) \geq 2(\log 2)\min\{p, 1 - p\} \geq 2(\log 2)p(1 - p)$, proving the second. If D is linear, $Y_\xi - \mathbb{E}_\xi Y_\xi = \sum_i (A_i^\xi - p_i)Dz_i$, and independence cancels all cross terms, giving the stated equality. For the soft-allocation output, Jensen's inequality and Lipschitzness give $\|D(\mathbb{E}_\xi v(A^\xi)) - \mathbb{E}_\xi D(v(A^\xi))\|_2 \leq L_D\sqrt{\mathbb{E}_\xi\|v(A^\xi) - \mathbb{E}_\xi v(A^\xi)\|_2^2}$. Independence cancels cross terms in this input-space variance, leaving $\sum_i \|z_i\|_2^2 p_i(1 - p_i)$.

B.7 Proof of Proposition 6 (Held-Out Threshold Policies)

For fixed τ, define

$$Z_\tau(a) = \epsilon(a)\mathbb{1}\{S(a) \leq \tau\} + \lambda c_1 \mathbb{1}\{S(a) > \tau\}.$$

The additive term λc_0 is constant and cancels from empirical-population deviations. Since Z_τ lies in an interval of length at most $M + \lambda c_1$, Hoeffding gives

$$\Pr\left(\left|\widehat{\mathbb{E}}Z_\tau - \mathbb{E}Z_\tau\right| > t\right) \le 2\exp\left(-\frac{2nt^2}{(M+\lambda c_1)^2}\right).$$

Conditional on the independently fixed score and candidate set, a union bound over the J candidates gives the stated simultaneous bound. The policy may minimize empirical risk-plus-cost over these candidates using the evaluation sample. For thresholds chosen from observed evaluation scores, a uniform empirical-process argument must account for candidate selection. An independent candidate-selection split recovers the fixed-set setting above. Here R_τ is unconditional accepted-error mass; conditional selective risk additionally divides by acceptance probability. Proposition 7 handles that conditional rate uniformly over all thresholds for a predeclared high-error event.

B.8 Proof of Proposition 7 (Accepted High-Error Rate)

Conditional on the independently fixed predictor, score, and error tolerance, the audit pairs (S_i, ϵ_i) are i.i.d. The empirical distribution function of S is $\widehat{A}$. The Dvoretzky–Kiefer–Wolfowitz inequality with its sharp constant (Massart, 1990) gives

$$\Pr\left(\sup_t |\widehat{A}(t) - A(t)| > r_n\right) \le 2e^{-2nr_n^2} = \delta/2.$$

Set $W_i = S_i$ when $\epsilon_i > e_\star$ and $W_i = +\infty$ otherwise. For every finite t, the empirical and population distribution functions of W equal $\widehat{B}(t)$ and $B(t)$. Applying the same inequality to these i.i.d. extended-real variables gives $\sup_t |\widehat{B}(t) - B(t)| \le r_n$ except with probability at most $\delta/2$. A union bound yields both inequalities simultaneously. Whenever $\widehat{A}(t) > r_n$, we have $A(t) > 0$ and

$$\frac{B(t)}{A(t)} \le \frac{\widehat{B}(t) + r_n}{\widehat{A}(t) - r_n}.$$

Uniformity permits threshold selection on the audit sample, including its error labels. The guarantee concerns future samples from the audited distribution; reusing the audit sample to orient or fit S, set $e_\star$, or select the predictor requires an independent split or a separate uniform analysis over those choices.

B.9 Proof of Corollary 1 (Validation-Fixed Threshold)

Conditional on the independently fixed threshold and on $K = k > 0$, the k accepted audit labels are independent Bernoulli draws with success probability $p(t) = \Pr(\epsilon > e_\star \mid S \le t)$. Thus $F \mid K = k \sim \mathrm{Binomial}(k, p(t))$. Inverting its lower-tail test gives the one-sided exact binomial upper confidence limit (Clopper & Pearson, 1934); it covers $p(t)$ with probability at least $1 - \delta$ for every k. Averaging over the random K preserves this coverage; $K = 0$ gives the vacuous bound one. When $F = 0$, the defining binomial tail is $(1 - U_{\mathrm{bin}})^K = \delta$. The threshold may use validation labels but cannot be selected from the audit labels for this corollary. Proposition 7 covers the latter case with a uniform bound. For example, at $\delta = 0.05$ with $n = 500$, empirical acceptance 0.8, and zero observed high-error cases, the audit-selected threshold bound is 0.0902; a validation-fixed threshold with $K = 400$ has exact upper limit 0.00746. These values illustrate the selection cost and are not measurements from the experiments.

B.10 Proof of Corollary 2 (Entropy-Scaled Coverage)

Let

$$\epsilon_i = \frac{\|u_i - \hat{u}_i\|_2}{\|u_i\|_2 + \eta}, \qquad r_i = \frac{\epsilon_i}{\sigma_\beta(a_i)}.$$

We fit the orientation, location, spread, and log-linear scale coefficients in σ_β on the first validation half. Conditional on that fit, the r_i on the disjoint calibration half and the corresponding test score are exchangeable under standard split-conformal assumptions. The usual conformal rank argument implies

$$\Pr(r_{n+1} \le q_{1-\alpha}) \ge 1 - \alpha.$$

Rearranging $r_{n+1} \leq q_{1-\alpha}$ gives

$$\frac{\|u_{n+1} - \hat{u}(a_{n+1})\|_2}{\|u_{n+1}\|_2 + \eta} \leq q_{1-\alpha}\sigma_\beta(a_{n+1}),$$

This event defines membership in a relative-error prediction set $\mathcal{C}(a) = \{v : \|v - \hat{u}(a)\|_2 \leq q_{1-\alpha}\sigma_\beta(a)(\|v\|_2 + \eta)\}$. The denominator depends on the candidate field v, making $q_{1-\alpha}\sigma_\beta(a)$ a relative-error threshold.

C Extended Theory: Risk Bounds and Budget-Aware Policies

C.1 From Prediction-Path Sensitivity to Risk Diagnostics

Propositions 3, 4, and 5 connect threshold entropy to active-set flips, prediction-path deviation, and perturbed-path output variance under their specified perturbations. Experiments then test the connection to task error on an independent split. Proposition 2 bounds pairwise ordering violations, and Proposition 6 bounds risk–cost deviations over independently fixed candidate thresholds. Proposition 7 certifies the high-error rate among accepted predictions when an independent audit set selects any threshold. The path bound characterizes sensitivity inside the operator; the held-out bounds quantify sampling uncertainty under their stated assumptions. Independent measurements establish the score's observed association with task error in each physical regime.

The gap between path sensitivity and physical error follows from an exact residual identity. Fix any reference allocation path with output $\hat{u}_0(a)$ and write $d(a) = \hat{u}(a) - \hat{u}_0(a)$ and $r_0(a) = \hat{u}_0(a) - u(a)$. Then

$$\|\hat{u} - u\|_2^2 = \|r_0\|_2^2 + \|d\|_2^2 + 2\langle r_0, d\rangle. \tag{16}$$

Equation (16) exposes two independent contributors to physical error: the reference-path residual r_0 and its alignment with the allocation-induced change d. A sharp allocation can retain a large r_0 even when d is small. An ambiguous allocation can produce a correct field when $r_0 = -d$, or zero output variance when its ambiguous components contribute zero to the decoder. The path bounds weight components by their contribution under specified hard-allocation perturbations; the deployed score averages unweighted soft-mask entropy. Evaluated relative error also depends on the input-varying denominator $\|u(a)\|_2 + \eta$. Held-out evaluation therefore measures the entropy–error association for each predictor and physical regime.

Finite-sample concentration for entropy-binned risk: Fix the score and an entropy bin B independently of an i.i.d. audit sample. Assume the centered errors conditional on membership in B are σ-sub-Gaussian. For $n_B \geq 1$ observations in B and conditional mean risk R_B, the empirical mean $\hat{R}_B$ satisfies, with probability at least $1 - \delta$:

$$|\hat{R}_B - R_B| \leq \sigma\sqrt{\frac{2\log(2/\delta)}{n_B}}.$$

Practical sample-complexity rule: to estimate bin risk within tolerance ϵ with confidence $1 - \delta$, it suffices that

$$n_B \geq \frac{2\sigma^2}{\epsilon^2}\log\frac{2}{\delta}.$$

C.2 Budget-Aware Escalation Strategy

For an escalation policy, let q be the routed fraction and let $\mathcal{C}_{\text{strong}}$ denote the additional backend cost after the AUD pass. Expected cost is

$$\mathcal{C}(q) = \mathcal{C}_{\text{AUD}} + q\mathcal{C}_{\text{strong}}.$$

The special case $1 + q(K - 1)$ assumes a unit-cost first pass, K equal-cost ensemble members, and reuse of that first prediction by the ensemble. The cost must also include separately measured readout overhead.

Using empirical risk maps from validation data, deployment selects:

$$q^\star = \arg\min_q \ \mathcal{R}(q) + \lambda\mathcal{C}(q),$$

Table 4: Five-seed PDEBench matrix. ID ρ_s uses validation-fixed entropy orientation; low/high columns report beta shifts for Darcy and viscosity shifts for Burgers.

Task	AUD adapter	ID $\rho_s \uparrow$	Low AUROC$\uparrow$	High AUROC$\uparrow$
Darcy	MWNO	0.193 ± 0.075	0.499 ± 0.011	0.469 ± 0.011
Darcy	Dense WNO	0.121 ± 0.097	0.490 ± 0.047	0.485 ± 0.044
Darcy	FNO	-0.014 ± 0.026	0.499 ± 0.011	0.497 ± 0.014
Darcy	Fourier-DeepONet	$\mathbf{0.241 \pm 0.009}$	0.500 ± 0.002	0.500 ± 0.002
Burgers	MWNO	0.302 ± 0.083	0.504 ± 0.155	0.485 ± 0.128
Burgers	Dense WNO	0.234 ± 0.024	0.529 ± 0.114	0.462 ± 0.092
Burgers	FNO	0.269 ± 0.126	0.505 ± 0.007	0.495 ± 0.005

with λ encoding latency sensitivity. Larger λ favors near-AUD-only operation; smaller λ favors more escalation.

C.3 Cross-PDE Empirical Validation of the Entropy-Risk Link

The five-seed PDEBench matrix stores validation-fixed per-sample scores for each task, architecture, and shift. The 200-cell sensitivity audit adds validation and test risk-coverage traces. Validation selects the entropy direction and operating point separately for each task.

D Held-Out Deployment Checks and Regime Identification

D.1 Three Diagnostic Outcomes

1. Allocation-margin profile. The high-confusion index (HCI) records the fraction of allocations in $[0.4, 0.6]$. The complete sensitivity surface maps HCI from 0.638 to 0.962 on EM, 0.275 to 0.692 on Burgers, and 0.470 to 0.874 on Darcy. HCI diagnoses the allocation field; held-out ranking criteria determine admissibility.

2. Conditional input-difficulty control. Input roughness and occupied area rank EM error at least as strongly as raw entropy in the factorial suite. Conditioning jointly on contrast, roughness, and occupied area preserves a positive within-stratum entropy–error correlation in every seed and at all six tested matching granularities (0.080–0.161), while linear residualization supplies a complementary excess-entropy check. These tests quantify entropy–error association conditional on the measured input descriptors.

3. Target specificity. ID ranking and shift detection depend on different score distributions. The FNO geometry and PDEBench coefficient-shift results show the resulting target dependence. Separate held-out ranking and shift-detection checks guide the corresponding deployment action.

D.2 Pre-Deployment Diagnostics

Before activating an entropy-based policy, we record:

HCI: HCI $= \#\{g_i \in [0.4, 0.6]\}/n$ records gate ambiguity and supports mechanism inspection.

Entropy-error spatial overlap: We measure intersection over union (IoU) between the top-20% allocation and error regions. The measured EM overlap is 0.184 ± 0.001, with an approximate chance reference of 0.111.

Held-out signal check: For each seed, validation fixes the adapter and score orientation. Ranking passes the signal check when the across-seed 95% confidence interval of validation Spearman correlation is positive. Risk–coverage curves then measure selective prediction performance. Shift detection separately requires the lower 95% confidence bound of validation AUROC to exceed 0.5 for the declared shift. These intervals describe training-seed variation on a fixed validation distribution; independent test fields evaluate each frozen score and policy.

Benchmark-aware policy choice: The signal ledger measures AUD's association with a named target. A disjoint policy-selection split compares its retained error with available input-only policies at the same coverage and label budget, turning signal detection into a comparative screening choice.

Optional error-rate certificate: Given a predeclared error tolerance and target accepted high-error rate q, fix the predictor and score on validation. An independent i.i.d. audit set can either select any threshold using the uniform bound in Proposition 7, or certify a validation-fixed threshold with the tighter exact binomial bound in Corollary 1. Activate when the applicable upper bound is at most q; otherwise route all cases for further checking. The first bound also lower-bounds accepted coverage. These certificates concern accepted high-error rate on the audited distribution; same-budget comparator tests determine the preferred score. The reported validation-selected results remain empirical diagnostics, with a separate audit required to invoke either certificate.

Table 5: Validation-side signal ledger. Intervals describe independently trained seeds on a fixed split; the signal check precedes the disjoint test evaluation. "Pass" records a positive validation association; deployment additionally compares same-budget policies. "Route" sends the declared target to the stronger backend under this check.

Adapter/target	Validation statistic	95% CI	Signal check
MWNO / factorial EM ranking	$\rho_s = 0.478$	$[0.473, 0.484]$	Pass
MWNO / compound EM shift	AUROC $= 0.852$	$[0.843, 0.861]$	Pass
Dense WNO / compound EM shift	AUROC $= 0.856$	$[0.824, 0.887]$	Pass
FNO / geometry shift	AUROC $= 0.547$	$[0.542, 0.552]$	Pass
FNO / frequency shift	AUROC $= 0.598$	$[0.495, 0.701]$	Route
Fourier-DeepONet / frequency	AUROC $= 1.000$	$[1.000, 1.000]$	Pass
Fourier-DeepONet / geometry	AUROC $= 0.496$	$[0.486, 0.506]$	Route
LaMO / horizon-10 ranking	$\rho_s = 0.436$	$[0.329, 0.543]$	Pass
MWNO / Maxwell horizons 1–7	$\rho_s = 0.250$–0.442	lower CI ≥ 0.084	Pass
Darcy 32^2 / ranking	$\rho_s = 0.039$	$[0.025, 0.053]$	Pass
PDEBench / 12 coefficient shifts	AUROC $= 0.464$–0.528	lower CI ≤ 0.488	Route

The intervals summarize training variability on a fixed split. Score and orientation selection use validation data; independent test fields evaluate each frozen choice. For factorial EM ranking, all five validation splits select the same entropy direction and the test risk–coverage curve is monotone in every seed. The compound EM and Fourier-DeepONet signal checks use independent 1,000-ID/1,000-OOD validation pairs, while the PDEBench row summarizes two shifts for each of six task–adapter combinations. FNO splits 1,000 ID validation samples into 500 reference and 500 activation samples and uses 1,000 OOD validation samples; the frozen score then evaluates independent 1,000-ID/1,000-OOD test sets. LaMO selects among step-size, state, and update entropy on each validation split; Maxwell uses horizon-specific validation orientation, whose lower confidence bound remains positive through horizon 7. Their independent test correlations appear in Figure 3 and Tables 6–7. Darcy 32^2 yields independent test $\rho_s = 0.022$ $[-0.002, 0.047]$ after its validation-only signal check, locating a weak-signal boundary of this criterion.

E Internal-Readout and Shift Details

E.1 Entropy Targets the Allocation Boundary

At the same readout cost and on the same fixed allocation field, allocation entropy improves by +0.08 in ρ_s and +0.06 in AUROC over gate variance in the EM deployment track. Gate variance is a second-moment summary of the allocation values, whereas binary entropy directly weights values near the decision boundary where allocation margins are smallest. Both statistics summarize the same allocation field.

E.2 Threshold-Level and Shift-Wise Behavior

At 80% coverage, allocation entropy rejects a larger fraction of high-error samples than alternative internal summaries in the EM deployment track. Table 14 reports fixed-coverage deployment results, and Table 2 reports the factorial comparison at measured inference cost.

The factorial suite evaluates geometry and frequency shifts separately (Table 1) (Hendrycks & Gimpel, 2017; Liu et al., 2020b). The separate deployment protocol reports aggregate OOD AUROC 0.82 ± 0.02.

F Expanded Protocol and Baseline Setup

Matched timing. In a common RTX 4090 benchmark with 20 warm-up and 100 synchronized timed iterations, batch-96 MWNO prediction and prediction-plus-AUD take 0.489 and 0.513 ms per sample, respectively; validation-tuned ten-pass MC-dropout takes 5.070 ms per sample. The entropy readout adds 4.8% to the batched pass; the full prediction-plus-AUD pass is $9.88\times$ faster than MC-dropout. At batch size one, the corresponding latencies are 5.382, 7.718, and 56.605 ms. We use AUD directly for selective prediction or as the first stage of a policy that escalates only the highest-risk samples to a stronger backend.

Non-wavelet readout details. FNO takes the retained 12×12 positive/negative-height Fourier quadrants after every spectral block; each retained coefficient is a component, channel-mean magnitude defines its energy, and a sample-wise top-r_g threshold defines the soft allocation. Fourier-DeepONet uses 256 fixed sensors, a 64-dimensional branch–trunk basis, and coordinate Fourier features; we average each absolute branch–trunk contribution over output channels and spatial points to define one energy per basis index before final summation. The official LaMO state-space adapter (Tiwari et al., 2025) traces tensors indexed by token, scan direction, head/group, and state; validation chooses the regularized energy-dispersion index for selective step size, state coupling, or their input-weighted product. For positive total energy, as the numerical stabilizer $\eta \to 0$, Eq. (6) tends to the Shannon entropy of a component index sampled in proportion to its energy, divided by $\log|\mathcal{C}_b|$. At finite η it remains a regularized dispersion index; no perturbed active-set probability is assigned to LaMO.

Algorithm 1 Single-pass MWNO prediction and allocation-entropy readout

Input: coefficient field a, trained parameters θ
$v \leftarrow \text{Lift}(a)$, $S \leftarrow []$
for $\ell = 1$ to B **do**
  $(v^{LL}, \{v_j^b\}) \leftarrow \text{DWT}(v)$
  for $j \in \{1, 2\}$ **do**
    Compute $a_j(x, y)$ by Eq. (13)
    $\tau_j \leftarrow \text{quantile}(a_j, 1 - r_g)$
    $g_j(x, y) \leftarrow \sigma((a_j - \tau_j)/T)$
    Append all cell values $\mathsf{h}(g_j(x, y))$ to S
  end for
  $\hat{v} \leftarrow \text{IDWT}(\text{Mix}(v^{LL}, \{g_j \odot v_j^b\}))$
  $v \leftarrow v + \hat{v}$
end for
$\hat{u} \leftarrow \text{Projection}(v)$
$H_L \leftarrow$ mean over all appended cells in S (Eq. (15))
Output: $(\hat{u}, H_L)$

Reproduction package. The anonymized supplementary code contains the MWNO, dense-WNO, FNO, Fourier-DeepONet, and crossed Haar/db2 readout implementations. The factorial suite adds dense WNO, FNO, and Fourier-DeepONet adapters, a supervised variance head, frozen-backbone layer probes, independent-checkpoint deep ensembles, MC-dropout, active-set perturbations, physics-conditioned sensitivity, and native 3D rollout evaluation.

Deployment metadata. Tables 14 and 15 contain descriptive end-to-end summaries. The retained records omit repeat counts, dispersion definitions, the Sel.@80% high-error cutoff, and the two-model

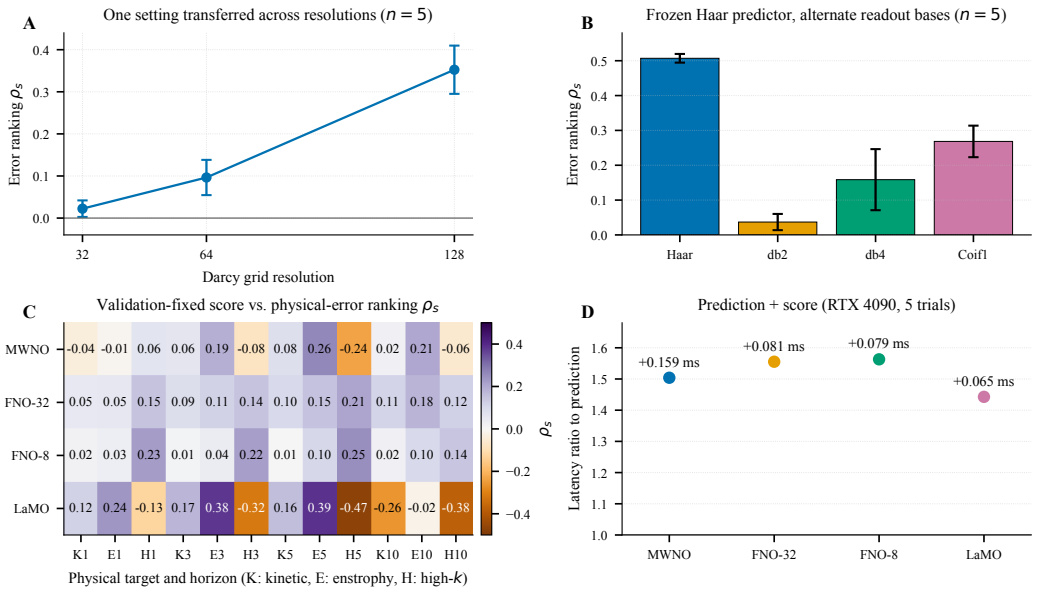

Figure 5: Resolution, basis, physical-target, and timing experiments. **A:** With one transferred Darcy setting, prediction error stays within 0.0536–0.0633 while ranking strengthens across 32^2, 64^2, and 128^2. **B:** Haar, db2, db4, and Coiflet-1 re-read identical frozen Haar-trained predictions as a prediction-fixed mechanism control. **C:** The validation-fixed field-error orientation ranks kinetic-energy (K), enstrophy (E), and high-wavenumber-energy (H) errors. **D:** Synchronized RTX 4090 timing computes only each validation-selected scalar. Error bars show sample SD; Panels A–C use independent model seeds.

Table 6: Ten-seed Navier–Stokes trajectory ranking at 64^2. Each model receives ten frames and predicts the next ten in one pass. Validation selects the allocation score and orientation; entries are mean±sample SD.

Adapter	**Params**	L_2**@10**↓	ρ_1	ρ_3	ρ_5	ρ_{10}	**E-AURC**$_{10}$ ↓
MWNO	157k	$.408 \pm .002$	$.368 \pm .228$	$.381 \pm .236$	$.383 \pm .236$	$.320 \pm .212$	$.0326 \pm .0098$
FNO-32	2365k	$\mathbf{.252 \pm .002}$	$.330 \pm .048$	$.318 \pm .049$	$.297 \pm .053$	$.265 \pm .053$	$.0287 \pm .0028$
FNO-8	148k	$.341 \pm .005$	$.347 \pm .074$	$.396 \pm .075$	$.391 \pm .079$	$.346 \pm .081$	$.0278 \pm .0049$
LaMO	134k	$.423 \pm .007$	$\mathbf{.485 \pm .237}$	$\mathbf{.521 \pm .235}$	$\mathbf{.517 \pm .222}$	$\mathbf{.470 \pm .176}$	$\mathbf{.0246 \pm .0089}$

Table 7: Five-seed native Maxwell-3D rollout ranking. Training uses 32^3 six-channel fields; validation fixes one entropy orientation per model, seed, and horizon before test. Entries report test Spearman mean±sample SD.

Adapter	**Val. rel.** L_2 ↓	**1 step**	**3 steps**	**5 steps**	**7 steps**
MWNO entropy	0.676 ± 0.009	$\mathbf{0.687 \pm 0.044}$	$\mathbf{0.584 \pm 0.048}$	0.493 ± 0.138	0.374 ± 0.217
Dense WNO entropy	0.678 ± 0.010	-0.455 ± 0.078	0.155 ± 0.495	$\mathbf{0.542 \pm 0.085}$	$\mathbf{0.442 \pm 0.110}$
FNO Fourier entropy	$\mathbf{0.616 \pm 0.001}$	0.051 ± 0.156	0.089 ± 0.055	0.049 ± 0.162	0.134 ± 0.121

combination rule, so statistical inference uses the fully specified factorial and trajectory protocols. In the factorial implementation, high-error recall uses the highest-error quintile and rejects the highest-score quintile.

Benchmark datasets. We analyze electromagnetic scattering as the primary interface-dominated case study on 64^2 fields. The ID generator supplies 10,000/1,000/1,000 train/validation/test samples with one to three simple objects, relative permittivity in $[2, 12]$, and $k_0 \in [35, 45]$. Each declared EM shift has two disjoint sets: 1,000 samples for activation validation and 1,000 for final test. Geometry

OOD uses unseen mixed-shape compositions at the ID frequency range; frequency OOD uses simple shapes with $k_0 \in [60, 80]$. The common raw-score comparison uses 1,000 ID validation samples to fix orientation before one evaluation on 1,000-ID/1,000-OOD test sets. The FNO two-sided validation check instead splits the 1,000 ID validation samples into 500 reference and 500 activation samples and combines them with 1,000 OOD validation samples; it then freezes Eq. (8) for one evaluation on the independent 1,000-ID/1,000-OOD test sets. The cross-physics matrix adds PDEBench Darcy flow and one-shot Burgers space–time prediction (Takamoto et al., 2022). Darcy uses $\beta = 1.0$ in-domain and $\beta \in \{0.1, 10.0\}$ as declared coefficient shifts, with fixed index ranges 0–6999/7000–7999/8000–9999 for train/validation/test. The matched Burgers experiment uses viscosity $\nu = 0.01$ in-domain and $\nu \in \{0.001, 0.1\}$ as declared shifts, with the same disjoint-index construction; the OOD activation audit uses validation indices disjoint from the frozen test indices. The global-dynamics track uses 1,200 standard 64^2 Navier–Stokes trajectories with 20 frames (Li et al., 2021); one fixed permutation assigns 800 training, 200 validation, and 200 test trajectories, and no temporal window crosses a split. Every model receives frames 1–10 and predicts frames 11–20 in one pass. The independent experimental unit is a trajectory; pixels and temporal windows remain nested observations. The volumetric track uses PDEArena Maxwell-3D (Gupta & Brandstetter, 2023; Brandstetter et al., 2023): three parameter files supply 300 training trajectories, and disjoint parameter files supply 75 validation and 75 test trajectories. Each trajectory contains eight frames of electric displacement and magnetic fields on a 32^3 grid; we roll one-step models autoregressively to horizons 1, 3, 5, and 7 and fix score orientation independently on validation for each declared horizon. OOD evaluation uses two controlled shifts: geometry shift (unseen shape compositions) and frequency shift ($k_0 \in [60, 80]$ vs training $[35, 45]$).

Baselines. The baselines include input contrast, roughness, occupied area, entropy, Gini, variance, HCI, supervised variance prediction, frozen-backbone layer-probe disagreement, and deep ensembles of independently trained checkpoints. The comparison uses two checkpoint pools. Pool A contains five deterministic factorial MWNO checkpoints; both the zero-head entropy readout and frozen layer probes use this pool, and their identical 0.709 prediction error confirms prediction preservation. Pool B contains five validation-selected dropout-trained MWNO checkpoints; the supervised variance head, last-layer Laplace/KFAC, and MC-dropout use this pool. The first two preserve each deterministic predictor, whereas MC-dropout averages ten stochastic passes. The deep ensemble combines independently trained Pool-A checkpoints, while we train SNGP end to end as its own predictor and place FNO in the matched cross-architecture track. Layer probes attach to blocks 2, 4, and 6 (one-based indexing) of each frozen MWNO. Three auxiliary heads train for 20 epochs to minimize mean relative L_2 prediction loss. Their score is the population variance across these three predictions and the unchanged final prediction, averaged over output channels and spatial cells; the final predictor supplies the reported field error. The supervised baseline freezes each predictor and trains a 1×1 convolution–GELU–1×1 convolution log-variance head by fixed-mean Gaussian negative log-likelihood (NLL) on the training split. Validation selects learning rate from $\{3 \times 10^{-4}, 10^{-3}\}$ and weight decay from $\{10^{-5}, 10^{-4}\}$ before five-seed test evaluation; the head has 4,290 parameters and preserves the prediction. The last-layer Laplace reference (Daxberger et al., 2021) with Kronecker-factored approximate curvature (KFAC) keeps the same prediction fixed, augments the actual feature entering the final 1×1 convolution with its bias coordinate, and estimates activation and output-gradient Kronecker factors from training features and residuals. The spectral-normalized neural Gaussian process (SNGP) reference (Liu et al., 2020a) trains a new operator with spectral normalization on all 50 hidden convolutional/linear layers. It replaces the final projection with 128 random Fourier features (RFFs) and a Gaussian-process (GP) mean layer, then fits the posterior precision from training random features. A dedicated validation sweep selects RFF length scale ℓ and spectral-norm bound b_{SN} by validation E-AURC; each evaluation seed then selects its GP ridge and score orientation on validation. Both baselines operate on predictor features.

These posterior references use the method-specific validation grids above. The uniform settings below govern only the architecture-control factorial runs.

Navier–Stokes architecture control. MWNO has 156,562 trainable parameters, the parameter-matched FNO-8 has 148,010, official LaMO has 133,666, and the strong predictive FNO-32 reference has 2,365,322. All four families receive the same four validation-only learning-rate/weight-decay candidates shown in Table 8, the same trajectory split, epoch cap, validation schedule, and early-stopping rule. NS runs use an 80-epoch cap, validation every epoch, and patience 12. MWNO/FNO-32 use width 32, four blocks, and batch 8; FNO-8 uses width 8 with the same block count and batch

Table 8: Validation-selected settings for learned and multi-pass baselines. Test evaluation uses the selected configuration.

Family	Candidate set	Validation objective	Selected
MWNO–AUD (NS)	lr $\in \{10^{-3}, 3\times10^{-4}\}$, wd $\in \{10^{-5}, 10^{-4}\}$	Relative L_2	$10^{-3}, 10^{-5}$
FNO-32 (NS)	Same four-candidate grid	Relative L_2	$10^{-3}, 10^{-5}$
FNO-8 (NS)	Same four-candidate grid	Relative L_2	$10^{-3}, 10^{-5}$
LaMO (NS)	Same four-candidate grid	Relative L_2	$10^{-3}, 10^{-5}$
MC-dropout (EM)	$p \in \{.05, .10, .20\}$, passes $\in \{10, 20, 40\}$	Oriented ρ_s; fewer passes break ties	$p = .20$, 10 passes
Variance head (EM)	lr $\in \{3\times10^{-4}, 10^{-3}\}$, wd $\in \{10^{-5}, 10^{-4}\}$	Gaussian NLL	$10^{-3}, 10^{-4}$
Deep ensemble (EM)	Independent seed prefixes of size $\{2, 3, 5\}$	Oriented ρ_s	MWNO: 3; FNO: 5
Laplace/KFAC (EM)	prior $\in \{10^{-4}, \ldots, 10\}$, damping $\in \{10^{-6}, \ldots, 10^{-2}\}$	E-AURC	Per-seed validation
SNGP (EM)	$\ell \in \{.5, 1, 2\}$; $b_{\mathrm{SN}} \in \{.95, 1, 1.05\}$; ridge $\in \{10^{-5}, \ldots, 1\}$	E-AURC	$\ell = 2$, $b_{\mathrm{SN}} = 1$; per-seed ridge

Table 9: Five-seed posterior baselines on EM. ID metrics use validation-selected score orientation. OOD columns use raw posterior variance, where larger values denote higher uncertainty, and report geometry/frequency AUROC.

Method	ID rel. $L_2 \downarrow$	ID $\rho_s \uparrow$	ID E-AURC$\downarrow$	OOD AUROC G/F$\uparrow$	ms/sample$\downarrow$
Laplace/KFAC	$.700 \pm .002$	$.484 \pm .077$	$.0867 \pm .0103$	$.548 \pm .009/.894 \pm .238$	$.541 \pm .001$
SNGP	$.777 \pm .013$	$.344 \pm .038$	$.0497 \pm .0078$	$.480 \pm .008/1.000 \pm .000$	$.889 \pm .002$

size; LaMO uses width 32, two blocks, and batch 64. After validation freezes the architecture configuration, we train seeds 0–9 independently and pair them by seed for comparison. For every trained model, validation selects the allocation statistic and its orientation before test evaluation at horizons 1, 3, 5, and 10. The state-space implementation follows the official LaMO model (Tiwari et al., 2025); the entropy instrumentation reads direction–state update groups, selective step sizes, and state couplings without adding a prediction head or altering the predicted field. The primary endpoint is horizon-10 raw E-AURC. Exact paired-seed sign-flip tests use Holm correction over all six architecture pairs, separately within each metric. Main-text p-values follow this six-pair correction. The ranking-difference interval is an unadjusted paired-seed Student-t interval for that comparison.

Maxwell-3D optimization. All three families use width 12, three blocks, batch 4, a 40-epoch cap, AdamW with learning rate 5×10^{-4} and weight decay 10^{-4}, and cosine learning-rate decay. Validation runs each epoch; improvement must exceed 10^{-4}, with patience eight. Wavelet models use two levels and FNO uses four retained modes per axis.

Architecture-control optimization. The matched EM architecture track uses:

- Same optimizer family (AdamW), learning rate (1e-3), weight decay (1e-5)
- Same scheduler (OneCycleLR)
- Same epoch budget (100 epochs)
- Same 16-bit floating-point (FP16) mixed precision
- Same locked 10,000/1,000/1,000 EM train/validation/test split
- Five independent seeds for the factorial suite
- Validation-only score orientation and operating-point selection

Runtime measurements include full uncertainty overhead. Every run retains checkpoints, raw and validation-fixed scores, per-sample errors, latency, and test metrics. The deep-ensemble track evaluates fixed seed-prefix ensembles of sizes 2, 3, and 5, then selects size using validation ranking.

Table 10: Five-seed conformal comparison on disjoint scale-fit and calibration halves. Costs use deterministic MWNO prediction as the reference.

Method	Coverage	Mean rel. threshold↓	Batch cost↓
Unscaled split conformal	$95.34 \pm 0.09\%$	0.959 ± 0.005	$\mathbf{1.00}\times$
MC-dropout conformal (10)	$95.80 \pm 0.60\%$	$\mathbf{0.950 \pm 0.016}$	$10.36\times$
Entropy-scaled conformal	$95.76 \pm 0.17\%$	0.964 ± 0.003	$1.08\times$

G Conformal Details and Shift Behavior

For entropy-scaled conformal prediction, we divide validation data into two disjoint parts. The first orients the score and fits ridge-regularized least squares from standardized score to $\log(\epsilon + 10^{-8})$:

$$\sigma_\beta(a) = \exp\left(\beta_0 + \beta_1 \frac{S_{\mathrm{rank}}(a) - \mu_S}{\sigma_S + \eta}\right), \tag{17}$$

and the second calibrates the quantile of relative residuals divided by $\sigma_\beta(a)$. This makes allocation entropy determine sample-wise relative-error threshold while preserving the standard split-conformal rank guarantee.

Conformal evaluation, summarized in Table 10, follows split conformal principles (Vovk et al., 2005; Lei et al., 2018; Angelopoulos & Bates, 2023) and gives near-nominal coverage for entropy-scaled relative-error nonconformity. The unscaled row uses the same deterministic MWNO predictions and the same odd-indexed conformal-calibration half, but fixes $\sigma_\beta(a) = 1$; it therefore isolates the contribution of entropy-dependent threshold allocation. Entropy scaling changes coverage from $95.34 \pm 0.09\%$ to $95.76 \pm 0.17\%$ and mean relative-error threshold from 0.959 ± 0.005 to 0.964 ± 0.003. For conditional-coverage evaluation, we sort test samples by their validation-fitted entropy-scaled threshold and form four equal-count groups. Relative to the unscaled threshold on the same groups, scaling reduces mean absolute deviation from 95% coverage from 1.80 to 1.44 percentage points. Worst-group undercoverage falls from 2.84 to 1.08 points, and coverage spread falls from 5.92 to 4.40 points over five seeds. These are empirical conditional diagnostics; the formal guarantee remains marginal coverage under exchangeability. Entropy scaling has a mean relative-error threshold of 0.964 at $1.08\times$ cost; tuned MC-dropout gives 0.950 at $10.36\times$ cost.

Even-indexed validation samples fit score direction and the log-linear scale σ_β in Eq. (17); odd-indexed validation samples calibrate $q_{1-\alpha}$ from $r_i = \epsilon_i / \sigma_\beta(a_i)$. For both entropy and MC-dropout, the score-specific scale uses the same ridge objective $\min_{\beta_0,\beta_1} \sum_i [\log(\epsilon_i + 10^{-8}) - \beta_0 - \beta_1 z_i]^2 + 10^{-6}\beta_1^2$; We standardize z_i as oriented entropy or predictive variance, respectively. The resulting relative-error prediction set is

$$\hat{C}_\alpha(a) = \left\{ v : \frac{\|v - \hat{u}(a)\|_2}{\|v\|_2 + \eta} \le q_{1-\alpha}\sigma_\beta(a) \right\}, \tag{18}$$

For n_{cal} calibration scores, $q_{1-\alpha}$ is their kth order statistic with $k = \lceil (n_{\mathrm{cal}} + 1)(1 - \alpha) \rceil$; set $q_{1-\alpha} = +\infty$ when $k > n_{\mathrm{cal}}$. A recent calibration buffer supports recalibration under shift (Gibbs & Candès, 2021). The marginal-coverage guarantee requires exchangeability between the buffer and future evaluation samples within the shifted regime. We fix one permutation, use nested prefixes of size $m \in \{25, 50, 100, 200\}$ only to update the normalized-residual quantile, and evaluate on the disjoint complement.

The geometry intervention reaches the target with 100–200 recent samples; the stronger frequency intervention reaches 95.30% with 200. The threshold increase responds to the larger residual scale, and the disjoint evaluation measures out-of-buffer coverage recovery. For $q_{1-\alpha}\sigma_\beta(a) > 1$, the relative-error set is unbounded; at equality it can also be unbounded. The recovered coverage rates can therefore accompany unbounded prediction sets. Related efficient variants include conformalized quantile regression (CQR) (Romano et al., 2019).

Table 11: Five-seed shifted entropy-scaled conformal audit at 95% target coverage. A recent buffer updates only the normalized-residual quantile; the predictor, entropy orientation, and scale map remain fixed.

Regime	Buffer m	Coverage	Mean rel. threshold
ID	0	$95.76 \pm 0.17\%$	0.964 ± 0.003
Geometry shift	0	$93.92 \pm 0.08\%$	0.948 ± 0.004
Geometry shift	100	$96.02 \pm 0.48\%$	0.988 ± 0.011
Geometry shift	200	$95.78 \pm 0.16\%$	0.977 ± 0.007
Frequency shift	0	$0.28 \pm 0.38\%$	0.943 ± 0.017
Frequency shift	25	$85.66 \pm 4.97\%$	2.023 ± 0.689
Frequency shift	50	$88.76 \pm 1.92\%$	2.090 ± 0.771
Frequency shift	100	$91.29 \pm 1.25\%$	2.144 ± 0.820
Frequency shift	200	$95.30 \pm 0.77\%$	2.237 ± 0.874

Table 12: Resolution transfer and frozen-predictor basis comparison. Every entry reports mean±sample SD over five seeds. The resolution suite transfers one configuration. The basis suite re-reads the same frozen Haar-trained predictors, leaving every prediction fixed.

A. Resolution transfer on Darcy

Resolution	Rel. $L_2 \downarrow$	$\rho_s \uparrow$	Norm. E-AURC↓	Samples/seed
32^2	$.0619 \pm .0012$	$.022 \pm .020$	$.449 \pm .094$	2000
64^2	$.0633 \pm .0014$	$.096 \pm .042$	$.414 \pm .068$	2000
128^2	$.0536 \pm .0010$	$.352 \pm .057$	$.207 \pm .073$	2000

B. Alternative readout bases on fixed EM predictions

Readout basis	$\rho_s \uparrow$	E-AURC↓
Haar	$.507 \pm .012$	$.0769 \pm .0026$
db2	$.037 \pm .023$	$.1577 \pm .0028$
db4	$.159 \pm .088$	$.1315 \pm .0098$
Coiflet-1	$.268 \pm .045$	$.1197 \pm .0054$

H Detailed Ablation Results

Descriptive deployment summaries. Tables 14 and 15 retain the recorded end-to-end operating points. Their source records omit repeat counts, dispersion definitions, the high-error cutoff, and the two-model combination rule (Appendix F). The two-pass configuration reports ranking 0.72 and AUROC 0.85 at $2.13\times$ cost.

The descriptive EM deployment summaries report higher entropy ranking than the alternative internal statistics. On EM, allocation entropy achieves $\rho_s = 0.67 \pm 0.03$, AUROC 0.82 ± 0.02, and Sel.@80% $31 \pm 2\%$. Alternative statistics are weaker: $\mathrm{Var}(g)$ $(0.59/0.76/25\%)$, Gini $(0.61/0.78/27\%)$, $\max(g)$ $(0.53/0.72/21\%)$, and feature norm $\|v\|$ $(0.56/0.74/23\%)$.

In the factorial control suite, excess entropy H_E serves as an anti-confound diagnostic, separate from the deployed score. Validation-only residualization against input contrast, roughness, and occupied area yields ID $\rho_s = 0.068 \pm 0.015$, which quantifies the association remaining after these measured covariates. The FNO adapter reads retained Fourier coefficients; Table 1 reports its five-seed ID ranking and geometry/frequency shift detection.

I Physics-Stratified Sensitivity Details

The architecture matrix contains 30 independently trained task–architecture runs. A separate sensitivity study uses ten top-r_g MWNO checkpoints: two tasks and five seeds per task. For each checkpoint, we evaluate $r_g \in \{0.05, 0.10, 0.15, 0.20, 0.30\}$ and $T \in \{0.10, 0.15, 0.20, 0.30\}$, giving 20 cells per seed and 200 seed–task–parameter evaluations. Each cell holds weights fixed, changes the prediction-

Table 13: Physical-target ranking at horizon 10 and validation-selected timing. The positive-interval count spans three targets at horizons 1, 3, 5, and 10. The physical-target comparison reuses the field-error orientation from validation; timing uses batch 8, ten warmups, and five trials of 20 synchronized repetitions.

A. Physical-error ranking

Architecture	Kinetic $\rho_s \uparrow$	Enstrophy $\rho_s \uparrow$	High-k $\rho_s \uparrow$	Positive CI cells
MWNO	$.020 \pm .107$	$.205 \pm .185$	$-.062 \pm .102$	$3/12$
FNO-32	$.111 \pm .078$	$.178 \pm .044$	$.119 \pm .048$	$12/12$
FNO-8	$.020 \pm .065$	$.101 \pm .062$	$.142 \pm .060$	$6/12$
LaMO	$-.263 \pm .314$	$-.020 \pm .363$	$-.378 \pm .144$	$5/12$

B. Prediction and selected-score latency (ms/sample)

Architecture	Selected score	Prediction↓	Pred.+score↓	Ratio↓
MWNO	Allocation entropy	.315	.473	$1.504\times$
FNO-32	Allocation entropy	.147	.228	$1.555\times$
FNO-8	Allocation entropy	.141	.220	$1.563\times$
LaMO	Update entropy	.148	.213	$1.443\times$

Table 14: Deployment-track EM triage. Costs use WNO at batch size 96 as the reference. The two-model row uses additional inference. Descriptive results; protocol metadata appear in Appendix F.

Method	Passes	$\rho_s \uparrow$	AUROC↑	Sel.@80%↑	Batch cost↓
Gate Gini	1	0.61 ± 0.03	0.78 ± 0.02	$27 \pm 2\%$	$\mathbf{1.13}\times$
Multi-head ensemble	1	0.61 ± 0.03	0.78 ± 0.03	$27 \pm 2\%$	$1.20\times$
AUD H_L	1	0.67 ± 0.03	0.82 ± 0.02	$31 \pm 2\%$	$\mathbf{1.13}\times$
AUD H_L + 2-model	2	$\mathbf{0.72 \pm 0.03}$	$\mathbf{0.85 \pm 0.02}$	$\mathbf{35 \pm 2\%}$	$2.13\times$

coupled masks, and recomputes predictions, entropy, and errors. Each cell selects score direction on ID validation and freezes it for ID and both shifted test sets. Figure 6 summarizes test ranking; selecting a new operating point also requires validation prediction error and risk–coverage checks. The grid measures frozen-checkpoint response on Burgers and Darcy; it contains no per-setting retraining or turbulent Navier–Stokes sensitivity measurements.

On Burgers, the 20-cell mean ID ρ_s range is 0.254 to 0.443; low/high viscosity-shift AUROC ranges are 0.480 to 0.522 and 0.465 to 0.516. On Darcy, the mean ID ρ_s range is 0.045 to 0.386; low/high beta-shift AUROC ranges are 0.486 to 0.508 and 0.473 to 0.491. The corresponding validation HCI ranges are 0.275 to 0.692 and 0.470 to 0.874. Temperature acts relative to component-magnitude scale, which varies across physical regimes. For a frozen energy field, r_g changes membership at the threshold and T changes softness; prediction-coupled mask changes can also alter subsequent block energies.

J Cross-Architecture Mechanism Summary

For the component-perturbation audit, we add independent Gaussian energy noise at relative root-mean-square levels $0.001, 0.005, 0.01, 0.05$ and recompute the top-r_g active set after each draw. Predictions and trained weights remain fixed. Re-estimating the threshold couples margin changes, so this empirical audit is distinct from the fixed-margin sub-Gaussian assumption in Proposition 3.

The mechanism analysis combines allocation flips (Figure 2), conditional input controls and a random-extractor comparison (Section 5), frozen-predictor basis changes (Table 12), and spatial overlap (Appendix D). These controls measure representation dependence and the association between entropy and instability. Identifying the contributions of epistemic uncertainty, physical competition, and representation entanglement requires interventions that vary those factors separately.

Table 15: Internal-score ablation in the EM deployment track. Costs use WNO at batch 96 as the reference, matching Table 14. The separate five-seed conformal protocol appears in Table 10.

Method	$\rho_s \uparrow$	**AUROC**$\uparrow$	**Sel.@80%**$\uparrow$	**Batch cost**$\downarrow$
Feature norm	0.56 ± 0.04	0.74 ± 0.03	$23 \pm 2\%$	$\mathbf{1.0}\times$
Max gate	0.53 ± 0.04	0.72 ± 0.03	$21 \pm 3\%$	$1.13\times$
Gate variance	0.59 ± 0.04	0.76 ± 0.03	$25 \pm 2\%$	$1.13\times$
Gate Gini	0.61 ± 0.03	0.78 ± 0.02	$27 \pm 2\%$	$1.13\times$
Multi-head ensemble	0.61 ± 0.03	0.78 ± 0.03	$27 \pm 2\%$	$1.20\times$
AUD H_L	0.67 ± 0.03	0.82 ± 0.02	$31 \pm 2\%$	$1.13\times$
AUD H_L + 2-model	$\mathbf{0.72 \pm 0.03}$	$\mathbf{0.85 \pm 0.02}$	$\mathbf{35 \pm 2\%}$	$2.13\times$

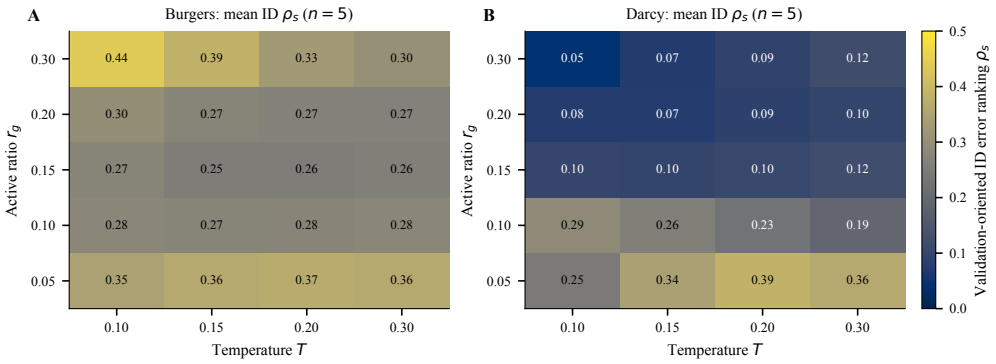

Figure 6: Physics-conditioned hyperparameter response over a five-seed grid. Burgers retains positive ranking throughout all 20 cells. Darcy favors a smaller retained fraction and peaks near $r_g = 0.05$. Each cell changes prediction-coupled masks on frozen top-r_g MWNO checkpoints. Colors show mean test ranking over five seeds on a shared scale.

K Input-Difficulty Risk Models

An exploratory five-seed EM comparison measures allocation entropy's contribution beyond input contrast, roughness, and occupied high-permittivity area. Each fixed MWNO predictor supplies 1,000 calibration examples and 1,000 disjoint ID test examples. For each seed and learner, we fit three risk models: entropy only, input descriptors only, and input descriptors plus entropy. All three predict the calibration percentile rank of relative field error; fitting and hyperparameter selection use calibration data only.

Ridge regression maps each covariate through its calibration empirical distribution, standardizes it, and selects the penalty from $\{10^{-4}, 10^{-2}, 1, 100\}$. Histogram gradient boosting uses raw covariates, learning rate 0.05, and 150 iterations; selection spans maximum leaf counts $\{7, 15, 31\}$, minimum leaf sizes $\{20, 50\}$, and regularization strengths $\{0, 1\}$. Five-fold shuffled calibration cross-validation selects by mean squared error, and the selected model refits on all calibration examples before test evaluation. The ridge empirical-distribution transform uses the full calibration split; each cross-validation training fold fits its own standardizer. This analysis uses the five existing seeds and reports exploratory test comparisons.

Adding entropy leaves mean ridge ranking unchanged at three-decimal precision and lowers mean gradient-boosting ranking. Both learners show higher mean E-AURC with the added entropy covariate. These comparisons place the representation-dependent associations in Section 5 alongside fitted input-difficulty models on the same EM task.

Table 16: Exploratory input-difficulty comparison on five fixed MWNO predictors. Every row fits an auxiliary risk model, including the entropy-only rows. Entries show test mean±sample SD across five seeds; the two metrics summarize error ranking.

Learner	Inputs to risk model	$\rho_s \uparrow$	E-AURC$\downarrow$
Ridge	Entropy only	0.4593 ± 0.0046	0.0768 ± 0.0015
Ridge	Input descriptors	0.5760 ± 0.0048	0.0637 ± 0.0015
Ridge	Input descriptors + entropy	0.5759 ± 0.0049	0.0639 ± 0.0015
Gradient boosting	Entropy only	0.4342 ± 0.0080	0.0802 ± 0.0020
Gradient boosting	Input descriptors	0.5588 ± 0.0049	0.0643 ± 0.0012
Gradient boosting	Input descriptors + entropy	0.5542 ± 0.0055	0.0654 ± 0.0019

L Independent Label-Budget Triage

Application and independent data. The label-budget experiment evaluates onboarding a frozen predictor with limited calibration labels and no access to its original training labels. We retain the five matched MWNO checkpoints and generate 1,000 new EM ID examples with seed 20260923 under the original 64×64 physical configuration. The configuration uses one to three simple objects, object permittivity in $[2, 12]$, and wavenumber in $[35, 45]$. All five models evaluate the same examples; each method uses the same model-specific field errors and calibration subsets. We fix the protocol, comparison groups, and endpoints before evaluating the new cohort. CUDA inference with batch size 96 and TensorFloat-32 (TF32) matches the stored calibration backend; each model passes an old-validation replay before new-test evaluation.

Shared label budgets and scalar policies. We use $m \in \{8, 16, 32, 64, 128, 256, 1000\}$ labels from the existing 1,000-example calibration split. Twenty random permutations define nested subsets for the first six budgets, shared across methods and models. The 1,000-label reference uses one fit per model. Every raw scalar chooses its sign by minimizing mean error among the lowest-score $\lfloor 0.8m \rfloor$ calibration examples; ties prefer positive orientation. Selected input chooses among contrast, roughness, and occupied area under the same criterion. Selected alternative score chooses among Gini, variance, active ratio, HCI, and the input norm. The active-ratio score averages soft allocation probabilities. HCI measures the fraction of allocations in $[0.4, 0.6]$. The input-norm control is the root mean square (RMS) of the model input tensor; the stored field name is `feature_norm`. Fixed random priorities, independent of scores and errors, resolve ranking ties.

Fitted risk models. Input ridge uses the three input descriptors; score-summary ridge uses the five allocation statistics and input RMS. Two further models combine input descriptors with entropy or all six scalar scores. Four-fold cross-validation within each label budget selects the ridge penalty from $\{10^{-4}, 10^{-2}, 1, 100\}$ by squared-error loss. Each training fold fits its own mid-empirical-distribution transforms for covariates and error targets, and its own covariate standardizer. Each selected model refits on its m calibration examples and predicts risk on the independent test set. The original fully supervised variance head in Table 2 uses a separate training budget.

Fixed endpoint and uncertainty. Every test policy retains the 800 lowest-score examples, giving 80% test coverage. Normalized retained error divides their mean relative L_2 error by the model's mean error over all 1,000 test examples. Uniform random screening has expected normalized error one. The primary endpoint averages this metric over $\log_2 m$ from 3 to 8 using trapezoidal integration divided by five; the 1,000-label reference contributes only a separate endpoint. We average subset repetitions within each model before summarizing the five independently trained models. Sample SD and descriptive paired t intervals quantify model-training variation conditional on the shared calibration and test examples. The experiment measures batch screening; solver escalation and online threshold transfer require separate evaluation.

Observed application value. AUD's primary value of 0.9693 corresponds to a 3.07% reduction in retained error relative to random screening (Table 17). Selected input, input ridge, and score-summary ridge achieve lower integrated error in all five models. Their paired differences from AUD are -0.00089, -0.00279, and -0.00113, with descriptive 95% intervals $[-0.00114, -0.00064]$,

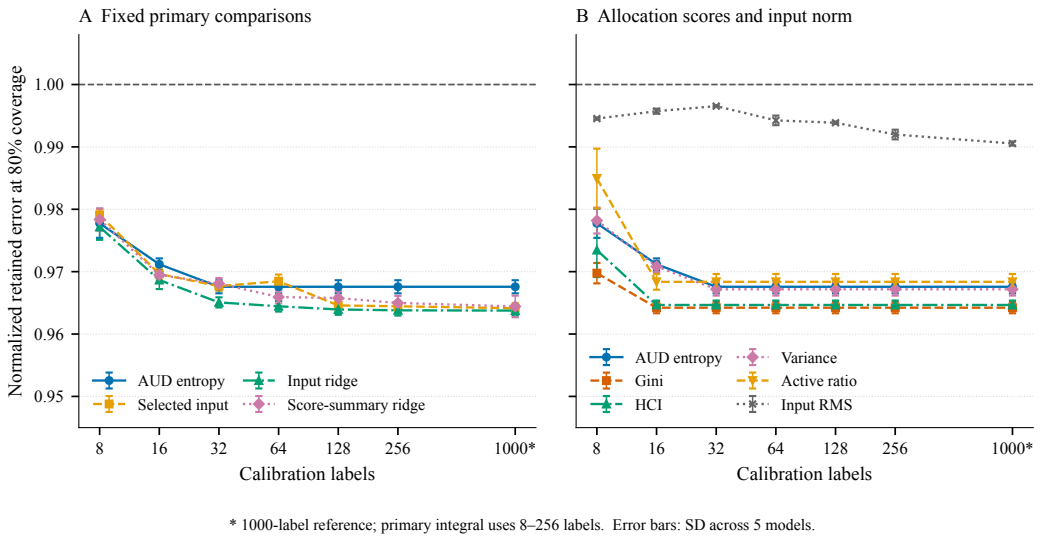

Figure 7: Independent EM triage with limited calibration labels. **A:** AUD and the three primary comparator policies. **B:** Five allocation statistics and the input-norm control. Lower normalized retained error indicates better screening at 80% test coverage; the dashed line marks the random-screening expectation. Error bars show sample SD over five models after averaging calibration subsets.

Table 17: All 15 policies on the independent 1,000-example EM test set. The primary column integrates normalized retained error over 8–256 calibration labels and reports mean±sample SD across five models. Endpoint columns report means at 8 and 1,000 labels. Every column measures retained error at 80% test coverage; lower is better.

Policy	**Budget integral↓**	**8 labels↓**	**1,000 labels↓**
Raw allocation readouts			
AUD entropy	0.9693 ± 0.0007	0.9778	0.9676
Gini	0.9648 ± 0.0007	0.9698	0.9642
Variance	0.9690 ± 0.0007	0.9782	0.9672
Active ratio	0.9700 ± 0.0013	0.9849	0.9684
HCI	0.9655 ± 0.0007	0.9735	0.9647
Raw input descriptors			
Contrast	0.9891 ± 0.0004	1.0018	0.9843
Roughness	0.9664 ± 0.0008	0.9778	0.9651
Occupied area	0.9650 ± 0.0008	0.9727	0.9641
Input RMS	0.9947 ± 0.0004	0.9945	0.9905
Calibration-selected scalar policies			
Selected input	0.9684 ± 0.0009	0.9791	0.9641
Selected alternative score	0.9674 ± 0.0007	0.9773	0.9646
Fitted risk models			
Input ridge	0.9665 ± 0.0009	0.9772	0.9638
Input + entropy ridge	0.9667 ± 0.0010	0.9762	0.9638
Score-summary ridge	0.9682 ± 0.0010	0.9784	0.9644
Input + score-summary ridge	0.9666 ± 0.0009	0.9756	0.9622

$[-0.00306, -0.00253]$, and $[-0.00155, -0.00071]$, respectively. Gini and HCI also achieve lower integrated error than entropy, at 0.9648 and 0.9655. Figure 7 shows that input ridge has lower mean retained error at every tested budget. These observations establish useful AUD screening in this setting, with stronger screening from inexpensive input and allocation statistics.

M Readout Choices and Triage Decisions

Fixed-prediction comparison. The four-basis experiment changes the readout of each frozen Haar-trained MWNO while retaining its predictions and errors. The basis and MWNO perturbation suites reuse five dropout-trained checkpoints (dropout rate 0.2) in deterministic evaluation mode; Table 1 uses the separate factorial checkpoint pool. For each input to a wavelet block, each basis produces two levels of detail coefficients, with circular padding where the filter length requires it. At each location, the readout takes the Euclidean norm across the three detail subbands and averages over channels. It then computes quantile-threshold entropy with $r_g = 0.15$ and $T = 0.2$ separately at each block and level, averages over locations within a level, and gives each block–level pair equal weight. This common pooling rule defines the basis suite; the deployed wavelet readout in Eq. (15) pools scale–location cells within blocks. Within-suite comparisons hold the pooling rule fixed.

The basis intervention changes filter support and coefficient-magnitude distributions at a fixed absolute temperature. For a uniform energy rescaling $E' = cE$ with $c > 0$, quantile equivariance gives $m' = cm$, so $\mathsf{h}(\sigma(cm/T)) = \mathsf{h}(\sigma(m/(T/c)))$. Thus magnitude scale sets the effective temperature. All predictors train with Haar, and alternative bases also change their alignment with that training representation. The comparison measures the combined effect of these readout choices on identical predictions.

Decision protocol. A secondary analysis of the existing scores quantifies their effect on triage. For each of five models and each basis, the original 1,000-example validation split fixes the entropy sign by its Spearman correlation with error. We rank all 1,000 test examples and accept the 800 lowest oriented scores. Stable ordering by sample index resolves ties without using errors; none of the 20 model–basis cases has a tie across the acceptance boundary. This rule implements batch screening at fixed empirical coverage. Normalized retained error divides the accepted examples' mean relative error by the same model's mean relative error over the full test set; random screening has expectation one. Decision disagreement is the fraction of all test examples whose accept/reject status differs from Haar. Table 18 reports every basis, including paired differences computed within each model. The five models share the same physical test examples; sample SD describes variation across these models conditional on that test set.

Table 18: Changing the readout changes decisions on fixed predictions. At 80% test coverage, each basis retains 800 of the same 1,000 examples. Entries report mean$\pm$sample SD across five frozen Haar-trained models. Paired differences and decision disagreement use Haar as the reference; disagreement counts both acceptance-to-rejection and rejection-to-acceptance changes.

Readout	Norm. retained error↓	Paired difference	Decisions changed
Haar	0.9651 ± 0.0021	0.0000 ± 0.0000	$0.0 \pm 0.0\%$
db2	0.9883 ± 0.0048	0.0231 ± 0.0045	$22.5 \pm 8.0\%$
db4	0.9879 ± 0.0075	0.0228 ± 0.0061	$22.2 \pm 5.8\%$
Coiflet-1	0.9763 ± 0.0033	0.0112 ± 0.0040	$11.8 \pm 2.8\%$

Switching from Haar to db2 or db4 changes more than one fifth of test decisions on average, while Coiflet-1 changes 11.8%. Every alternative has higher retained error than Haar in all five models under this protocol. These decision changes connect the score distributions in Figure 4 to the examples a screening policy accepts.

Signed mechanism comparison. Allocation instability and task error have different empirical relationships with raw entropy. We compute the three Spearman correlations among raw entropy, allocation-flip rate, and relative field error from the original perturbation records. MWNO raw entropy has validation error correlation -0.525 ± 0.013 and test error correlation -0.507 ± 0.012; each of the five validation splits selects negative orientation. Fourier-DeepONet has validation and test correlations 0.485 ± 0.034 and 0.477 ± 0.040, respectively, with positive orientation in all five models. Table 19 retains raw entropy and flip rates when computing the mechanism correlations.

Table 19: Signed mechanism correlations on the existing perturbation test set. Relative RMS gives the energy-noise amplitude. Entries report mean±sample SD of Spearman correlations over five models; raw entropy and flip rates retain their original signs. Figure 2B plots the entropy–flip column.

Adapter	Relative RMS	Entropy–flip	Flip–error
MWNO	0.001	0.842 ± 0.023	-0.548 ± 0.005
MWNO	0.005	0.850 ± 0.019	-0.550 ± 0.004
MWNO	0.01	0.858 ± 0.015	-0.552 ± 0.004
MWNO	0.05	0.901 ± 0.009	-0.553 ± 0.004
Fourier-DeepONet	0.001	0.010 ± 0.042	-0.011 ± 0.036
Fourier-DeepONet	0.005	0.031 ± 0.054	0.008 ± 0.058
Fourier-DeepONet	0.01	0.030 ± 0.058	0.014 ± 0.060
Fourier-DeepONet	0.05	0.004 ± 0.086	0.003 ± 0.062

MWNO entropy tracks allocation flips positively at every perturbation level, while flip rates track field error negatively in every model at all four levels. Fourier-DeepONet's near-zero mean flip correlations coexist with positive raw entropy–error ranking. The perturbation analysis therefore distinguishes the allocation-stability endpoint from the task-error endpoint. The validation step in Eq. (7) determines which direction of the readout supports the latter. These paired descriptive analyses reuse existing test data and quantify observed relationships; the independent label-budget evaluation remains the separate experiment in Appendix L.

Error modes on independent fields. We inspect the two error modes directly on the independent 1,000-example EM test set of Appendix L; this is a post-hoc diagnostic on the same five frozen MWNO models. Within each model, the upper and lower relative-error quintiles define high and low error. The lower and upper *raw* entropy quintiles define relatively sharper and more ambiguous mean allocations; a low mean does not imply that every component has a large threshold margin. The original 1,000-example calibration split fixes the sign by Eq. (7) before we form test risk quintiles. Each quintile contains 200 examples; the five models share the same physical test fields.

Table 20: Two error modes on independent EM fields. Each cell counts intersections of two 200-example quintiles among 1,000 test examples per frozen model. Raw columns test whether relatively sharp allocation predicts low error and ambiguous allocation predicts high error; oriented columns count contrary-end cases after calibration fixes the score sign. Random ordering expects 40 cases per intersection. Models share physical fields.

Model seed	High error, low raw H	Low error, high raw H	High error, low risk	Low error, high risk
0	70	94	11	5
1	74	96	9	5
2	70	95	13	5
3	67	94	11	5
4	66	96	13	4
Mean±SD	69.4 ± 3.1	95.0 ± 1.0	11.4 ± 1.7	4.8 ± 0.4

All five calibration splits select negative orientation; independent test raw entropy–error correlations range from -0.477 to -0.462. Orientation reduces the two contrary-end intersections to 9–13 and 4–5 cases per model, while leaving individual errors on both sides of the ranking.

Architecture-specific error modes. The matched EM cohort in Table 1 supplies a second, earlier held-out test set with the same 1,000 physical examples for all four adapters. Within each model, its calibration split fixes the entropy sign; we then count the high-error/low-risk and low-error/high-risk quintile intersections using the protocol above. Table 21 reports means and sample SD across five training seeds per adapter. The signed raw correlation shows why orientation differs between wavelet/Fourier and branch–trunk representations.

Table 21: Adapter-specific error modes on the matched EM test set. Each contrary-end cell counts examples in the intersection of two 200-example quintiles among 1,000 shared test fields, reported as mean±sample SD over five trained models. Calibration fixes each sign before test evaluation. Random ordering expects 40 cases per intersection.

Adapter	Cal. sign	Raw entropy–error ρ_s	High error, low risk	Low error, high risk
MWNO	$-$	-0.459 ± 0.005	15.2 ± 1.1	4.8 ± 0.4
Dense WNO	$-$	-0.492 ± 0.003	14.2 ± 2.6	2.8 ± 0.8
FNO	$-$	-0.238 ± 0.025	10.0 ± 2.4	36.6 ± 3.1
Fourier-DeepONet	$+$	0.477 ± 0.040	30.8 ± 2.2	1.2 ± 0.8

FNO's 36.6 low-error/high-risk cases per model approach the random-order expectation of 40, whereas Fourier-DeepONet's 30.8 high-error/low-risk cases reveal a different blind spot. These two error modes differ despite positive validation-oriented correlations in all four adapters. The raw-sign reversal also shows that allocation ambiguity has no architecture-independent direction as a physical-error signal.

Validation-based energy-scale control. A global energy-scale correction leaves differences between the readout policies on the fixed predictors. For each model, basis b, and block–level group g, we compute $s_{b,g} = \sqrt{\mathrm{mean}_{i,c}\, E_{b,g,c}(a_i)^2}$ over all validation inputs and locations. We set $T_{b,g} = 0.2 s_{b,g}/s_{\mathrm{Haar},g}$, which matches the validation RMS of temperature-normalized energies E/T to Haar while retaining the common $r_g = 0.15$ and pooling rule. The Haar temperature remains 0.2. Scale estimation uses validation inputs alone; each temperature setting separately selects its score direction from validation errors. Both settings then accept the 800 lowest-scoring test examples.

We fix this control before collecting its corrected scores and reuse the original five predictors and data splits. Replaying the raw readouts reproduces all original acceptance sets; maximum absolute differences are 1.79×10^{-7} for raw scores and 2.38×10^{-7} for relative errors. Haar's raw and corrected scores agree elementwise. The primary descriptive comparison is the change in each alternative's normalized retained-error gap from Haar, paired within model and measured against the same run's raw readout. Table 22 reports all eight conditions and all three paired gap changes.

The mean db2 gap decreases by 0.0047, with sample SD 0.0071; three of five models improve. Its validation-selected orientation changes from negative to positive in four models, while the fifth remains positive. The reported policy effect therefore includes this direction selection. The db4 mean gap changes by -0.0003, and the Coiflet-1 mean gap increases by 0.0023. Every corrected alternative retains higher error than Haar in all five models. Global validation-RMS scale matching thus leaves the observed policy differences. Appendix N tests training-basis alignment in new models; sample-specific scale, margin distributions, and filter support remain distinct explanations.

Table 22: Validation-RMS temperature matching on the same five frozen predictors. Panel A reports test mean±sample SD at 80% coverage; decision disagreement uses Haar and counts all test examples. Panel B reports alternative-minus-Haar error gaps and their paired change (RMS-matched minus fixed-T). Each condition selects its orientation on validation.

A. Complete readout comparison

Readout	Temperature	$\rho_s \uparrow$	Norm. retained error↓	Decisions changed
Haar	Fixed T	0.507 ± 0.012	0.9651 ± 0.0021	$0.0 \pm 0.0\%$
Haar	RMS-matched	0.507 ± 0.012	0.9651 ± 0.0021	$0.0 \pm 0.0\%$
db2	Fixed T	0.037 ± 0.023	0.9883 ± 0.0048	$22.5 \pm 8.0\%$
db2	RMS-matched	0.176 ± 0.101	0.9836 ± 0.0073	$31.0 \pm 3.2\%$
db4	Fixed T	0.159 ± 0.088	0.9879 ± 0.0075	$22.2 \pm 5.8\%$
db4	RMS-matched	0.156 ± 0.100	0.9877 ± 0.0079	$23.0 \pm 6.9\%$
Coiflet-1	Fixed T	0.268 ± 0.045	0.9763 ± 0.0033	$11.8 \pm 2.8\%$
Coiflet-1	RMS-matched	0.262 ± 0.048	0.9786 ± 0.0046	$13.8 \pm 2.8\%$

B. Paired change in the error gap from Haar

Readout	Fixed-T gap	RMS-matched gap	Paired gap change
db2	0.0231 ± 0.0045	0.0185 ± 0.0072	-0.0047 ± 0.0071
db4	0.0228 ± 0.0061	0.0226 ± 0.0063	-0.0003 ± 0.0013
Coiflet-1	0.0112 ± 0.0040	0.0135 ± 0.0046	0.0023 ± 0.0023

Table 23: Training-basis matching depends on the task. Five paired seeds: prediction error and retained/full-test mean error at 80% coverage (mean±sample SD). Both readouts use neutral calibration-RMS temperatures. Δ subtracts matched from mismatched ratios within a predictor; positive values favor matching.

Task	Train basis	Rel. L_2	Haar readout	db2 readout	Δ
EM	Haar	0.89993 ± 0.00130	0.99647 ± 0.00019	0.99685 ± 0.00041	0.00039 ± 0.00032
EM	db2	0.89141 ± 0.00408	0.99581 ± 0.00026	0.99488 ± 0.00039	0.00093 ± 0.00017
Darcy	Haar	0.09386 ± 0.00458	0.96543 ± 0.01336	0.98816 ± 0.01972	0.02273 ± 0.02220
Darcy	db2	0.08656 ± 0.00419	0.96982 ± 0.00633	0.98988 ± 0.01292	-0.02006 ± 0.01179

N Crossed Training and Readout Representations

Question and experimental unit. We test whether matching training and readout bases lowers retained error in both training arms and both tasks. All 20 models complete before test analysis. Five seeds pair identical learned initial tensors and minibatch order; readouts share predictions within each model. This width-32, four-block, two-level, no-dropout cohort differs from the main benchmark and historical mechanism models.

Independent data and fixed roles. Each task shares 2,000 training, 500 checkpoint-selection, 500 calibration, and 1,000 test examples across models. EM uses the first 2,000 original training files and splits the original validation files into disjoint halves; seed 2026092401 generates a fresh test set from the original physical configuration. Inputs and per-example maximum-magnitude target normalization follow the EM task, with eight boundary cells excluded from relative error.

Controlled Darcy is a new smooth-coefficient task, separate from PDEBench. We solve $-\nabla\cdot(a\nabla u) = 1$ on the unit square with zero Dirichlet boundaries and 32^2 interior nodes at $x_i = (i+1)/33$. For each sample, standard Gaussian grid noise is Fourier-filtered by $\exp[-2\pi^2(0.15)^2(k_x^2 + k_y^2)]$, centered, and divided by its spatial RMS to obtain z; $a = \exp(z)$. Independent split streams descend from seed 2026092402. A five-point flux discretization uses harmonic interior-face coefficients and the adjacent interior coefficient at boundary faces. Float64 sparse direct solves have maximum relative residual 2.90×10^{-13}. Constant- and variable-coefficient manufactured solutions show convergence toward second order on $16^2/32^2/64^2$ grids. Training data alone fit the four input-channel normalizers and one scalar target RMS; evaluation uses all interior nodes.

Matched training and score construction. Orthogonal periodic transforms with adjoint boundary folding pass reconstruction, energy, gradient, and Haar-compatibility checks. The parameter counts are 156,042 for EM and 156,009 for Darcy. Every model completes 100 epochs, 3,200 updates, batch 64, AdamW with learning rate 3×10^{-4} and weight decay 10^{-4}, cosine decay to 10^{-6}, and gradient clipping at 1. The loss is mean relative L_2; we retain the earliest minimum selection-error checkpoint. Training uses float32, soft top-0.15 masks at $T = 0.2$, no auxiliary losses, and no mixed precision or TF32. Paired arms share a GPU and randomize their training order.

From block inputs, component energy is $\sqrt{\sum_{d=1}^{3} h_d^2 + 10^{-8}}$ averaged over channels. The retained count per level is $\max(1, \lfloor 0.15n \rfloor)$, matching the prediction gate. We average statistics equally over the eight block–level groups. The primary neutral-RMS condition sets $T_{b,g} = 0.2\sqrt{\mathrm{mean}_{\mathrm{calibration}}\, E_{b,g}^2}$, making calibration RMS of E/T equal to 5 for either basis. This symmetric rule differs from Appendix M's Haar anchor. Fixed secondary conditions comprise raw $T = 0.2$, distribution Gini of the probabilities, and HCI in $[0.4, 0.6]$. Calibration errors select each condition's sign; stable sample order resolves score ties. We accept exactly 800 test examples and normalize their mean error by the same predictor's full-test mean.

Primary contrast and finding. Write $R_{tr,rd}$ for normalized retained error, where H and D denote Haar and db2. The primary matched-basis effect is $D_{\mathrm{match}} = \frac{1}{2}[(R_{H,D} - R_{H,H}) + (R_{D,H} - R_{D,D})]$. Positive values favor matching on average. EM favors matching in every seed of both arms, with mean effect 0.000656 ± 0.000202. Darcy has $D_{\mathrm{match}} = 0.001333 \pm 0.008049$, with three positive

seeds; Haar wins in every model under either training basis. The opposing arm effects appear in Table 23. We report descriptive model SD conditional on the common data.

Complete secondary comparisons. Table 25 retains every mode and statistic. Neutral Gini shares the EM matching pattern and Darcy arm asymmetry; neutral HCI has positive Darcy arm means as a secondary observation. Boundary ties occur in 48 of the 240 policies, all using HCI; primary entropy has neither boundary ties nor constant vectors.

The controlled comparison uses a smaller EM training cohort, which yields the prediction errors in Table 23; controlled Darcy samples one smooth coefficient distribution. Records retain all histories, checkpoints, scores, errors, and accepted indices. Seed-0 replay reproduces errors and all 12 readout scores for 16 test examples in each of the four cells.

Table 24: Absolute retained relative errors for the primary crossed-basis comparison (mean±sample SD), selection epochs in seed order 0–4, and fraction of all test decisions changed by swapping the readout. All models complete 100 epochs.

Task	Train	Retained: Haar	Retained: db2	Selected epochs	Decisions changed
EM	Haar	0.89675 ± 0.00133	0.89709 ± 0.00147	97,99,99,94,100	$16.32 \pm 1.63\%$
EM	db2	0.88767 ± 0.00424	0.88684 ± 0.00435	97,99,99,97,100	$14.00 \pm 1.05\%$
Darcy	Haar	0.09065 ± 0.00518	0.09274 ± 0.00453	98,99,100,100,99	$29.32 \pm 4.21\%$
Darcy	db2	0.08394 ± 0.00397	0.08569 ± 0.00438	99,97,99,90,99	$25.04 \pm 3.25\%$

Table 25: All crossed readout comparisons at 80% coverage: normalized retained error and paired matching effect, mean±sample SD over five seeds. $R_{tr,rd}$ indexes training then readout basis (H: Haar; D: db2). Neutral entropy is the primary condition; raw temperature, Gini, and HCI are secondary.

Temperature	Score	$R_{H,H}$	$R_{H,D}$	$R_{D,H}$	$R_{D,D}$	D_{match}
EM						
Raw	Entropy	0.9970 ± 0.0001	0.9989 ± 0.0009	0.9980 ± 0.0020	0.9975 ± 0.0007	0.00121 ± 0.00136
Raw	Gini	0.9960 ± 0.0001	0.9985 ± 0.0007	0.9995 ± 0.0014	0.9956 ± 0.0004	0.00323 ± 0.00112
Raw	HCI	0.9964 ± 0.0004	0.9989 ± 0.0002	0.9994 ± 0.0011	0.9976 ± 0.0013	0.00217 ± 0.00074
Neutral	Entropy	0.9965 ± 0.0002	0.9969 ± 0.0004	0.9958 ± 0.0003	0.9949 ± 0.0004	0.00066 ± 0.00020
Neutral	Gini	0.9966 ± 0.0003	0.9968 ± 0.0002	0.9958 ± 0.0001	0.9948 ± 0.0003	0.00057 ± 0.00023
Neutral	HCI	0.9962 ± 0.0003	0.9963 ± 0.0002	0.9964 ± 0.0003	0.9955 ± 0.0003	0.00046 ± 0.00038
Controlled Darcy						
Raw	Entropy	0.9783 ± 0.0167	0.9885 ± 0.0208	0.9768 ± 0.0107	0.9924 ± 0.0145	-0.00266 ± 0.01833
Raw	Gini	0.9732 ± 0.0164	0.9886 ± 0.0209	0.9728 ± 0.0151	0.9897 ± 0.0055	-0.00073 ± 0.00996
Raw	HCI	0.9778 ± 0.0149	0.9836 ± 0.0151	0.9704 ± 0.0048	0.9832 ± 0.0105	-0.00355 ± 0.01095
Neutral	Entropy	0.9654 ± 0.0134	0.9882 ± 0.0197	0.9698 ± 0.0063	0.9899 ± 0.0129	0.00133 ± 0.00805
Neutral	Gini	0.9659 ± 0.0146	0.9896 ± 0.0209	0.9706 ± 0.0059	0.9847 ± 0.0130	0.00485 ± 0.00928
Neutral	HCI	0.9630 ± 0.0157	0.9856 ± 0.0093	0.9875 ± 0.0065	0.9865 ± 0.0084	0.01178 ± 0.01050

Table 26: Every primary paired contrast. Positive Δ_H or Δ_D favors the matched readout within that training arm; D_{match} averages the two. The replication unit is a paired training-seed block, giving five blocks per task.

Task	Seed	Δ_H	Δ_D	D_{match}
EM	0	0.000081	0.000714	0.000397
EM	1	0.000840	0.001056	0.000948
EM	2	0.000582	0.000809	0.000695
EM	3	0.000243	0.001117	0.000680
EM	4	0.000183	0.000936	0.000560
Darcy	0	0.018559	−0.031494	−0.006467
Darcy	1	0.058207	−0.031246	0.013480
Darcy	2	0.005317	−0.003813	0.000752
Darcy	3	0.027933	−0.019723	0.004105
Darcy	4	0.003629	−0.014035	−0.005203

O PROSPECTIVE TRANSFER TO ADVECTION–DIFFUSION

Question and locked cohorts. We test an existing allocation readout on a newly generated physical task and newly trained predictors. We specified the primary cohort of five MWNO and five FNO training seeds before generating fields or training models. Both use the same split, selection rule, and fixed entropy parameters. A subsequent five-seed FNO width-8 cohort tests the primary FNO result at approximately matched predictor capacity. We fixed its design after the primary outcomes and before training any FNO-8 model. The trained model is the replication unit; all models share the physical test examples.

Controlled discrete problem. We solve $-\nu\Delta u + \mathbf{v} \cdot \nabla u + 0.1u = f$ on the unit square with zero Dirichlet boundary values. A five-point diffusion stencil and first-order upwind advection on a 30×30 interior grid define the discrete target; sparse LU gives relative linear-system residuals below 2×10^{-14}. Manufactured-solution checks give relative continuum errors 0.0007 with zero advection and 0.0941 in an advection case. Smooth random forcing uses Dirichlet sine modes 1–5 with coefficients decaying as $(m^2 + n^2)^{-1}$. The input channels are forcing, horizontal and vertical velocity, diffusivity, and spatial coordinates. The ID regime crosses $\nu \in \{0.03, 0.06\}$, speed $\in \{0.5, 1.0\}$, and four cardinal directions. The named shift changes diffusivity to $\{0.015, 0.03\}$ and speed to $\{1.5, 2.0\}$.

Independent random streams generate 2,000 training, 500 checkpoint-selection, 500 score-calibration, 1,000 ID-test, 500 shift-validation, and 1,000 shift-test fields. We balance each regime's 16 parameter cells within every role. One training-set target RMS normalizes every target. The ID test is independent of the EM, Darcy, Burgers, Navier–Stokes, and Maxwell datasets elsewhere in this paper.

Training and frozen readouts. The primary models have four blocks, width 32, and either two-level Haar MWNO masks or 12-mode FNO spectral layers. AdamW trains each model for 80 full epochs with learning rate 3×10^{-4}, weight decay 10^{-4}, batch 64, cosine decay to 10^{-6}, and mean per-example relative L_2 loss. The earliest minimum checkpoint-selection error chooses the frozen model. The capacity follow-up uses four FNO blocks at width 8, with the same optimizer, data, epochs, and seeds. AUD retains $r_g = 0.15, T = 0.2$ in both families. Calibration fixes the risk-score sign separately for each model and policy; test labels do not select checkpoints, signs, or hyperparameters.

The input baselines are forcing roughness and speed divided by diffusivity. A two-feature ridge predictor fits log error on the same 500 calibration labels, choosing its penalty by four-fold calibration-only cross-validation. Gini and high-ambiguity-cell incidence use the same frozen allocation masks. FNO also receives a Haar readout of frozen block inputs; it leaves predictions fixed. Stable sample order resolves tied scores at the 80% coverage boundary. For the named shift, an independent ID/shift validation pair selects the raw-score sign, then a disjoint ID/shift pair supplies AUROC. The directly available speed/diffusivity ratio is a shift baseline.

Primary outcomes and capacity control. Table 27 reports every trained model. Raw entropy has negative calibration correlation with ID error in all 15 models, so each calibration split selects the negative sign. The average ID correlations are 0.313 ± 0.043 (MWNO-32), 0.320 ± 0.025 (FNO-32), and 0.310 ± 0.050 (FNO-8). The seed-paired MWNO-32 minus FNO-8 correlation difference is 0.003 ± 0.066; their predictive relative errors differ by -0.066 ± 0.020. FNO-32 has 2,364,833 parameters versus 156,073 for MWNO-32, while FNO-8 has 147,881. Similar rank correlations therefore survive the capacity control, but predictive accuracy and retained-error ratios differ. All five seeds in each cohort complete 80 epochs; the readout changes no prediction on replay (maximum absolute difference zero).

Table 27: All advection–diffusion seeds on the same 1,000 ID-test fields. Normalized retained error accepts 800 fields; the input ridge uses the same 500 score-calibration labels. Shift AUROC compares disjoint 1,000-ID and 1,000-shift test fields. The FNO-8 rows are a capacity follow-up.

Model	Seed	Rel. L_2	AUD ρ_s	AUD retained	Input ridge retained	AUD shift AUROC
MWNO-32	0	.2782	.3470	.9507	.9351	.6070
	1	.2751	.3213	.9514	.9353	.7332
	2	.2683	.2422	.9634	.9345	.5133
	3	.2858	.3480	.9518	.9404	.7440
	4	.2760	.3047	.9558	.9389	.6757
FNO-32	0	.1684	.3332	.9285	.9322	.5125
	1	.1713	.2852	.9563	.9291	.5175
	2	.1690	.3328	.9382	.9284	.5087
	3	.1642	.3451	.9398	.9323	.5108
	4	.1717	.3023	.9493	.9333	.5137
FNO-8	0	.3199	.3106	.9616	.9514	.4909
	1	.3320	.2432	.9801	.9646	.5246
	2	.3538	.3372	.9558	.9528	.5248
	3	.3724	.3748	.9527	.9491	.4902
	4	.3343	.2848	.9752	.9560	.4886

Input and allocation controls. Table 28 reports the fixed comparison policies. The input ridge gives the lowest mean retained error in MWNO-32 and FNO-32; input roughness and ridge tie to displayed precision in FNO-8. The FNO Fourier readout ranks ID error more strongly than the frozen Haar readout at both capacities. A post-hoc exploratory ridge fit adds raw entropy to the two input descriptors using the same calibration-only cross-validation. Its retained-error change relative to input-only ridge is $+0.0005 \pm 0.0025$ for MWNO-32, $+0.0002 \pm 0.0021$ for FNO-32, and -0.0031 ± 0.0029 for FNO-8, where negative favors adding entropy. This exploratory addition improves two, three, and five of the five seeds, respectively; the prespecified standalone comparison remains the primary endpoint.

Table 28: All prespecified advection–diffusion screening controls: mean±sample SD of normalized retained error at 80% coverage (lower is better). Random-order expectation is 1.000. Input ridge uses two input descriptors and calibration labels. Haar is a fixed-prediction FNO representation control.

Score	MWNO-32	FNO-32	FNO-8
AUD entropy	$.955 \pm .005$	$.942 \pm .011$	$.965 \pm .012$
Input roughness	$.947 \pm .002$	$.934 \pm .002$	$.955 \pm .006$
Input speed/diffusivity	$.963 \pm .002$	$1.022 \pm .003$	$1.022 \pm .006$
Input ridge	$\mathbf{.937 \pm .003}$	$\mathbf{.931 \pm .002}$	$\mathbf{.955 \pm .006}$
Allocation Gini	$.949 \pm .004$	$.954 \pm .017$	$.971 \pm .009$
High-ambiguity-cell incidence	$.951 \pm .003$	$.982 \pm .010$	$.989 \pm .006$
Frozen Haar readout	n/a	$.988 \pm .010$	$.995 \pm .012$

Error modes and named shift. The AUD score rejects $38.5 \pm 3.3\%$, $32.9 \pm 3.4\%$, and $31.5 \pm 2.5\%$ of the highest-error quintile in MWNO-32, FNO-32, and FNO-8, respectively, while rejecting 20% overall. Among 200 high-error fields, the low-risk quintile still contains 16.8 ± 4.1, 15.8 ± 1.9, and 14.8 ± 3.0 fields; among 200 low-error fields, the high-risk quintile contains 14.8 ± 1.8, 12.6 ± 3.2, and 16.2 ± 6.1. These counts directly quantify confidently wrong and conservatively flagged correct predictions after calibration-based orientation. The named-shift AUROC is 0.655 ± 0.096, 0.513 ± 0.003, and 0.504 ± 0.019 for AUD, compared with 1.000 for the input speed/diffusivity ratio in all 15 models. The shift deliberately changes the support of these supplied physical parameters; this baseline has direct access to that shift. The 80% cutoff has no AUD, roughness, Gini, or Haar score ties; the discrete speed/diffusivity score ties in all 15 models, and high-ambiguity-cell incidence ties in 14. Stable input order fixes those policies' acceptance sets.