FROM LEARNING OPERATORS TO INFERRING OPERATORS: ORIA FOR BAYESIAN REASONING IN FUNCTION SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific inference requires reasoning about unknown systems from limited observations and using uncertainty to guide further discovery. We introduce ORIA, a continuum framework for Bayesian operator inference in function space. Trained under priors over operators, ORIA conditions on a few input–output function pairs to produce joint posterior predictive distributions for unseen systems, without test-time optimization. Its continuum architecture supports prediction across discretizations, while its joint predictive distribution captures uncertainty and dependencies across entire response fields. This enables uncertainty-aware prediction of global field-structure-dependent physical quantities and guides the selection of informative experiments. Comparisons with exact Bayesian references assess posterior fidelity, and experiments on diverse physical systems demonstrate up to an order of magnitude lower relative error than baselines. Posterior-guided measurement selection reduces average prediction error by 13–18% relative to random selection. We further characterize how reliability depends on the training prior and how prior revision can improve transfer. ORIA paves a promising path toward efficient foundation models for scientific computing that know what they do not know.

1 INTRODUCTION

Neural operators have begun to extend scientific machine learning beyond finite-dimensional input–output mappings toward function-to-function mappings. Once trained, they can make fast predictions for new inputs across different grids and resolutions (Lu et al., 2021; Li et al., 2021; Kovachki et al., 2023). These benefits, however, usually come with an implicit assumption: the target operator, or operator family, is fixed by the training task; that is, the model is trained on known physical systems with varying inputs and parameters. The operator itself is not generally treated as a latent object to be inferred. A substantially different operator therefore typically requires retraining or task-specific adaptation (Yang et al., 2023). This can be costly, as training neural operators often relies on large amounts of high-fidelity simulation data, and becomes more challenging still when data must be collected through physical experiments (Azizzadenesheli et al., 2024).

Scientific discovery and engineering inference often proceed differently. In many cases, the governing system is not fully known in advance; instead, it must be inferred from limited observations and prior knowledge (Brunton et al., 2016; Raissi & Karniadakis, 2018). This naturally suggests a Bayesian view, in which observations update our beliefs about the possible underlying systems. Predictions can then account for the remaining uncertainty rather than committing to a single fitted model (Kennedy & O'Hagan, 2001; Stuart, 2010; Yang et al., 2020). Toward such scientific intelligence, a central question is: can a frozen model infer a previously unseen operator from only a few input–output function pairs and predict its response to new inputs? More importantly, how reliable is the predictive distribution?

Recently, scientific foundation models have begun to answer this question through operator learning across equations and physical regimes, thereby increasing the diversity of systems represented within their frozen weights (Liu et al., 2024; Sun et al., 2025; Hao et al., 2024). Among these efforts, ICON takes a more consequential step by moving adaptation from gradient updates into

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.

context, showing that a frozen model can infer the response of an unseen operator from only a few functional demonstrations (Yang et al., 2023). This progress, however, leaves unresolved whether the predictive distribution faithfully represents uncertainty, since scientific inference requires more than good model fit.

Neural processes provide a probabilistic framework, amortizing context observations into predictive distributions over functions, with Gaussian and transformer variants further capturing dependencies across target locations (Garnelo et al., 2018; Bruinsma et al., 2021; Markou et al., 2022; Nguyen & Grover, 2022). Their standard formulation, however, infers an unknown function from point observations rather than an unknown operator from multiple functional demonstrations. Prior-fitted networks make the Bayesian target explicit: specifying the distribution over training tasks as a prior defines the posterior predictive induced by the observed context (Müller et al., 2022; Hollmann et al., 2025), providing a reference against which learned uncertainty can be audited. Such auditing is especially important in function space, where uncertainty is a joint distribution over the entire response field, encoding both uncertainty at individual locations and its coupling across space or time. A model may recover the correct mean and calibrated uncertainty at each location while still assigning the wrong probabilities to complete response functions, as we demonstrate experimentally. Such errors matter whenever scientific conclusions depend on correlations across the field.

In this work, our contributions are:

1. **We introduce ORIA (OpeRator Inference Architecture), a continuum framework for Bayesian operator inference in function space.** Trained under operator priors, ORIA produces an explicit joint posterior predictive density over response fields from a few input–output function pairs, supporting prediction across discretizations without test-time optimization. Exact Bayesian comparisons on a conjugate operator family, for which the posterior predictive can be computed exactly, validate joint predictive fidelity and reveal how pointwise calibration can conceal errors in uncertainty over field distribution.
2. **We demonstrate accurate few-shot prediction and the value of joint uncertainty on unseen physical systems.** Frozen ORIA achieves predictive gains over classical estimators, task-fitted neural operators, transformer neural processes, in-context operator predictors, and fine-tuned pretrained models, with error reductions of up to an order of magnitude in representative comparisons. ORIA's joint predictive distribution captures cross-location dependencies, supporting calibrated uncertainty quantification for global physical quantities such as spatially aggregated responses.
3. **We turn posterior uncertainty into scientific action and characterize its reliability under prior shifts.** Posterior-guided experiment selection reduces prediction-error AUC by 13.1–18.3% relative to random selection. Controlled prior shifts further reveal how reliability depends on the training prior and how observed demonstrations can help diagnose noise and structural mismatch. Revising the prior improves external transfer but degrades performance on tasks drawn from the original prior.

2 Related Work

In-context operator learning. ICON predicts the responses of unseen operators from functional demonstrations without updating model parameters (Yang et al., 2023). GenICON and Zebra extend this paradigm with generative and autoregressive formulations, respectively (Zhang et al., 2025; Serrano et al., 2025), inducing predictive distributions through generated response samples or distributions over discrete output tokens. ORIA instead represents predictions at queried locations with an explicit joint Gaussian density. This representation provides a tractable joint likelihood and preserves dependencies needed to quantify uncertainty in functionals of the response.

Functional uncertainty. Neural processes model distributions over functions conditioned on pointwise observations, with Gaussian and transformer-based variants capturing dependencies across target locations (Bruinsma et al., 2021; Nguyen & Grover, 2022). Function-valued uncertainty has also been studied in LUNO and probabilistic neural operators (Magnani et al., 2025; Bülte et al., 2025). In in-context operator inference, however, the conditioning structure is different: multiple input–output function pairs share a common unknown operator, and prediction is required for

a new input function. We study joint predictive uncertainty under this shared-operator conditioning structure, including uncertainty in spatial aggregates derived from the predicted response.

Prior-conditioned prediction and continuous representations. PFNs and TabPFN learn amortized posterior predictive approximations from synthetic task priors (Müller et al., 2022; Hollmann et al., 2025), motivating our use of sampled operator priors to train in-context predictors. Separately, work on representation equivalence and continuum attention studies how models can respect the relationship between continuous functions and their discrete representations (Bartolucci et al., 2023; Calvello et al., 2025). ORIA combines prior-conditioned learning with quadrature-aware attention to operate on discretized functional observations while targeting a predictive law over the underlying response.

3 Problem Setting

Task. An unknown operator $\mathcal{G} : \mathcal{F} \to \mathcal{U}$ maps an input function f to a response function $u = \mathcal{G}[f]$ on Ω. Given a few input–response pairs $(f_i, u_i)_{i=1}^m$ from this operator, called *demonstrations*, and a query $f_\star$, we must predict a noisy measurement of $u_\star = \mathcal{G}[f_\star]$ and its uncertainty.

Discretization. Functions are recorded by their values on a measurement grid $X = (x_1, \dots, x_M)$: we write $v(X) = (v(x_1), \dots, v(x_M))$, and likewise $v(Y)$ at the N locations $Y = (y_1, \dots, y_N)$ where predictions are requested. The grid carries quadrature weights $w = (w_1, \dots, w_M)$, the share of the domain each x_k covers, taken from the trapezoidal rule on the sorted grid so that $\sum_k w_k\, v(x_k) \approx \int_\Omega v \,\mathrm{d}x$. They are what lets one predictor treat coarse and fine grids consistently.

Input. Everything the predictor observes is the tuple

$$\mathcal{C} = \big(f_\star(X),\ \mathcal{D}_m^X,\ X,\ w,\ Y\big)\,, \qquad \mathcal{D}_m^X = \{(f_i(X),\ u_i(X) + \boldsymbol{\varepsilon}_i)\}_{i=1}^m\,, \tag{1}$$

with $\mathcal{D}_m^X$ the demonstrations ($f_i(X)$ the sampled inputs, $u_i(X)+\varepsilon_i$ the noisily measured responses), $f_\star(X)$ the sampled query input, and X, w, Y the grid information. The noise $\boldsymbol{\varepsilon}_i$ has zero mean and variance σ_ε^2, because real acquisition is noisy, and a noisy target also keeps the predictive law well defined.

Output. The predictor returns a law $q(\,\cdot \mid \mathcal{C})$ on $\mathbb{R}^N$ for a future noisy measurement of the query response at Y:

$$\tilde{u}_\star = u_\star(Y) + \boldsymbol{\varepsilon}_\star \ \sim\ q(\,\cdot \mid \mathcal{C}). \tag{2}$$

With m demonstrations the operator is not pinned down, so an honest answer is a *joint* law rather than N marginals.

Bayesian target. We assume the unknown system comes from a distribution over plausible systems. The operator $\mathcal{G}$ and noise variance σ_ε^2 are drawn from a prior Π, which captures our beliefs before observing any data. Conditioning on the demonstrations $\mathcal{C}$ shifts these beliefs toward systems that better explain the observations, and averaging their predictions for the query gives the Bayes posterior predictive $P^\star(\cdot \mid \mathcal{C})$. This is the optimal prediction under the assumed prior. Each draw of $(\mathcal{G}, \sigma_\varepsilon^2)$, together with $\mathcal{C}$, forms an *episode*; episodes may be simulated from Π or collected experimentally.

Learning problem. Computing $P^\star$ from scratch for every new system is impractical, so we learn across episodes an amortized predictor that approximates the *Bayes prediction map*

$$\Phi^\star :\ \mathcal{C} \mapsto P^\star(\,\cdot \mid \mathcal{C}). \tag{3}$$

At deployment, prediction for an unseen system conditions on m demonstrations without test-time optimization. Different discretizations of the same episode should yield consistent predictive laws.

4 METHOD: ORIA

ORIA uses an encoder–decoder architecture (Figure 6, Appendix G) to approximate the map $\Phi^\star$ of equation 3, returning a joint Gaussian from demonstrations in a single forward pass with frozen weights.

Throughout the following detailed description,indices i range over examples $\{1, \dots, m, \star\}$, j over demonstrations $\{1, \dots, m\}$, k and l over grid points $\{1, \dots, M\}$, and n over evaluation locations $\{1, \dots, N\}$; the residual-stream representation is $h_i(x_k) \in \mathbb{R}^d$.

4.1 TOKENS

Each example contributes one token per grid point, containing the input $f_i(x_k)$, observed response $u_i(x_k) + \varepsilon_{ik}$, a demonstration flag, and the Fourier positional encoding

$$\psi(x) = \big(1,\ \cos 2\pi x,\ \dots,\ \cos 2\pi K x,\ \sin 2\pi x,\ \dots,\ \sin 2\pi K x\big) \in \mathbb{R}^{2K+1}. \tag{4}$$

Separate positional channels give a continuous, grid-independent encoding. The query's M tokens zero the response channel and clear the demonstration flag (Fig. 6b); a linear map embeds every token into $\mathbb{R}^d$.

4.2 QUADRATURE ATTENTION

ORIA uses scaled dot-product attention with a quadrature-weighted softmax,

$$\mathrm{softmax}_w(s)_l = \frac{w_l e^{s_l}}{\sum_{l'} w_{l'} e^{s_{l'}}}, \tag{5}$$

equivalently adding $\log w_l$ to the logits, so the discrete attention sum approximates a continuous-domain integral. Grid dependence enters only through $\psi(x_l)$ and w_l, so the same parameters transfer across grid sizes and spacings.

4.3 ENCODER

Each encoder block applies two quadrature-attention sub-layers (Fig. 6c): intra-function attention across each function's domain, followed by inter-function attention across examples. Both use pre-norm residual updates:

$$\begin{aligned} \bar{h}_i(x_k) &:= \mathrm{LN}\big(h_i(x_k)\big), & h_i^{\mathrm{intra}}(x_k) &:= h_i(x_k) + \Delta_i^{\mathrm{intra}}(x_k), \\ \bar{h}_i^{\mathrm{intra}}(x_k) &:= \mathrm{LN}\big(h_i^{\mathrm{intra}}(x_k)\big), & h_i^{\mathrm{inter}}(x_k) &:= h_i^{\mathrm{intra}}(x_k) + \Delta_i^{\mathrm{inter}}(x_k). \end{aligned} \tag{6}$$

Intra-function attention. Within each function, evaluation points attend to one another:

$$\begin{gathered} s_{kl}^i = \tfrac{1}{\sqrt{d_h}} \big\langle W_Q^{\mathrm{intra}} \bar{h}_i(x_k),\ W_K^{\mathrm{intra}} \bar{h}_i(x_l) \big\rangle, \qquad \alpha_{kl}^i = \frac{w_l\, e^{s_{kl}^i}}{\sum_{l'} w_{l'}\, e^{s_{kl'}^i}}, \\ \Delta_i^{\mathrm{intra}}(x_k) = \sum_l \alpha_{kl}^i\, W_V^{\mathrm{intra}} \bar{h}_i(x_l). \end{gathered} \tag{7}$$

With quadrature weights w_l, the update discretizes $\int \alpha^i(x_k, x)\, W_V^{\mathrm{intra}} \bar{h}_i(x)\, \mathrm{d}\mu(x)$.

Inter-function attention. Each example attends to the m demonstrations, integrating pointwise similarities at matching locations:

$$\begin{gathered} s_{ij} = \tfrac{1}{\sqrt{d_h}} \sum_k w_k \big\langle W_Q^{\mathrm{inter}} \bar{h}_i^{\mathrm{intra}}(x_k),\ W_K^{\mathrm{inter}} \bar{h}_j^{\mathrm{intra}}(x_k) \big\rangle, \qquad \beta_{ij} = \frac{e^{s_{ij}}}{\sum_{j' \le m} e^{s_{ij'}}}, \\ \Delta_i^{\mathrm{inter}}(x_k) = \sum_{j \le m} \beta_{ij}\, W_V^{\mathrm{inter}} \bar{h}_j^{\mathrm{intra}}(x_k). \end{gathered} \tag{8}$$

The score s_{ij} approximates the L^2 inner product $\langle W_Q^{\mathrm{inter}} \bar{h}_i^{\mathrm{intra}},\ W_K^{\mathrm{inter}} \bar{h}_j^{\mathrm{intra}} \rangle$; β_{ij} is therefore one function-level affinity shared across locations. Restricting keys and values to $j \le m$ enforces one-way demonstration-to-query flow; uniform context replication and its Bayesian scope are discussed in Appendix F.

4.4 DECODER

Let $z_\star(x_k)$ be the query's final encoder representation. At each target y_n, the decoder initializes $t_n = \psi(y_n)$. In each block, $\bar{t}_n = \mathrm{LN}(t_n)$ attends to $z_\star$ through evaluation attention (Fig. 6d):

$$s_{nk} = \tfrac{1}{\sqrt{d_h}}\big\langle W_Q^{\mathrm{eval}}\bar{t}_n,\ W_K^{\mathrm{eval}} z_\star(x_k)\big\rangle, \qquad \gamma_{nk} = \frac{w_k e^{s_{nk}}}{\sum_{k'} w_{k'} e^{s_{nk'}}},$$
$$\Delta_n^{\mathrm{eval}} = \sum_k \gamma_{nk}\, W_V^{\mathrm{eval}} z_\star(x_k). \tag{9}$$

The residual update is $t_n^{\mathrm{eval}} := t_n + \Delta_n^{\mathrm{eval}}$. Coordinates enter through $\psi(y_n)$, so output grids may differ from X in size and placement; target tokens do not attend to one another.

All attention modules concatenate heads of width d_h and project back to width d; each block then applies the pointwise pre-norm update $u \leftarrow u + \mathrm{FFN}(\mathrm{LN}(u))$, with $u = h_i^{\mathrm{inter}}(x_k)$ in the encoder and $u = t_n^{\mathrm{eval}}$ in the decoder.

4.5 PREDICTIVE DISTRIBUTION

Final normalization at y_n produces a decoder state $z_n \in \mathbb{R}^d$. Three heads map z_n to a mean μ_n, residual standard deviation $\sigma_n > 0$, and covariance feature $\ell_n \in \mathbb{R}^\rho$:

$$\mu_n = a^\top z_n + a_0, \qquad \sigma_n = \mathrm{softplus}(b^\top z_n + b_0) + \sigma_{\min}, \qquad \ell_n = W_L z_n/\sqrt{\rho}. \tag{10}$$

Stacking these outputs across the N target locations defines the joint predictive distribution

$$q_\theta\big(\tilde{u}_\star \mid \mathcal{C}\big) = \mathcal{N}(\mu, \Sigma), \qquad \Sigma = LL^\top + \mathrm{diag}(\sigma^2), \qquad L = (\ell_1, \ldots, \ell_N)^\top. \tag{11}$$

The covariance combines correlations $LL^\top$ induced across target locations by the unknown operator with location-specific residual variance $\mathrm{diag}(\sigma^2)$ that includes measurement noise (Fig. 6e), so $\Sigma_{nn} = \|\ell_n\|^2 + \sigma_n^2$; with $\rho \ll N$ storage falls from $O(N^2)$ to $O(N\rho)$, and $\sigma_{\min} > 0$ keeps Σ strictly positive definite.

4.6 TRAINING

Training draws episodes online from $\mathcal{G} \sim \Pi$ and uses each once (Fig. 6a), independently sampling operator structure and parameters, input functions, grid, quadrature weights, target locations, and noise level. The Gaussian in equation 11 is trained with

$$\mathcal{L} = \mathbb{E}_{\mathrm{episode}}\left[-\log q_\theta\big(\tilde{u}_\star \mid \mathcal{C}\big) + \frac{\lambda}{N}\|\mu - \tilde{u}_\star\|^2\right], \tag{12}$$

Joint likelihood supervises the mean and covariance; the auxiliary term further supervises the mean.

5 EXPERIMENTS

Four questions organize the experiments, each handing off to the next: ***(1)*** *is the inferred joint posterior faithful, and does pointwise calibration certify it?* ***(2)*** *can it make accurate predictions on unseen physical systems?* ***(3)*** *can posterior-guided measurement selection improve prediction with fewer observations?* ***(4)*** *where does reliability end under prior mismatch, and is that boundary detectable?*

5.1 JOINT POSTERIOR FIDELITY TO EXACT BAYES

We first evaluate ORIA trained on random operators drawn from a nine-mode linear-Gaussian prior, where its inferred joint predictive can be compared directly with the closed-form Bayesian posterior predictive (Appendix A.1).

For a fixed unseen operator and query, increasing m brings the predictive mean closer to the true response and contracts the uncertainty band (Fig. 1a). Across the test cohort, mean relative L_2 error continues to fall beyond the $m \leq 8$ training range, while joint KL against exact Bayes,

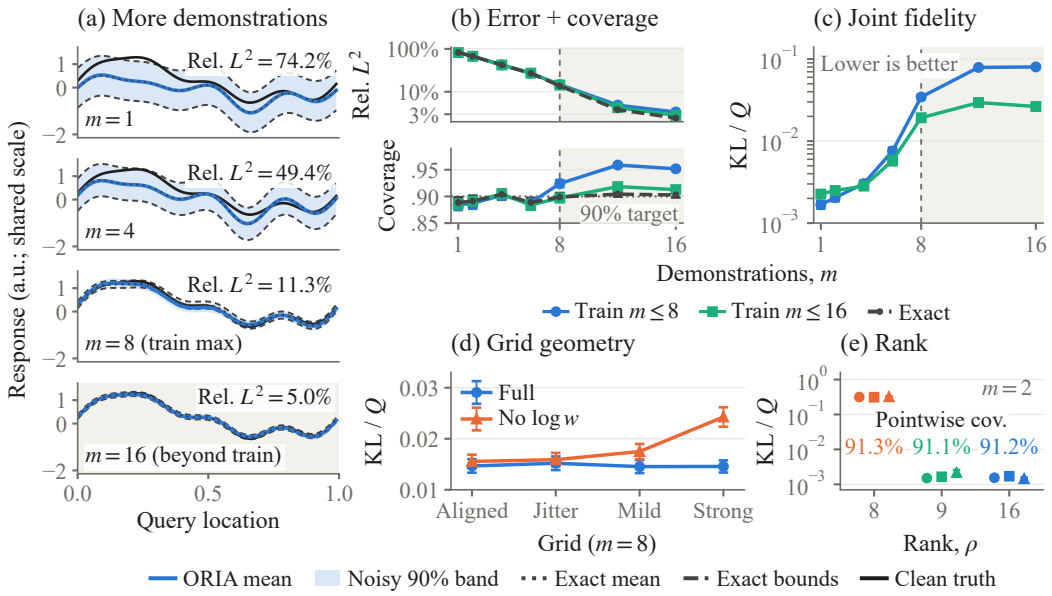

Figure 1: **Joint posterior fidelity to exact Bayes.** (a) ORIA and exact-Bayes predictions and uncertainty with increasing demonstrations. (b,c) Accuracy, calibration, and joint fidelity for ORIA trained with $m \le 8$ or $m \le 16$. (d) ORIA and its no-$\log w$ ablation across measurement grids. (e) Covariance capacity and the limits of pointwise calibration; points show three training seeds per rank, with coverage averaged over seeds.

$\mathrm{KL}(P^\star \| q_\theta)/Q$, rises from $m = 8$ to $m = 16$ but remains below 10^{-1} (Fig. 1b,c). Extending training to $m \le 16$ reduces the $m = 16$ error from 0.0337 to 0.0281, joint KL from 0.0802 to 0.0264 nats/query, and moving noisy pointwise 90% coverage from 0.952 to 0.913. Thus, gains in mean accuracy beyond the training range can coexist with insufficient posterior contraction, a gap reduced under broader-context training (Appendices A.2–A.3).

The joint fidelity extends across the grid discretizations and higher context resolutions. On Aligned, Jitter, Mild, and Strong grids (Appendix Fig. 5), ORIA's mean KL/Q stays near 0.015 and below the no-$\log w$ ablation, whose KL/Q rises from 0.0156 on Aligned to 0.0243 on Strong (Fig. 1d). Appendix A.4 further shows KL/Q of 0.0014–0.0018 on 1D grids with up to 192 context points (24–64 during training), versus 0.0137–0.0165 for a smaller canonical-grid baseline that resamples onto a fixed grid. Low joint KL also persists in 2D (Appendix A.4).These results together support the robustness of ORIA's continuum architecture across arbitrary nonuniform discretizations.

Beyond demonstrating robust discretizations, the joint perspective also reveals structural failures that marginal calibration alone can miss. We compare ORIA models with covariance-factor ranks $\rho \in \{16, 9, 8\}$, corresponding to excess, sufficient, and deliberately insufficient capacity, respectively, to represent the rank-nine correlation structure. At $m = 2$, three training seeds per rank yield KL/Q of 0.0014–0.0017 for $\rho = 16$ versus 0.312–0.319 for $\rho = 8$, while coverage remains near 0.91 for $\rho \in \{8, 9, 16\}$ (Fig. 1e; Appendix A.5). Coverage barely changes despite joint KL differing by roughly two orders of magnitude. This mismatch can materially affect uncertainty in aggregate responses on physical systems, as revealed in Section 5.2.

5.2 Few-shot prediction on unseen physical systems

To evaluate ORIA on practical physical systems, we first consider three FEM families: heterogeneous conduction, transient cooling, and an indefinite Helmholtz cavity. FEM baselines cover in-context probabilistic prediction (TNP-ND), per-episode neural-operator fitting (scratch FNO, FNO ensembles, and last-layer Bayes), pretrained operator adaptation (same-prior FNO with fine-tuning), and simple non-neural estimation (spectral fitting and context mean). We then evaluate ORIA on the public Navier–Stokes (NS) benchmark, where we additionally compare against a classical estimator

(affine) and a pretrained PDE foundation model (Poseidon-B, fine-tuned using the official recipe under observation parity). Implementation details for each baseline are given in Appendices B.1–B.2, B.5, and C.1.

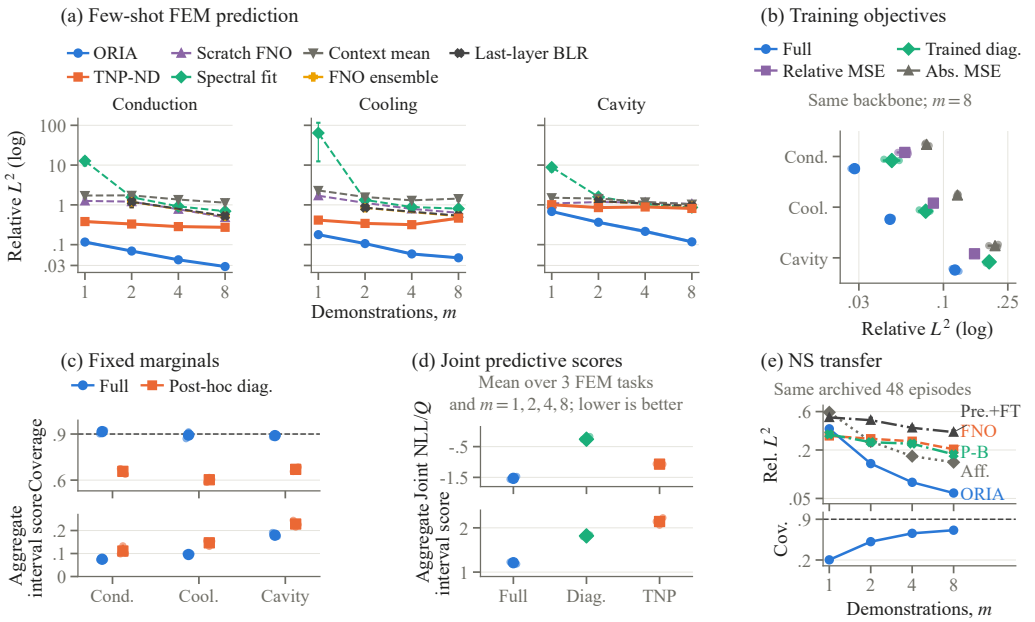

Figure 2: **Few-shot accuracy, joint uncertainty, and transfer.** (a) Prediction across ORIA and baselines with increasing demonstrations. (b) Effects of ORIA's training objective. (c) Uncertainty over spatial averages with full and post-hoc diagonalized ORIA covariance. (d) Joint and aggregate uncertainty for full ORIA, trained-diagonal ORIA, and TNP-ND. (e) ORIA and baseline predictions, and ORIA coverage, on public Navier–Stokes data. In (b–d), small points show three training seeds per setting; large markers and bars show their means and ranges.

ORIA's errors decrease with demonstrations (Fig. 2a). Although adapted TNP-ND has the lowest reported mean error among the displayed baselines at $m = 8$, its errors are still $9.69\times$, $9.95\times$, and $6.90\times$ ORIA's for conduction, cooling, and the cavity, respectively.

ORIA's joint Gaussian NLL+MSE objective (Full) contributes to this accuracy. With backbone, prior, and budget fixed, relative MSE improves on absolute MSE but still gives 1.32–$2.05\times$ Full's errors at $m = 8$. Replacing the full joint Gaussian NLL with a diagonal Gaussian NLL increases prediction errors by 60–70% relative to Full, suggesting that learning cross-location dependence during training benefits the predictive mean itself (Fig. 2b; Appendix B.3).

Extending the joint-structure audit of Section 5.1 to physical systems, we diagonalize the learned covariance. This leaves the predictive mean, marginal variances, and pointwise coverage unchanged, but lowers family-level coverage over six predefined spatial-average responses from 0.890–0.916 to 0.602–0.670 at $m = 8$ (Fig. 2c), demonstrating the importance of cross-location covariance for calibrated aggregate uncertainty. (Appendix B.3.)

Full-model comparisons reinforce this distinction. ORIA achieves lower average joint NLL and aggregate interval scores (penalizing wide intervals and missed coverage; Appendix B.3) than trained-diagonal ORIA and TNP-ND (Fig. 2d). Trained-diagonal ORIA undercovers the spatial averages, while TNP-ND attains similar average aggregate coverage to ORIA but produces wider intervals. ORIA therefore provides a better average balance of coverage and sharpness. (Appendix B.4.)

On public NS, a separately trained ORIA model retains the prediction advantage over all evaluated baselines from $m = 2$ onward. At $m = 8$, its relative error is 0.058, compared with 0.141 for the best baseline, affine estimation (Fig. 2e). However, its 90% pointwise intervals achieve only 71.2% coverage of noisy responses, so improved prediction accuracy coexists with overconfident uncertainty on this external benchmark (Appendix C.2). A discussion of ORIA's reliability will be given in Section 5.4.

5.3 POSTERIOR-GUIDED SCIENTIFIC ACTION

Beyond prediction, we test whether ORIA's posterior can guide measurement selection. Figure 3a shows the measurement loop: ORIA's predictive distributions guide the next measurement, and the resulting observation is added to the context to update those distributions.

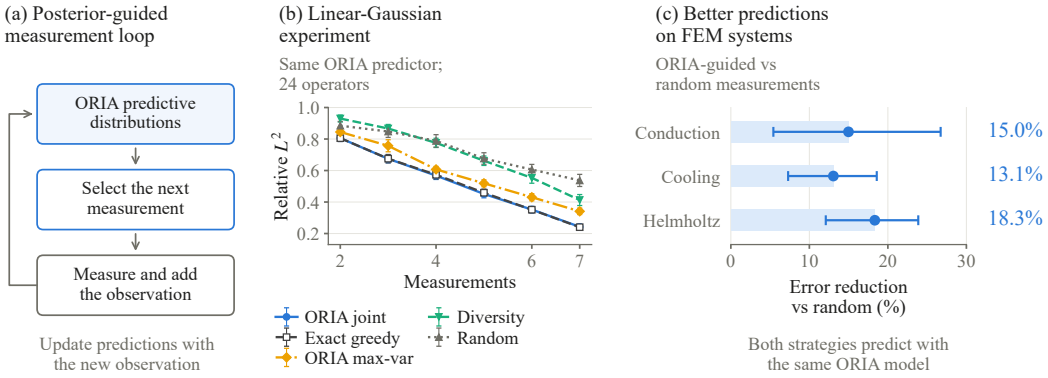

Figure 3: **Posterior-guided measurement selection.** (a) The measurement-selection loop. (b) Prediction accuracy under ORIA joint, Exact greedy, ORIA max-var, Diversity, and Random selection. (c) Prediction gains of ORIA max-var over Random on physical systems, averaging three fixed training seeds within each specimen; bars show paired-bootstrap 95% intervals.

In the linear-Gaussian setting introduced in Section 5.1, we compare five strategies for selecting the next measurement: ORIA joint and Exact greedy maximize expected one-step reduction in target uncertainty using learned and exact posteriors, respectively. ORIA max-var selects the most uncertain predicted response; Diversity favors inputs far from those already measured; Random samples uniformly.

All acquired data are evaluated with the same frozen ORIA predictor. The ORIA joint and Exact greedy error curves nearly overlap (Fig. 3b), showing that using the learned posterior to select measurements gives nearly the same prediction accuracy as using the exact posterior. With five measurements in total, ORIA joint already achieves lower mean error than random selection with seven (0.451 versus 0.537). ORIA max-var gives the next lowest mean error; Diversity improves on Random at larger budgets.

For the physical tasks, we use the simpler max-var rule. Across 64 new specimens per FEM task and three fixed training seeds, it reduces error-curve AUC over budgets 2–6 by 15.0%, 13.1%, and 18.3% versus random for conduction, cooling, and Helmholtz, respectively (Fig. 3c). Full protocols, comparisons, and intervals appear in Appendix D.

5.4 RELIABILITY BOUNDARIES UNDER PRIOR MISMATCH

To assess how far posterior reliability extends beyond the training prior and whether prior revision can improve transfer, we evaluate a frozen ORIA model pretrained on synthetic 1D operators generated by composing classical operators and pointwise nonlinearities (Appendix E.1). We use eight noisy demonstrations per episode and evaluate 128 episodes per condition (Fig. 4a). Nominal 90% intervals retain pointwise coverage of 0.884–0.922 under roughness ($\alpha = 1.0$–0.5) and scale ($\approx 3\times$) shifts, but undercover at 0.785–0.803 and 0.668–0.706 under noise ($\sigma_\varepsilon = 0.2$–0.4) and structural (unseen nonlinearity or Burgers dynamics) shifts, respectively. Against a matched control, an unseen nonlinearity increases error while predictive intervals narrow (Appendix E.2).

ORIA's joint predictions support diagnosing mismatch from observed demonstrations. We score leave-one-demonstration-out residuals across 48 locations using the full predictive covariance. Across three same-recipe checkpoints, AUROC (area under the receiver operating characteristic curve) against in-distribution episodes is 0.887–0.9999 for noise and structural shifts, versus 0.479–0.631 for roughness and scale (Fig. 4b). This diagnostic signal may flag potential failures of frozen inference and prompt consideration of more accurate alternatives.

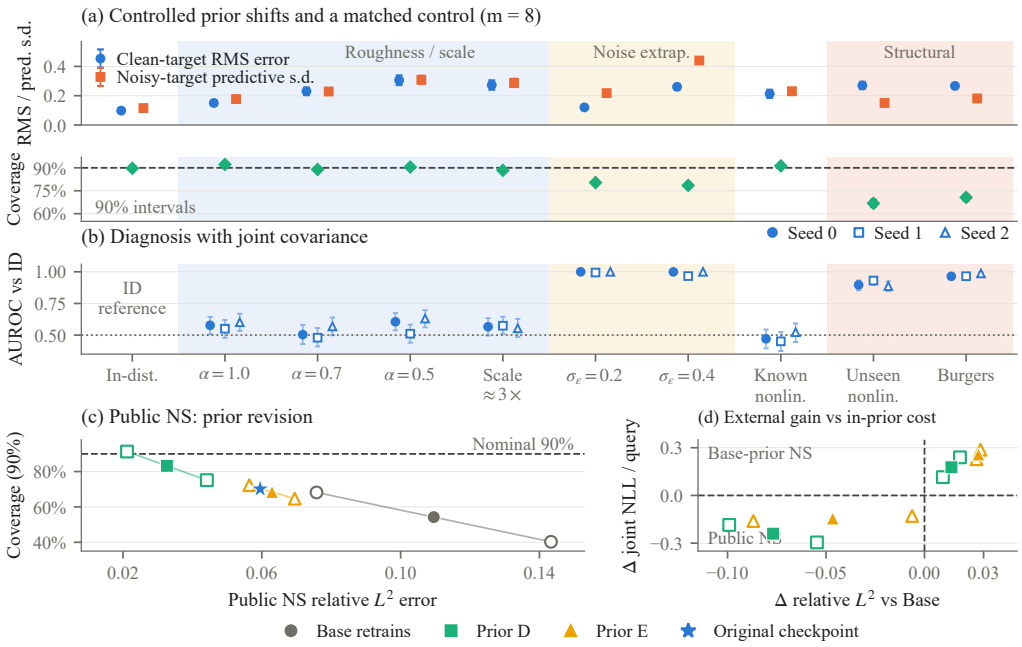

Figure 4: **Reliability under prior mismatch and revision.** (a) ORIA prediction and uncertainty under prior shifts. (b) Diagnosing mismatch from observed demonstrations, with one point per training seed and condition across three seeds. (c) External transfer for Base and revised priors D/E; the star marks the original checkpoint. (d) External versus original-prior tradeoffs for D/E relative to Base. In (c,d), open markers show two training seeds per prior and filled markers their means; differences in (d) are paired by seed.

We next test whether revising the prior improves transfer, using the separate 2D physics-mixture model at fixed architecture and training budget. Revision D broadens NS input and forcing laws; E additionally widens viscosity and forcing-amplitude support. Relative to Base, D reduces public-NS mean error from 0.110 to 0.033 and raises coverage from 0.54 to 0.83, with both seeds improving (Fig. 4c). Both revisions improve external error and joint NLL, but worsen both scores on fresh draws from the original NS prior (Fig. 4d). E yields smaller gains and larger costs than D. At this training budget, the D–E comparison shows that broader parameter support does not ensure better transfer (Appendix C.3).

6 Conclusion and Discussion

ORIA performs Bayesian operator inference from a few functional demonstrations with frozen weights across discretizations. Exact-Bayes comparisons and physical experiments show that pointwise calibration can conceal joint errors affecting aggregate uncertainty. Posterior-guided measurement selection improves prediction. A privileged benchmark using unobserved responses and target truth indicates further potential gains.

Reliability remains prior-dependent: structural novelty can increase error without wider intervals, diagnostics are weak for roughness and scale shifts, and prior revision trades external gains for in-prior performance. Gaussian predictives, finite covariance rank, and limited operator coverage constrain generality. Future work should develop more expressive predictives, preserve in-prior performance during revision, validate on laboratory data, and connect reliability diagnostics to measurement value.

AI USE STATEMENT

We used generative AI tools to aid and polish the writing of this manuscript and to help prepare and organize the appendix descriptions based on code written by the authors. The authors manually reviewed and verified all AI-assisted content for accuracy and take responsibility for the final content of this submission.

REFERENCES

Kamyar Azizzadenesheli, Nikola Kovachki, Zongyi Li, Miguel Liu-Schiaffini, Jean Kossaifi, and Anima Anandkumar. Neural operators for accelerating scientific simulations and design. *Nature Reviews Physics*, 6(5):320–328, 2024. doi: 10.1038/s42254-024-00712-5.

Francesca Bartolucci, Emmanuel de Bézenac, Bogdan Raonić, Roberto Molinaro, Siddhartha Mishra, and Rima Alaifari. Representation equivalent neural operators: a framework for alias-free operator learning. In *Advances in Neural Information Processing Systems*, 2023.

Wessel P Bruinsma, James Requeima, Andrew YK Foong, Jonathan Gordon, and Richard E Turner. The Gaussian neural process. *Advances in Approximate Bayesian Inference*, 2021.

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

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

Edoardo Calvello, Nikola B Kovachki, Matthew E Levine, and Andrew M Stuart. Continuum attention for neural operators. *Journal of Machine Learning Research*, 26(300):1–52, 2025.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In *International Conference on Machine Learning*, 2018.

Zhongkai Hao, Chang Su, Songming Liu, Julius Berner, Chengyang Ying, Hang Su, Anima Anandkumar, Jian Song, and Jun Zhu. DPOT: Auto-regressive denoising operator transformer for large-scale PDE pre-training. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 17616–17635. PMLR, 2024. URL https://proceedings.mlr.press/v235/hao24d.html.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025.

Marc C. Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001. doi: 10.1111/1467-9868.00294.

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.

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.

Yuxuan Liu, Jingmin Sun, Xinjie He, Griffin Pinney, Zecheng Zhang, and Hayden Schaeffer. PROSE-FD: A multimodal PDE foundation model for learning multiple operators for forecasting fluid dynamics. *arXiv preprint arXiv:2409.09811*, 2024. URL https://arxiv.org/abs/2409.09811.

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. Linearization turns neural operators into function-valued Gaussian processes. In *International Conference on Machine Learning*, 2025.

Stratis Markou, James Requeima, Wessel P Bruinsma, Anna Vaughan, and Richard E Turner. Practical conditional neural processes via tractable dependent predictions. In *International Conference on Learning Representations*, 2022.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do Bayesian inference. In *International Conference on Learning Representations*, 2022.

Tung Nguyen and Aditya Grover. Transformer neural processes: Uncertainty-aware meta learning via sequence modeling. In *International Conference on Machine Learning*, 2022.

Maziar Raissi and George Em Karniadakis. Hidden physics models: Machine learning of nonlinear partial differential equations. *Journal of Computational Physics*, 357:125–141, 2018. doi: 10.1016/j.jcp.2017.11.039.

Louis Serrano, Armand Kassaï Koupaï, Thomas X Wang, Pierre Erbacher, and Patrick Gallinari. Zebra: In-context generative pretraining for solving parametric PDEs. In *International Conference on Machine Learning*, 2025.

Andrew M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010. doi: 10.1017/S0962492910000061.

Jingmin Sun, Yuxuan Liu, Zecheng Zhang, and Hayden Schaeffer. Towards a foundation model for partial differential equations: Multioperator learning and extrapolation. *Physical Review E*, 111: 035304, 2025. doi: 10.1103/PhysRevE.111.035304.

Liu Yang, Xuhui Meng, and George Em Karniadakis. B-PINNs: Bayesian physics-informed neural networks for forward and inverse PDE problems with noisy data. *arXiv preprint arXiv:2003.06097*, 2020. URL `https://arxiv.org/abs/2003.06097`.

Liu Yang, Siting Liu, Tingwei Meng, and Stanley J Osher. In-context operator learning with data prompts for differential equation problems. *Proceedings of the National Academy of Sciences*, 120(39):e2310142120, 2023.

Benjamin J Zhang, Siting Liu, Stanley J Osher, and Markos A Katsoulakis. Probabilistic operator learning: generative modeling and uncertainty quantification for foundation models of differential equations. *arXiv preprint arXiv:2509.05186*, 2025.

A JOINT POSTERIOR FIDELITY TO EXACT BAYES

This appendix gives the protocols and numerical evidence for Section 5.1. We use the main-text notation: $\mathcal{G}$ is the unknown operator, m is the number of demonstrations, X contains M measurement locations, and Y contains N prediction locations. The experimental figures write the same prediction count as Q; throughout this appendix, KL per prediction location is $\mathrm{KL}(P^\star \| q_\theta)/N$.

A.1 LINEAR-GAUSSIAN PRIOR, EXACT REFERENCE, AND TRAINING

The companion model uses a separate linear-Gaussian operator prior on $\Omega = [0, 1)$. Its nine modes are the constant function and the first four sine and cosine pairs, normalized to unit L^2 norm. The operator maps the input coefficients in this basis linearly to output coefficients in the same basis. Its 81 coefficients are independent zero-mean Gaussians. Their variances are proportional to the product of the input- and output-mode frequency weights: for either mode, the weight is one plus its frequency, raised to $-3/2$. The variances are normalized to sum to nine. Input functions have independent zero-mean Gaussian coefficients with standard deviations proportional to the reciprocal of one plus the mode frequency, normalized so that the expected squared L^2 norm is one.

Both demonstration and query responses receive independent Gaussian noise with the fixed standard deviation $\sigma_\varepsilon = 0.05$. Consequently, $P^\star$ is obtained by Gaussian conditioning on the observed demonstrations and then adding the variance of a future noisy measurement. This experiment fixes the noise level rather than inferring an unknown episode-specific variance. The exact covariance of the clean response has rank at most nine; the noisy predictive has a nonsingular covariance and is the reference for joint KL and coverage. The bands in Figure 1a are pointwise intervals for noisy measurements, displayed alongside the clean response; they are not simultaneous bands for the full field.

During training, m is uniform on $\{1, \ldots, 8\}$ and M on $\{24, \ldots, 64\}$, with $N = 48$. With probability 0.7, the measurement grid contains one point in each of M equal cells, sampled uniformly between 10% and 90% of the cell width; otherwise, the grid contains M sorted independent uniform points. Prediction locations are sorted independent uniform points. All demonstrations and the query input share X within an episode. Periodic trapezoidal weights are computed from these measurement locations.

The model has width 256, six encoder blocks, two decoder blocks, four attention heads, and covariance-factor rank $\rho = 16$. Training uses 100,000 updates and batch size 64. The broader-context comparison changes only the training range to $m \in \{1, \ldots, 16\}$. The no-$\log w$ comparison retains $m \leq 8$ and removes only the additive spatial and decoder attention bias; the inter-function attention still integrates using the actual quadrature weights. Each comparison uses three independent training seeds and the final checkpoint, without test-based selection. Training with $m \leq 16$ uses more demonstrations and compute at the same update budget; the comparison does not match those resources. The asynchronous training streams are not exactly paired.

The single example in Figure 1a holds the operator, query, and noisy demonstration pool fixed at $m = 1, 4, 8, 16$, with $M = 48$ and $N = 96$. The displayed errors are for this example, not the cohort means. The Gaussian prior permits unbounded responses; the common plotted range is only a display choice.

A.2 METRIC CONVENTIONS AND THE CONTEXT-SIZE COMPARISON

Relative error is the unweighted discrete L^2 distance from μ to the clean response u, divided by the response norm with a 10^{-9} floor. Joint KL is directed from the exact noisy predictive $P^\star$ to q_θ and uses natural logarithms. Dividing the total joint KL by N gives nats per prediction location; this normalization does not bound full-field discrepancy independently of the number of prediction locations.

Noisy pointwise coverage uses the intervals $\mu_n \pm 1.64485\sqrt{\Sigma_{nn}}$. It averages the coverage indicators over prediction locations and then over episodes. Mean interval width uses the same order of averaging. The nominal coverage is 90%. A numerical variance of 10^{-8} is added to the diagonals of both covariance matrices only when computing KL; coverage and widths use the original covariances.

The context-size comparison uses 192 independent specimens, $M = 48$, $N = 24$, and the mixed measurement-grid distribution described above. Each specimen retains its operator, query, query noise, and pool of 16 demonstrations. The predictor and exact reference receive only the first $m \in \{1, 2, 4, 6, 8, 12, 16\}$ demonstrations. All training seeds use the same evaluation states. Metrics are averaged across the three fixed seeds within each specimen before calculating means and specimen standard errors. Repeated context sizes, locations, and seeds do not increase the independent specimen count, and these standard errors do not quantify uncertainty over new training runs.

The exact reference reconstructs input-function coefficients by least squares from the same float32-visible samples supplied to ORIA and conditions on the visible noisy responses. Reference calculations use float64 and a continuous Gaussian likelihood at the recorded values. On this bank, the maximum relative differences from a reference using the generator's input coefficients are 6.51×10^{-7} in the mean and 3.66×10^{-7} in the clean-response covariance. These checks support negligible numerical effects on this bank, not equivalence between sampled and continuously observed inputs in general.

Table 1: Context-size comparison on 192 paired specimens with three fixed training seeds per model. Joint KL per location is in nats.

Training range	m	Rel. L^2	KL/location
$m \leq 8$	1	0.8228	0.00167
	2	0.6706	0.00205
	4	0.4226	0.00304
	6	0.2717	0.00757
	8	0.1475	0.03447
	12	0.0488	0.07908
	16	0.0337	0.08023
$m \leq 16$	1	0.8223	0.00226
	2	0.6707	0.00250
	4	0.4215	0.00282
	6	0.2709	0.00577
	8	0.1433	0.01926
	12	0.0427	0.02948
	16	0.0281	0.02640

For the original training range, clean-response error improves beyond $m = 8$, while joint KL increases. Broader-context training reduces joint KL at $m = 16$, but does not eliminate posterior error or improve every context size: at $m = 1$, KL per location rises from 0.00167 to 0.00226.

A.3 Noisy predictive intervals and the residual variance

The interval comparison below uses the total covariance $\Sigma = LL^\top + \mathrm{diag}(\sigma^2)$ of the noisy predictive in the main text. At $m = 16$, broader-context training changes coverage from 0.9518 to 0.9128 and width from 0.2516 to 0.2013, compared with the exact width 0.1859. Thus the improved KL accompanies reduced excess width and coverage closer to the nominal level.

Table 2: Noisy pointwise 90% intervals on the same 192 specimens. Each entry is coverage / width; widths are in response units. These are marginal intervals, not simultaneous field bands.

m	Train $m \leq 8$	Train $m \leq 16$	Exact Bayes
1	0.8843 / 3.1313	0.8864 / 3.1403	0.8885 / 3.1875
2	0.8870 / 2.4795	0.8909 / 2.4972	0.8913 / 2.5172
4	0.9022 / 1.5042	0.9043 / 1.5158	0.9039 / 1.5053
6	0.8880 / 0.9268	0.8840 / 0.9158	0.8887 / 0.9212
8	0.9240 / 0.5730	0.8971 / 0.5090	0.8989 / 0.4898
12	0.9588 / 0.3063	0.9184 / 0.2357	0.9043 / 0.2126
16	0.9518 / 0.2516	0.9128 / 0.2013	0.9028 / 0.1859

The implemented residual standard deviation uses a softplus followed by the floor $\sigma_{\min} = 0.001$. All 96,768 recorded residual standard deviations for the original model on this cohort lie between 0.047694 and 0.052531, well above that floor and close to the fixed noise level 0.05. These observations do not support floor activation as the cause of the remaining contraction error. Because

the noise level is constant throughout this prior, the result does not establish recovery of unknown episode-specific noise or identify a unique separation of the two covariance contributions.

A.4 Attention bias, grid geometry, and resolution transfer

The attention-bias comparison in Figure 1d uses 128 independent specimens with $m = 2, 8$ and $M = N = 48$. The prediction locations are fixed uniform midpoints. Aligned uses these same locations for measurement. Jitter uses the within-cell sampling rule above. Mild and Strong displace each uniform midpoint by the sine of 2π times that midpoint, divided by 2π, and multiplied by 0.5 or 0.9, respectively. These two fixed transformations increasingly cluster measurement locations near the center while preserving their order.

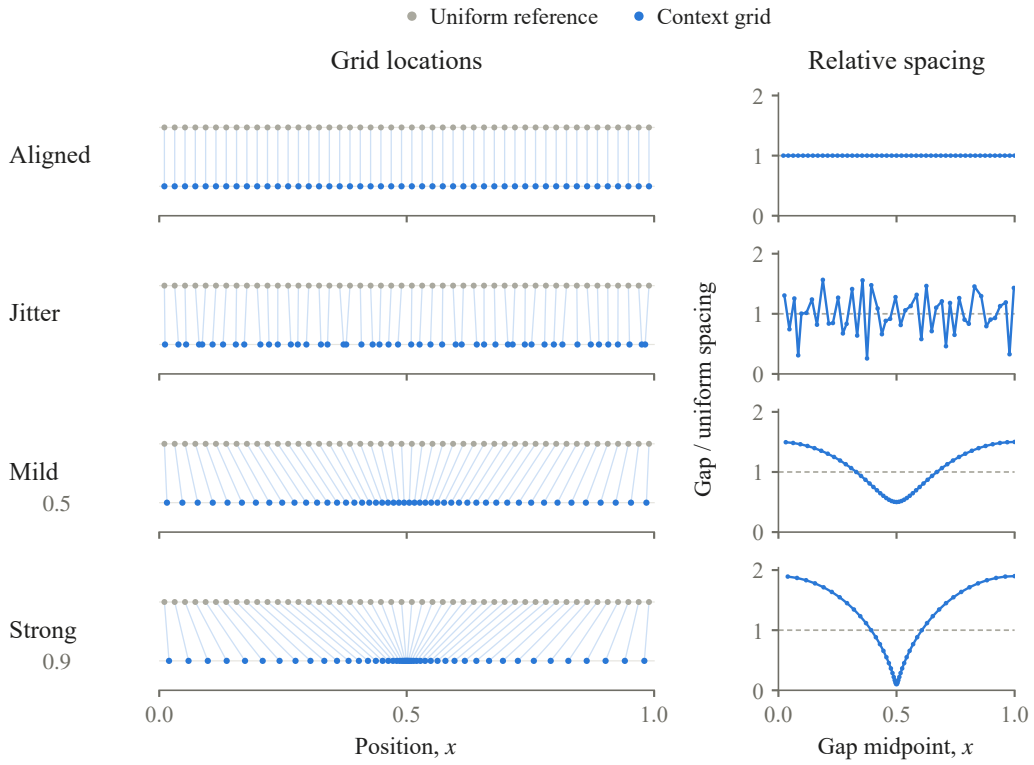

Figure 5: **The four measurement-grid geometries used in Figure 1d.** Each row has $M = 48$ measurement locations, with the same $N = 48$ uniform prediction locations. Left: gray dots are uniform midpoints and colored dots are the actual measurement locations of specimen 0 at $m = 8$; lines connect corresponding points. Right: adjacent-point spacing divided by the uniform spacing $1/48$, including the periodic boundary gap. A value of one denotes uniform spacing.

The same operators, continuous input functions, and standard-normal noise draws are reused across geometries, but the measurement locations change. The exact reference is recomputed for each geometry's actual observations using the same input-reconstruction convention as above. Full ORIA and no-$\log w$ have the same architecture, prior, update budget, and training range. The comparison averages three fixed training seeds within each specimen. Repeated geometries and seeds do not increase the independent specimen count.

At $m = 8$, the no-$\log w$ minus full ORIA KL difference on Strong grids is 0.00970 ± 0.00105 nats per prediction location (specimen standard error). Mean point error differs less, and no-$\log w$ has slightly lower point error on Aligned and Jitter grids. The evidence therefore concerns joint fidelity on the specified nonuniform grids; it does not show improvement in every metric on every grid.

Table 3: Attention-bias comparison on 128 paired specimens. Each entry is full ORIA / no-$\log w$. Coverage and width concern noisy pointwise 90% intervals. Both models retain quadrature weights in inter-function attention.

Grid	m	Rel. L^2	KL/location	Coverage	Width
Aligned	2	$0.6108/0.6110$	$0.00125/0.00137$	$0.8939/0.8990$	$2.3613/2.3862$
Aligned	8	$0.1496/0.1481$	$0.01466/0.01555$	$0.9196/0.9244$	$0.5723/0.5843$
Jitter	2	$0.6107/0.6110$	$0.00127/0.00139$	$0.8938/0.8993$	$2.3616/2.3861$
Jitter	8	$0.1504/0.1484$	$0.01520/0.01589$	$0.9187/0.9244$	$0.5726/0.5847$
Mild	2	$0.6110/0.6131$	$0.00126/0.00223$	$0.8925/0.8937$	$2.3610/2.3713$
Mild	8	$0.1517/0.1509$	$0.01454/0.01749$	$0.9183/0.9223$	$0.5726/0.5876$
Strong	2	$0.6108/0.6176$	$0.00128/0.00476$	$0.8915/0.8959$	$2.3612/2.3958$
Strong	8	$0.1551/0.1576$	$0.01457/0.02427$	$0.9119/0.9255$	$0.5724/0.6237$

One-dimensional resolution transfer. The resolution comparison fixes $m = 4$ for 96 continuum episodes. Measurement grids with $M = 24, 48, 96$ are stride subsets of one jittered 192-point grid, and all resolutions share $N = 48$ uniform prediction locations. One noise realization on the finest measurement grid is inherited by every subset. At each resolution, the exact reference conditions on that resolution's measurements. Additional points therefore change both discretization and available information.

Table 4: One-dimensional resolution transfer: KL per prediction location (mean $\pm$ episode standard error), with 96 episodes, $m = 4$, and $N = 48$. ORIA was trained with $24 \le M \le 64$.

M	ORIA	Canonical-grid baseline
24	0.001793 ± 0.000193	0.013727 ± 0.001679
48	0.001404 ± 0.000153	0.014974 ± 0.001703
96	0.001357 ± 0.000149	0.016038 ± 0.001747
192	0.001357 ± 0.000149	0.016477 ± 0.001823

The canonical-grid baseline resamples observations onto 48 fixed points and has 2.205 million parameters, compared with ORIA's 7.916 million. This comparison does not match capacity or isolate the effect of quadrature. Its KL differences may reflect interpolation and noise processing as well as architecture and capacity.

Two-dimensional resolution transfer. The 2D linear-Gaussian companion uses a 5×5 tensor-product trigonometric basis, 625 Gaussian operator coefficients, and a rank-32 head. Training measurement grids contain 12^2–20^2 points. Each tested resolution uses 48 episodes with $m = 4$ and $N = 48$, with an exact predictive recomputed for that resolution. Episode sets differ across resolutions because sampling consumes different numbers of random draws, despite reuse of the random seed. The results compare average fidelity across resolutions rather than paired refinement of the same episodes. No 2D canonical-grid comparison is included.

Table 5: Two-dimensional resolution transfer: KL per prediction location (mean $\pm$ episode standard error), with 48 episodes per resolution. Episode sets are not paired across resolutions.

M	KL/location
8^2	0.08295 ± 0.01946
12^2	0.01517 ± 0.00308
16^2	0.00652 ± 0.00158
20^2	0.00461 ± 0.00050
24^2	0.00481 ± 0.00124

Joint KL remains low at 24^2 points, beyond the training range, whereas the 8^2 case is less accurate. The geometry, 1D-resolution, and 2D-resolution comparisons use different cohorts and checkpoints; their results are not pooled.

A.5 THREE TRAINING SEEDS PER COVARIANCE-FACTOR RANK

The rank comparison trains three independent models at each $\rho \in \{8, 9, 16\}$, changing only the covariance-factor rank. All nine models use the same stream of 256 episodes at each of $m = 2, 8$, with $M = 48$, $N = 24$, and jittered measurement grids. Figure 1e plots the three seed means at $m = 2$, each with its own episode standard error; these bars do not represent variation over training seeds. Noisy pointwise coverage in that panel averages the three archived seed means within each rank. Per-episode coverage was not saved, so the displayed coverage uses the archived summaries.

Table 6: Rank comparison on 256 shared episodes per context size. KL per prediction location is in nats, with episode standard errors. Coverage concerns noisy pointwise 90% intervals.

ρ	Seed	KL/location, $m = 2$	Coverage	KL/location, $m = 8$	Coverage
8	0	0.31571 ± 0.01067	0.911	0.10384 ± 0.00443	0.939
8	1	0.31201 ± 0.01104	0.911	0.09881 ± 0.00433	0.926
8	2	0.31922 ± 0.01059	0.916	0.09676 ± 0.00424	0.925
9	0	0.00151 ± 0.00011	0.912	0.02116 ± 0.00133	0.924
9	1	0.00165 ± 0.00015	0.911	0.02378 ± 0.00144	0.924
9	2	0.00217 ± 0.00039	0.911	0.03698 ± 0.00240	0.929
16	0	0.00155 ± 0.00018	0.912	0.02291 ± 0.00145	0.924
16	1	0.00171 ± 0.00014	0.909	0.02578 ± 0.00157	0.928
16	2	0.00144 ± 0.00013	0.914	0.02198 ± 0.00136	0.929

At $m = 2$, every rank-eight model has much higher joint KL, although coverage over all nine models remains between 0.909 and 0.916. Rank nine suffices to express the exact clean-response covariance with the correct diagonal noise. This does not prove that a rank-eight factor with an unrestricted diagonal cannot represent the total covariance. The comparison demonstrates a learned joint-predictive failure that pointwise coverage misses, rather than a general representation impossibility. The physical experiments in Section 5.2 separately assess the consequences for aggregate-response uncertainty.

B PHYSICAL-SYSTEM PREDICTION: PROTOCOLS AND SUPPORTING RESULTS

This appendix supports the FEM comparisons in Section 5.2 and Figure 2a–d. The three families are heterogeneous conduction, transient cooling, and an indefinite Helmholtz cavity on $\Omega = [0, 1]^2$. Each specimen fixes an operator $\mathcal{G}$; its demonstrations and query differ in the input function f. Training resamples $m \in \{1, \dots, 8\}$ and 12–20 observation locations per axis. Evaluation uses $M = 14^2$ observations in X, $N = 48$ prediction locations in Y, and $m \in \{1, 2, 4, 8\}$. The demonstration functions and query input share X; Y is sampled separately and uniformly. Prediction error is relative L^2 error against the clean response, whereas coverage and density scores use noisy query responses.

Responses are divided by the standard deviation of the clean demonstration responses before independent Gaussian noise is added. The episode noise standard deviation σ_ε is drawn log-uniformly between 0.01 and 0.08 in these normalized units. This normalization uses simulator information and is part of the recorded protocol; it is not an additional observable supplied to the predictor. The archived pretrained-FNO control uses a different normalization during pretraining, as specified below.

B.1 SHARED-BANK BASELINES AND TRAINING BUDGETS

The controlled comparison contains Full ORIA, trained-diagonal ORIA, relative-MSE ORIA, absolute-MSE ORIA, and adapted TNP-ND, each trained with seeds 0, 1, and 2. All fifteen runs use their completed 100,000-step checkpoints. They share a fixed test bank of 768 episodes: 64 fresh specimen/excitation episodes in each of twelve combinations of family and m. Reusing this bank across training seeds does not increase the number of independent test specimens. Figure 2a (solid curves), b–d shows three-seed means and variation; the additional baseline curves have separate archived cohorts and episode SEMs.

All five models receive the sampled query input, every demonstrated input–response pair, coordinates, and observation weights. They use the same prior and 2,500-specimen training bank, with a separate 500-specimen validation shard, noisy targets, and batch size 24. Each final run therefore has 2.4 million episode presentations, with specimens reused. The models share the learning rate and training schedule. ORIA and TNP-ND remain frozen during evaluation.

The four ORIA objectives use the same architecture: width 256, six encoder blocks, two decoder blocks, four attention heads, and covariance rank $\rho = 32$, giving 7,936,802 allocated parameters. The point-only objectives leave the covariance heads untrained; trained-diagonal ORIA sets the covariance factor to zero and freezes that head. Their backbones are matched, but the number of parameters receiving gradients differs. The selected TNP-ND has 6,470,946 parameters; neither equal wall-clock time nor equal memory use is implied.

Adapted TNP-ND. The baseline adapts TNP-ND to functional demonstrations. It receives all sampled values of each input function, their coordinates and observation weights, and the demonstrated responses. Its learned function representation is finite-dimensional and need not preserve all information in the sampled function. Its predictive distribution is a joint Gaussian with low-rank-plus-diagonal covariance, scored directly. This is an adapted operator baseline, not the unmodified point-observation benchmark.

Four pilot configurations cross width 128 or 256 with Gaussian NLL plus MSE weight 0 or 1. Each receives 10,000 steps. Selection minimizes mean joint NLL per prediction location on a disjoint 384-episode validation bank, equally weighting its twelve family/context-size combinations. The selected configuration has width 256, unit MSE weight, six transformer layers, four heads, and rank 32. All three final runs start afresh and are evaluated at 100,000 steps. The pilot search adds 40,000 steps beyond these final budgets; it does not use the test bank or establish exhaustive optimization of TNP-ND.

Table 7 retains the seed-specific prediction errors. At $m = 8$, Full's three-seed means are $0.0282/0.0467/0.1177$ for conduction/cooling/cavity, compared with $0.2734/0.4647/0.8127$ for adapted TNP-ND. The corresponding ratios are $9.69/9.95/6.90$. TNP-ND's cooling error rises from 0.3181 at $m = 4$ to 0.4647 at $m = 8$, so improvement with more demonstrations is not uniform across methods.

Table 7: Completed shared-bank clean-target relative L^2. Each row uses 64 common episodes; seeds reuse the same bank. All five models train for 100,000 steps.

Family	m	Seed	Full	Trained diag.	Relative MSE	Absolute MSE	TNP-ND
Conduction	1	0	0.1249	0.1762	0.2153	0.2370	0.3794
Conduction	1	1	0.1170	0.1927	0.2035	0.2184	0.3646
Conduction	1	2	0.1073	0.1724	0.2104	0.2436	0.3989
Conduction	2	0	0.0677	0.1005	0.1290	0.1700	0.3316
Conduction	2	1	0.0709	0.1194	0.1162	0.1734	0.3149
Conduction	2	2	0.0688	0.1071	0.1208	0.1784	0.3442
Conduction	4	0	0.0402	0.0671	0.0817	0.1151	0.2892
Conduction	4	1	0.0430	0.0755	0.0789	0.1099	0.2739
Conduction	4	2	0.0416	0.0665	0.0818	0.1102	0.2903
Conduction	8	0	0.0289	0.0477	0.0606	0.0816	0.2791
Conduction	8	1	0.0268	0.0533	0.0523	0.0767	0.2628
Conduction	8	2	0.0288	0.0429	0.0602	0.0776	0.2781
Cooling	1	0	0.1809	0.2570	0.2383	0.3117	0.4055
Cooling	1	1	0.1828	0.2626	0.2518	0.2826	0.4179
Cooling	1	2	0.1720	0.2457	0.2581	0.2904	0.4205
Cooling	2	0	0.1044	0.1457	0.1776	0.1973	0.3569
Cooling	2	1	0.1084	0.1585	0.1659	0.1986	0.3267
Cooling	2	2	0.1065	0.1535	0.1749	0.1988	0.3425
Cooling	4	0	0.0548	0.0908	0.1104	0.1448	0.3231
Cooling	4	1	0.0608	0.0975	0.0998	0.1393	0.3169
Cooling	4	2	0.0598	0.0900	0.1057	0.1462	0.3144
Cooling	8	0	0.0459	0.0795	0.0860	0.1204	0.4836
Cooling	8	1	0.0460	0.0826	0.0888	0.1235	0.4306
Cooling	8	2	0.0483	0.0708	0.0851	0.1224	0.4799
Cavity	1	0	0.6919	0.8801	0.6980	0.9254	1.0352
Cavity	1	1	0.6745	0.8875	0.7330	0.8932	0.9990
Cavity	1	2	0.6789	0.9765	0.7070	1.0192	0.9941
Cavity	2	0	0.3796	0.5356	0.4218	0.5402	0.8692
Cavity	2	1	0.3743	0.5155	0.4179	0.5500	0.8506
Cavity	2	2	0.3420	0.5535	0.4006	0.5237	0.8438
Cavity	4	0	0.2293	0.3653	0.2757	0.3887	0.8887
Cavity	4	1	0.2188	0.3852	0.2800	0.3919	0.8719
Cavity	4	2	0.1990	0.3777	0.2932	0.3708	0.8979
Cavity	8	0	0.1247	0.1835	0.1518	0.2163	0.8158
Cavity	8	1	0.1154	0.1973	0.1560	0.1907	0.8221
Cavity	8	2	0.1130	0.1946	0.1591	0.2182	0.8002

B.2 Archived point-prediction baselines

The scratch FNO, spectral fit, and context mean in Table 8 share noisy episodes within that table. Its 192 episodes per m are randomly divided among the three families. These episodes and its original ORIA checkpoint differ from the fixed-bank three-seed comparison above. Shared axes in Figure 2a do not make these cohorts paired.

Spectral fitting and context mean. The spectral fit projects inputs and responses onto 12×12 sine modes and fits one multiplier per mode using the noisy demonstrations, without an intercept or tuned regularization. The context mean interpolates the average demonstrated response and ignores the query input. Neither estimator is identified with the Bayesian posterior predictive under ORIA's training prior.

Scratch FNO. Eight configurations vary width, modes, layers, padding, optimization steps, and learning rate. Selection uses six fresh episodes per family/context-size combination at $m = 2, 8$, two initializations, and common validation episodes across candidates, disjoint from test evaluation. The criterion equally weights the six cell-mean errors. The selected model has width 64, 16 modes, four layers, padding 16, and 2,000 training steps at learning rate 10^{-3}. Test errors score the mean prediction of three initializations. Inputs are interpolated onto a 48×48 full-domain grid. Training loss is computed at the actual observed response coordinates; no boundary response labels are added.

Table 8: Archived FEM relative L^2 errors against the clean query response, mean $\pm$ episode SEM. All methods in this table use identical episodes: 192 per m, randomly divided among families. The original ORIA checkpoint and this cohort are distinct from the shared-bank three-seed comparison.

Family	m	ORIA	Scratch FNO	Spectral fit	Context mean	Episodes
Conduction	1	0.095 ± 0.011	1.258 ± 0.130	12.723 ± 2.932	1.716 ± 0.212	64
	2	0.069 ± 0.006	1.206 ± 0.157	1.556 ± 0.228	1.735 ± 0.197	73
	4	0.040 ± 0.004	0.776 ± 0.071	0.916 ± 0.101	1.362 ± 0.122	63
	8	0.025 ± 0.002	0.483 ± 0.028	0.693 ± 0.062	1.135 ± 0.071	69
Cooling	1	0.239 ± 0.038	1.699 ± 0.286	64.102 ± 51.682	2.301 ± 0.427	60
	2	0.099 ± 0.011	1.122 ± 0.085	1.334 ± 0.147	1.579 ± 0.140	67
	4	0.051 ± 0.005	0.792 ± 0.060	0.881 ± 0.094	1.302 ± 0.103	56
	8	0.047 ± 0.007	0.632 ± 0.061	0.806 ± 0.092	1.436 ± 0.153	58
Cavity	1	0.684 ± 0.052	1.085 ± 0.029	8.811 ± 1.854	1.512 ± 0.091	68
	2	0.460 ± 0.037	1.179 ± 0.072	1.596 ± 0.158	1.442 ± 0.101	52
	4	0.229 ± 0.021	1.147 ± 0.080	1.037 ± 0.070	1.185 ± 0.074	73
	8	0.119 ± 0.012	1.053 ± 0.053	0.864 ± 0.052	1.043 ± 0.044	65

Pretrained FNO with fine-tuning. The pretrained FNO uses the same specimen prior and full-domain input encoding, with 30,000 pretraining steps followed by 800 updates at learning rate 10^{-4} per episode. This adaptation recipe has no independent sweep or early stopping. Pretraining uses clean targets and divides each response by its own standard deviation, unlike ORIA's noisy episode task and 100,000-step budget. Thus the prior exposure is shared, but the training targets, normalization, and budgets are not matched. Table 9 reports the adapted predictions. Its nominal allocation is 24 episodes per family, with the actual random family counts shown. The shorter seeded stream shares only a prefix with the larger archived cohort, preventing a fully paired comparison of their aggregate results. These values evaluate the recorded recipe and do not isolate the cause of its larger errors.

Table 9: Pretrained FNO with fine-tuning: clean-target relative L^2 error, mean $\pm$ episode SEM. The actual family counts belong to a shorter episode stream than Table 8; the two tables do not form a fully paired comparison.

Family	m	Pretrained + FT	Episodes
Conduction	1	0.725 ± 0.046	27
	2	1.381 ± 0.310	27
	4	1.631 ± 0.292	25
	8	1.291 ± 0.259	27
Cooling	1	0.914 ± 0.085	22
	2	0.831 ± 0.088	30
	4	1.338 ± 0.237	25
	8	1.486 ± 0.314	24
Cavity	1	2.165 ± 0.321	23
	2	1.736 ± 0.237	15
	4	1.994 ± 0.197	22
	8	1.904 ± 0.179	21

B.3 Training objectives and covariance diagonalization

The four objectives in Figure 2b use the common training and evaluation protocol above. Absolute MSE averages squared errors over the noisy query target. Relative MSE divides that episode error by the episode's mean squared target value plus 10^{-6} before averaging. Full and trained-diagonal ORIA add unit-weight MSE to Gaussian NLL per prediction location. Full uses $\Sigma = \mathrm{diag}(\sigma^2) + LL^\top$; trained-diagonal ORIA uses $\mathrm{diag}(\sigma^2)$ and reoptimizes the predictive mean and variance. Both losses add 10^{-6} to each covariance diagonal entry for numerical evaluation. The MSE-only models have untrained covariance heads and receive no probability scores.

Across the twelve equally weighted family/context-size combinations and three seeds, mean relative error is 0.1688 for Full, 0.2502 for trained-diagonal, 0.2208 for relative MSE, and 0.2860 for absolute MSE. At $m = 8$, relative-MSE/Full error ratios are $2.05/1.85/1.32$, and trained-diagonal/Full

ratios are $1.70/1.66/1.63$. These comparisons support a benefit of the full joint training objective in these runs; they do not separately isolate every effect of the architecture, optimization, or loss weighting.

Six spatial averages. The six responses in Figure 2c are the global mean, four quadrant means, and a mean over points within distance 0.15 of the boundary. Each uses equal weights over the prediction locations inside its region; at least three such locations are required for scoring. Predictive means and variances are obtained by averaging the corresponding entries of μ and Σ, respectively. These are finite-sample averages of noisy responses at Y.

Coverage records whether each observed average lies in its nominal 90% predictive interval. Reported coverage averages the six indicators within each episode, then averages episodes within each physical family and training seeds. It is not simultaneous coverage of all six averages or the entire field. Interval score (IS) equals interval width plus twenty times the distance of an observation outside the interval to the nearest endpoint, with zero additional penalty for an observation inside. Scores and widths use the normalized response units; lower IS gives a better balance between width and missed coverage.

Post-hoc diagonalization deletes off-diagonal entries of Σ after training. It preserves μ, marginal variances, and pointwise coverage, and is distinct from training with a diagonal covariance. Table 10 gives the $m = 8$ values underlying Figure 2c. Full covariance improves both mean aggregate coverage and mean interval score in all three families.

Table 10: Post-hoc diagonalization at $m = 8$: three-seed means on 64 shared episodes per family. Each entry averages the six observed spatial averages. The predictive mean and marginal variances are fixed.

Family	Covariance	Coverage	Width	IS
Conduction	Full	0.916	0.0637	0.0752
	Diagonalized	0.657	0.0353	0.1109
Cooling	Full	0.893	0.0725	0.0958
	Diagonalized	0.602	0.0396	0.1465
Cavity	Full	0.890	0.1296	0.1798
	Diagonalized	0.670	0.0821	0.2283

B.4 Joint NLL and aggregate interval scores

Table 11 retains each training-seed/cell joint NLL and mean aggregate interval score for Full ORIA, trained-diagonal ORIA, and TNP-ND. Joint NLL evaluates $-\log q_\theta(\tilde{u}_\star \mid \mathcal{C})/N$ with $N = 48$ on the noisy response. It scores the actual joint Gaussian, not a KL divergence to an available exact posterior.

Across the twelve equally weighted cells and three seeds, Full/trained-diagonal/TNP-ND joint NLL is $-1.532/-0.259/-1.070$, and mean aggregate IS is $1.212/1.818/2.143$. Pointwise coverages are $0.877/0.890/0.876$, aggregate coverages are $0.884/0.607/0.874$, and aggregate widths are $0.408/0.294/1.031$. TNP-ND therefore has similar average aggregate coverage to Full but wider intervals and a larger interval score. These average comparisons do not imply that one method dominates in every cell, spatial average, or training seed. All observations remain in the reported averages.

Table 11: Joint NLL and aggregate interval score by training seed. Each row uses the same 64 episodes; joint NLL is divided by $N = 48$, and IS averages the six observed spatial averages. Diag. denotes trained-diagonal ORIA. Lower is better for both scores.

Family	m	Seed	Full NLL	Diag. NLL	TNP NLL	Full IS	Diag. IS	TNP IS
Conduction	1	0	-1.330	0.061	-0.684	1.212	2.592	2.260
Conduction	1	1	-1.345	0.085	-0.678	1.402	2.439	2.423
Conduction	1	2	-1.279	0.044	-0.674	0.994	2.454	2.921
Conduction	2	0	-1.932	-1.109	-1.316	0.222	0.544	0.807
Conduction	2	1	-1.906	-1.031	-1.326	0.249	0.616	0.773

Table 11 (continued)

Family	m	Seed	Full NLL	Diag. NLL	TNP NLL	Full IS	Diag. IS	TNP IS
Conduction	2	2	-1.906	-1.109	-1.326	0.281	0.598	0.821
Conduction	4	0	-1.993	-1.485	-1.418	0.097	0.217	0.537
Conduction	4	1	-1.993	-1.419	-1.427	0.112	0.244	0.539
Conduction	4	2	-1.999	-1.485	-1.428	0.108	0.241	0.617
Conduction	8	0	-2.033	-1.703	-1.487	0.073	0.134	0.517
Conduction	8	1	-2.030	-1.650	-1.501	0.080	0.152	0.442
Conduction	8	2	-2.037	-1.754	-1.490	0.073	0.115	0.536
Cooling	1	0	-1.271	0.259	-0.973	1.405	3.269	3.015
Cooling	1	1	-1.277	0.261	-0.979	1.273	2.715	2.747
Cooling	1	2	-1.260	0.251	-0.969	1.287	3.124	3.488
Cooling	2	0	-1.446	1.402	-1.159	1.648	2.221	2.582
Cooling	2	1	-1.232	-0.061	-1.203	1.716	2.265	2.248
Cooling	2	2	-0.297	2.089	-1.172	1.843	2.330	2.564
Cooling	4	0	-2.001	-1.349	-1.638	0.114	0.235	0.612
Cooling	4	1	-1.985	-1.283	-1.637	0.133	0.298	0.590
Cooling	4	2	-1.985	-1.333	-1.651	0.129	0.263	0.585
Cooling	8	0	-1.945	-1.519	-1.573	0.099	0.185	0.635
Cooling	8	1	-1.934	-1.502	-1.577	0.090	0.179	0.565
Cooling	8	2	-1.936	-1.564	-1.590	0.099	0.173	0.574
Cavity	1	0	-0.482	1.968	-0.249	6.384	7.256	7.306
Cavity	1	1	-0.664	1.995	-0.264	6.407	7.918	7.144
Cavity	1	2	-0.593	2.312	-0.299	6.602	7.795	7.226
Cavity	2	0	-1.140	1.051	-0.593	2.784	3.220	3.729
Cavity	2	1	-1.173	1.148	-0.618	2.678	3.335	3.600
Cavity	2	2	-1.161	1.220	-0.655	2.174	3.503	3.574
Cavity	4	0	-1.517	0.043	-0.786	0.498	1.411	2.260
Cavity	4	1	-1.552	0.099	-0.834	0.389	1.286	2.266
Cavity	4	2	-1.564	-0.017	-0.775	0.434	1.128	2.270
Cavity	8	0	-1.649	-0.764	-0.860	0.191	0.343	1.458
Cavity	8	1	-1.649	-0.746	-0.879	0.174	0.315	1.389
Cavity	8	2	-1.642	-0.728	-0.822	0.174	0.346	1.548

B.5 FNO ENSEMBLES AND LAST-LAYER BAYES

The two additional FNO methods use the selected full-domain architecture and 2,000-step training recipe above. The ensemble has five members. Last-layer Bayes performs Bayesian linear regression, including an intercept, on the first ensemble member's final features. Fitting uses the actual observation locations. Both uncertainty baselines receive the true episode noise level, which is added to their predictive variance.

Table 12 reports the prediction errors used in Figure 2a, together with the original ORIA checkpoint evaluated on those same episodes. Table 13 reports pointwise noisy-response coverage, averaging locations within episodes before computing episode means and SEMs. These separate historical cohorts contain only $m = 2, 8$ and are not pooled with the shared-bank experiment. The coverage values concern marginal intervals and do not establish joint predictive quality.

Table 12: FEM clean-target relative L^2 errors for the FNO ensemble and last-layer Bayes in Figure 2a, with the same-episode original ORIA checkpoint. Entries are mean $\pm$ episode SEM. This cohort is separate from the shared-bank comparison.

Family	m	ORIA	FNO ensemble	Last-layer Bayes	Episodes
Conduction	2	0.110 ± 0.061	1.185 ± 0.309	1.200 ± 0.330	17
Conduction	8	0.031 ± 0.008	0.515 ± 0.103	0.527 ± 0.108	15
Cooling	2	0.093 ± 0.021	0.840 ± 0.122	0.850 ± 0.122	17
Cooling	8	0.050 ± 0.020	0.518 ± 0.059	0.535 ± 0.069	21
Cavity	2	0.481 ± 0.085	1.237 ± 0.204	1.244 ± 0.200	14
Cavity	8	0.081 ± 0.021	0.926 ± 0.108	0.939 ± 0.112	12

Table 13: FEM pointwise coverage of nominal 90% noisy-response intervals, mean $\pm$ episode SEM on the same episodes as Table 12. The FNO ensemble and last-layer Bayes receive the true observation-noise level.

Family	m	ORIA	FNO ensemble	Last-layer Bayes	Episodes
Conduction	2	0.902 ± 0.023	0.104 ± 0.027	0.061 ± 0.017	17
	8	0.913 ± 0.015	0.285 ± 0.032	0.169 ± 0.026	15
Cooling	2	0.864 ± 0.031	0.179 ± 0.034	0.066 ± 0.020	17
	8	0.894 ± 0.025	0.272 ± 0.025	0.144 ± 0.021	21
Cavity	2	0.851 ± 0.055	0.164 ± 0.021	0.068 ± 0.011	14
	8	0.889 ± 0.042	0.151 ± 0.017	0.043 ± 0.012	12

C PUBLIC NAVIER–STOKES: PROTOCOLS AND SUPPORTING RESULTS

C.1 DATA, EPISODE CONSTRUCTION, AND BASELINE SELECTION

The public NS comparison in Section 5.2 uses the vorticity trajectories accompanying the Fourier neural operator release, `ns_V1e-3_N5000_T50.mat`. The file contains 5,000 trajectories on a 64×64 periodic grid at viscosity 10^{-3}, with a fixed forcing and unit-time stored frames. The comparison predicts frame 4 from frame 2. The file therefore supplies many trajectories of one fixed physical parameter setting, not a separate operator for every trajectory.

Each episode draws $m + 1$ distinct trajectories: m provide demonstrations and the remaining trajectory provides the query. Within each context size, consecutive groups from a seeded permutation are disjoint; trajectories may be reused at different context sizes. Fields are divided by the pooled standard deviation of demonstration inputs, and responses receive an additional scale estimated from the noisy demonstration responses. Demonstration functions and the query input are observed at $M = 16^2$ locations on a jittered product grid X with periodic trapezoidal weights. The response is predicted at $N = 64$ uniformly sampled locations Y. Independent observation noise has standard deviation 0.03 relative to the clean demonstration response scale. Error uses the clean query response; pointwise coverage uses the noisy query response.

The original data split reserves trajectory indices 0–999 for development and 1000–4999 for the registered evaluation block (zero-based indexing). Figure 2e uses the first 48 registered episodes at each $m \in \{1, 2, 4, 8\}$, shared by every displayed method. Figure 4c,d uses 96 episodes at $m = 8$. The registered block was subsequently reused for protocol corrections and prior revisions; it is not an untouched test set after selection of the revised prior.

Affine estimator. The affine baseline fits a separate complex affine relationship between input and response coefficients in each Fourier mode, using the m demonstrations and fixed regularization 10^{-3} on its slope and intercept. Inputs are evaluated on a 32×32 uniform grid and noisy responses are interpolated onto that grid. Predictions are evaluated at the same $N = 64$ query locations as ORIA.

FNO baselines. Scratch FNO is fitted to each episode's noisy demonstration pairs. Its width 32, 12 modes, three layers, 1,000 training steps, and learning rate 10^{-3} were selected on independent synthetic validation episodes. Scores average errors over three initializations. The pretrained FNO uses the same NS training prior followed by per-episode fine-tuning. A 48-episode synthetic validation bank selects from zero updates, 100 or 300 steps at 10^{-4}, 300 steps at 3×10^{-4}, and 800 steps at either rate. The selected recipe uses 300 steps at 3×10^{-4}. The retained pretraining result estimates response normalization from the other available pairs rather than the pair being predicted.

Poseidon-B. Poseidon-B is reset to its pretrained state for every episode and all weights are adapted to the m demonstrations. Its pretrained input and output layers are retained. The observation-parity comparison receives the same noisy 16×16 observations and query input as ORIA, interpolated to 128×128.

Adaptation follows the official recipe, with input/output-layer learning rates ten times the body rate. Twelve development episodes at $m = 4$ compare 100 or 300 steps and body rates 5×10^{-5} or 2×10^{-4}. Mean relative error selects 300 steps at 5×10^{-5}, used at every m. The reported runs use the first 48 registered episodes. FNO and Poseidon-B have different validation resources and pretraining exposure; observation parity does not imply equal training or tuning budgets.

C.2 MATCHED ACCURACY AND COVERAGE RESULTS

Table 14 gives the five prediction curves and ORIA's noisy-response coverage in Figure 2e. All methods use exactly the same 48 episodes per context size. At $m = 8$, ORIA's relative error is 0.058 and nominal 90% pointwise coverage is 0.712. Better mean prediction therefore coexists with undercoverage. Episode SEM describes variation within this cohort, not across trained models or independently drawn physical operators.

The original-checkpoint star in Figure 4c is the same trained model evaluated on all 96 episodes: its error is 0.060 and coverage is 0.702. The $0.058/0.712$ pair uses the first 48 of those episodes. Neither pair describes revision D or E.

Table 14: Public NS prediction from frame 2 to frame 4 at viscosity 10^{-3}: the first 48 episodes in Figure 2e. Values are clean-target relative L^2 errors (mean $\pm$ episode SEM), except the final row, which reports noisy-query pointwise coverage at nominal 90%. Every method uses the same episodes.

Method	$m = 1$	$m = 2$	$m = 4$	$m = 8$	Episodes
ORIA	0.367 ± 0.018	0.136 ± 0.008	0.079 ± 0.004	0.058 ± 0.003	48
Affine	0.589 ± 0.021	0.255 ± 0.016	0.167 ± 0.006	0.141 ± 0.005	48
FNO	0.300 ± 0.013	0.276 ± 0.014	0.257 ± 0.013	0.204 ± 0.012	48
Pretrained FNO + FT	0.508 ± 0.036	0.469 ± 0.034	0.377 ± 0.024	0.334 ± 0.020	48
Poseidon-B (parity)	0.313 ± 0.016	0.250 ± 0.012	0.239 ± 0.014	0.177 ± 0.012	48
ORIA noisy coverage	0.202 ± 0.012	0.515 ± 0.030	0.658 ± 0.019	0.712 ± 0.017	48

C.3 Prior revisions and the original-prior tradeoff

The original frozen model uses the 2D physics-mixture prior. Base denotes new training-seed controls under that same prior, with regenerated NS banks. The original NS bank contained 24,000 episodes; each Base control and revision instead uses 18,000 training episodes and a separate 6,000-episode synthetic validation shard. The controlled comparison is between these new Base runs and the revisions at fixed architecture and training budget. The original checkpoint is an additional reference, not a different prior.

Revision D broadens the NS input spectral-exponent range from $[1, 2]$ to $[1, 5]$, permits input amplitudes in $[0.6, 1.4]$ independently across pairs, and adds low-mode forcing to the original random-field forcing law. E additionally lowers the viscosity floor from 10^{-3} to 10^{-5} and extends forcing amplitudes to zero. Only the NS component of the mixture is changed; its other components retain the Base prior. Each condition is trained with two seeds.

Table 15 gives the relative error, pointwise coverage, and joint NLL values behind Figure 4c,d. Public NS uses the 96 reused episodes at $m = 8$. The original-prior evaluation uses 48 fresh synthetic NS episodes per seed at the same context size. Figure 4d subtracts Base from the revision with the same training seed before averaging. The displayed ranges span the two seed means and are distinct from episode SEMs.

D reduces external mean error from 0.110 to 0.033 and raises coverage from 0.542 to 0.832. Both D and E improve external error and joint NLL, while both worsen these scores under the original NS prior. E produces smaller external gains and larger original-prior costs than D at this budget.

The synthetic validation shard does not provide an independent public-data validation set. These revisions were informed by earlier external outcomes and evaluated on the previously used public block. Corresponding D/E retrains of the external baselines are unavailable. The comparison therefore measures ORIA's response to the declared prior changes, without establishing a fresh blinded external test or equal prior-selection opportunities across methods.

Table 15: Source values for Figure 4c,d at $m = 8$: two-seed means, with ranges of seed means in brackets and episode counts per seed. Public NS predicts frame 4 from frame 2; original-prior NS uses fresh synthetic draws. Joint NLL is per prediction location in nats. Same-seed differences from Base give Figure 4d.

Prior	Evaluation	Relative error	Noisy coverage	Joint NLL	Episodes
Base	Public NS	$0.110\ [0.076, 0.143]$	$0.542\ [0.403, 0.682]$	$-1.659\ [-1.665, -1.652]$	96
Base	Original-prior NS	$0.040\ [0.036, 0.043]$	$0.915\ [0.913, 0.916]$	$-1.657\ [-1.706, -1.607]$	48
D	Public NS	$0.033\ [0.021, 0.044]$	$0.832\ [0.751, 0.913]$	$-1.899\ [-1.948, -1.851]$	96
D	Original-prior NS	$0.053\ [0.053, 0.054]$	$0.940\ [0.939, 0.940]$	$-1.480\ [-1.492, -1.467]$	48
E	Public NS	$0.063\ [0.056, 0.069]$	$0.685\ [0.647, 0.723]$	$-1.804\ [-1.826, -1.782]$	96
E	Original-prior NS	$0.067\ [0.065, 0.070]$	$0.920\ [0.919, 0.920]$	$-1.399\ [-1.419, -1.380]$	48

D POSTERIOR-GUIDED MEASUREMENT SELECTION

Measurement and selection rules. Each measurement selects an input function and records its noisy response on the measurement grid X, adding one complete demonstration to $\mathcal{D}_m^X$. A fixed set of target input functions and prediction locations Y defines the evaluation task. Their inputs and locations are available during selection; their responses are withheld. Every acquired history in Figure 3 is evaluated by the same frozen ORIA predictor, conditioned on that history.

ORIA joint selects the candidate with the largest expected reduction in the mean marginal predictive variance over target functions and locations. For each candidate, it draws noisy response fields jointly from the learned Gaussian, temporarily conditions on each draw, and averages the resulting target variances. The observed response, rather than an imagined draw, is then added to the actual context. Exact greedy uses the same one-step variance-reduction objective with the exact Bayesian posterior. ORIA max-var selects the candidate with the largest mean marginal variance across its response locations. Diversity selects the input farthest from its nearest measured input, and Random samples uniformly. The joint and max-var rules differ in their use of target inputs, imagined observations, and reconditioning; their comparison does not isolate the contribution of off-diagonal covariance.

Linear-Gaussian experiment. The experiment uses 24 unseen operators from the nine-mode prior of Appendix A.1, with 16 unit-energy candidate inputs, six target inputs, $M = 48$ measurement locations, and $N = 24$ prediction locations. The pool contains nine basis modes, four random mixtures, and three composites spanning low, high, and broad frequencies. An independently drawn initial input and its noisy response are shared across the five strategies. Six subsequent choices give $m = 2, \ldots, 7$; candidates remain available, so repeated measurements are permitted. Subsequent observation noise is drawn separately for each strategy, although the operator and initial observation are shared. The errors below are means over operators; the error bars in Figure 3b are operator standard errors.

ORIA joint uses six imagined response fields per candidate. Under the fixed linear-Gaussian prior, exact posterior covariance depends on the chosen inputs and noise law, not on the measured response values. Exact greedy therefore ranks candidates without observing those values. It is an exact one-step reference for this pool, not an optimal multistep design. The noisy-response variance includes a fixed observation-noise term that cancels when taking its reduction.

Table 16: Clean-target relative L^2 error in the 24-operator linear-Gaussian experiment. Entries are operator means $\pm$ standard error for the displayed checkpoint. All histories use the same frozen ORIA predictor; subsequent observation noise differs across strategies.

m	ORIA joint	Exact greedy	ORIA max-var	Diversity	Random
2	0.806 ± 0.021	0.805 ± 0.021	0.845 ± 0.033	0.930 ± 0.022	0.885 ± 0.027
3	0.675 ± 0.026	0.675 ± 0.026	0.757 ± 0.038	0.867 ± 0.026	0.848 ± 0.037
4	0.568 ± 0.025	0.573 ± 0.027	0.607 ± 0.019	0.776 ± 0.029	0.788 ± 0.039
5	0.451 ± 0.025	0.459 ± 0.023	0.519 ± 0.022	0.662 ± 0.030	0.676 ± 0.037
6	0.352 ± 0.020	0.351 ± 0.020	0.431 ± 0.023	0.553 ± 0.034	0.606 ± 0.033
7	0.243 ± 0.018	0.240 ± 0.013	0.341 ± 0.021	0.413 ± 0.035	0.537 ± 0.038

These trajectories support the numerical proximity reported in the main text, rather than equivalence of the two selection procedures. The final prediction errors combine selection with a learned ORIA predictor; they are not errors of the exact posterior mean. The recorded implementation uses 193 model forward calls for one ORIA joint choice, including repeated candidate predictions for the six imagined responses, versus 16 individual candidate predictions for ORIA max-var. These are implementation counts, not measured hardware-normalized speedups.

FEM protocol. Each of conduction, cooling, and Helmholtz has a separate bank of 64 specimens evaluated with three fixed training checkpoints. A specimen supplies 12 candidate input functions, four target functions, $M = 196$ response locations, and $N = 48$ target locations. One randomly selected demonstration is shared across strategies. Five further choices without replacement give $m = 1, \ldots, 6$. Candidate noise is generated once and reused across strategies. Selection uses noisy-response uncertainty; evaluation averages clean-target relative L^2 error over the four target

functions. Response normalization is recomputed from the noisy observed history. The simulator's clean response scale is not given to the selector. Input diversity uses quadrature-weighted squared input distance with the weights normalized to sum to one.

Within a conduction specimen, the coefficient field and boundary conditions remain fixed while forcing varies. Cooling keeps the material, boundary conditions, and terminal time fixed and resets the system for each new initial temperature field. Helmholtz keeps the medium, frequency, and boundary conditions fixed while the spatial source varies. New specimens use the same priors and FEM pipeline as training. These results therefore assess selection within those physical families, rather than transfer to an independent simulator.

AUC and uncertainty. The error-curve AUC is a normalized trapezoidal average, with weights $(1/8, 1/4, 1/4, 1/4, 1/8)$ to budgets $m = 2, \ldots, 6$, excluding the shared initial observation. Three checkpoint values are averaged within each specimen before strategies are paired. The independent sample size is consequently 64 per task, not 192. Relative reductions are one minus the ratio of aggregate mean AUCs, expressed as percentages, rather than the mean of specimen-wise percentages. The intervals use 10,000 paired specimen bootstrap resamples after checkpoint averaging. They are descriptive, conditional on these three checkpoints, and unadjusted for multiple comparisons; they do not estimate training-seed uncertainty.

Table 17: FEM AUC under the common frozen ORIA predictor. Each entry averages the same three fixed checkpoints within each of 64 specimens. The relative reductions and paired-bootstrap 95% intervals compare ORIA max-var with the named strategy.

Task	Strategy	Mean AUC	Reduction (%)	95% interval (%)
Conduction	ORIA max-var	0.03724	–	–
	Diversity	0.03824	2.63	$[-1.27, 6.23]$
	Random	0.04380	14.98	$[5.42, 26.72]$
Cooling	ORIA max-var	0.05580	–	–
	Diversity	0.05947	6.18	$[2.32, 10.02]$
	Random	0.06418	13.06	$[7.29, 18.58]$
Helmholtz	ORIA max-var	0.19581	–	–
	Diversity	0.22262	12.04	$[7.13, 17.00]$
	Random	0.23978	18.34	$[12.09, 23.87]$

The ORIA-versus-Random intervals exclude zero in all three tasks. Against Diversity, the conduction interval includes zero, while cooling and Helmholtz have positive reductions throughout their intervals. These results concern the common ORIA predictor and the stated pools; they do not imply equal benefits for other predictors. Lower AUC does not imply an equal percentage reduction in required measurements. Because the FEM strategy uses only marginal variance for selection, its gains do not independently identify a benefit of cross-location covariance in the acquisition rule.

FEM computational cost. ORIA max-var evaluates the remaining candidates in a batch at each choice. The following recorded costs include normalization, data movement, and scoring, and exclude physical measurement and subsequent target prediction. Diversity and Random run on the CPU; ORIA runs on the GPU named in the table. Hardware differs across tasks, so these numbers do not support a speed comparison between tasks.

Privileged benchmark. The non-attainable benchmark mentioned in the conclusion uses these same 64 specimens per task, the three frozen checkpoints, and the common ORIA predictor. At each step it temporarily adds each remaining candidate's actual noisy response, evaluates prediction error using the clean target truth, and chooses the smallest realized next-step error. Compared with ORIA max-var, it reduces mean AUC by 27.86%, 27.45%, and 36.80% for conduction, cooling, and Helmholtz. The reductions are ratios of checkpoint-averaged specimen means. Although every specimen improves after averaging checkpoints, some individual checkpoint and budget comparisons worsen. This benchmark requires unobserved responses and target truth, so it is not a deployable strategy. It is a greedy comparison in the fixed pool, not a global optimum or evidence that a selector with access only to observed demonstrations can realize those gains.

Table 18: Mean cumulative acquisition seconds per specimen over five choices, separated by fixed checkpoint. ORIA's recorded peak allocated GPU memory is 51.7 MiB in each case.

Task and GPU	Strategy	Seed 0	Seed 1	Seed 2
Conduction	ORIA max-var	0.433479	0.430047	0.460361
RTX 4080 SUPER	Diversity	0.001005	0.001000	0.000933
	Random	0.000404	0.000426	0.000389
Cooling	ORIA max-var	0.422076	0.425409	0.434124
RTX 4080 SUPER	Diversity	0.001342	0.000874	0.000932
	Random	0.000379	0.000382	0.000375
Helmholtz	ORIA max-var	0.292759	0.295480	0.300482
RTX 4090	Diversity	0.000645	0.000709	0.000858
	Random	0.000273	0.000286	0.000278

E CONTROLLED PRIOR MISMATCH AND RELIABILITY DIAGNOSTICS

E.1 GRAMMAR AND EVALUATION CONDITIONS

Figure 4a,b uses the frozen 1D grammar model, separate from the linear-Gaussian model and the 2D physics-mixture model. Its periodic functions on $[0, 1)$ are represented by finite Fourier series. Training samples compositions of one or two classical operator and pointwise-nonlinearity components, with continuously varying parameters. The input functions use independent Gaussian Fourier coefficients up to mode 10. Their amplitudes decay with frequency; the roughness exponent ranges uniformly from 1.2 to 2.5 during training, with smaller values producing rougher functions. The coefficient variances are normalized to sum to one. The observation-noise standard deviation σ_ε is sampled log-uniformly from $[0.01, 0.1]$ for each episode. This is the noise in the data, distinct from the predicted residual standard deviation σ_n of Eq. equation 10. Outputs are normalized using independent input functions to a target standard deviation between 0.5 and 1.5.

The test cohort uses $m = 8$ demonstrations, $M = 48$ measurement locations, $N = 24$ query locations, and 128 episodes per condition. Roughness conditions set the exponent to 1.0, 0.7, or 0.5; the scale condition changes the target-standard-deviation range to $[2.5, 3.5]$. The approximately threefold scale change refers to the centres of these ranges, not multiplication of every episode by exactly three. Noise conditions set σ_ε to 0.2 or 0.4. Structural conditions use an unseen pointwise nonlinearity or the Burgers solution map shown in Figure 4a. These conditions represent different kinds of departure from the prior, rather than ordered levels of a common shift magnitude.

The episode bank and evaluation conditions were fixed before inference. The error, predictive standard deviation, and coverage results use one frozen checkpoint. The entries below are means with episode standard errors; they quantify variation within this cohort, not uncertainty over training seeds. Root-mean-square (RMS) error compares the predictive mean with the clean query response, as in Figure 4a. Predictive standard deviation and nominal 90% pointwise coverage refer to noisy query measurements.

Table 19: Controlled mismatch with eight demonstrations and 128 episodes per condition. Entries are means $\pm$ episode standard error for the checkpoint in Figure 4a. Roughness entries give the input spectral-decay exponent.

Condition	RMS error	Predictive s.d.	Coverage
In distribution	0.098 ± 0.015	0.116 ± 0.014	0.897 ± 0.010
Roughness 1.0	0.150 ± 0.018	0.177 ± 0.018	0.922 ± 0.007
Roughness 0.7	0.230 ± 0.026	0.229 ± 0.023	0.889 ± 0.011
Roughness 0.5	0.306 ± 0.033	0.309 ± 0.028	0.905 ± 0.010
Output scale $\approx 3\times$	0.273 ± 0.032	0.287 ± 0.028	0.884 ± 0.014
Noise $\sigma_\varepsilon = 0.2$	0.120 ± 0.013	0.218 ± 0.014	0.803 ± 0.011
Noise $\sigma_\varepsilon = 0.4$	0.260 ± 0.020	0.440 ± 0.020	0.785 ± 0.015
Known nonlinearity (matched)	0.213 ± 0.028	0.230 ± 0.023	0.913 ± 0.012
Unseen nonlinearity	0.270 ± 0.025	0.150 ± 0.009	0.668 ± 0.025
Burgers	0.267 ± 0.016	0.181 ± 0.009	0.706 ± 0.022

E.2 MATCHED STRUCTURAL CONTROL

The unseen-nonlinearity condition fixes the composition, input roughness exponent at 1.8, and observation-noise standard deviation at 0.03. Its comparison with the general in-distribution condition therefore changes more than the pointwise nonlinearity. The matched condition keeps the same composition structure, input law, grid protocol, noise, and normalization, while using a pointwise function drawn from the training pool. These construction rules match, but the realized episodes are independent rather than paired. The matched condition is consequently a structural control, not an additional sample from the general in-distribution law.

Replacing the known nonlinearity with the unseen one increases mean clean relative L^2 error from 0.244 to 0.372, while mean predictive standard deviation falls from 0.230 to 0.150 and coverage from 0.913 to 0.668. Thus the observed loss of accuracy is accompanied by narrower predictive intervals. These comparisons are conditional on the same single checkpoint as Figure 4a.

Diagnostic using observed demonstrations. For each of the eight demonstrations, ORIA conditions on the other seven and predicts the held-out noisy response at its 48 observed locations. The diagnostic forms the residual from the observed response and predictive mean μ, weights its squared magnitude by the inverse full predictive covariance Σ, and divides by 48. Averaging this quantity over the eight held-out demonstrations gives one score per episode. The eight folds are repeated evaluations within an episode, not independent samples.

The diagnostic uses only observed demonstrations. It does not use the new query input or response, clean fields, simulator parameters, or the true noise level. The learned covariance already describes noisy responses; no known noise term is added or subtracted. The same score is used for every test condition and every checkpoint. AUROC compares 128 episodes of a condition with 128 independent in-distribution test episodes. The in-distribution episodes are shared across these comparisons; the known-nonlinearity condition remains a construction control. The reported 95% intervals use 1,000 stratified episode-bootstrap resamples.

Table 20: Full-covariance diagnostic AUROC across three fixed checkpoints trained with the same recipe. The checkpoints score identical episodes; columns are paired. Intervals are descriptive, marginal, and unadjusted for multiple comparisons. Entries are rounded to three decimals; rounded values of 1.000 do not indicate a perfect unrounded AUROC.

Condition	Seed 0	Seed 1	Seed 2
Roughness 1.0	$0.577\ [0.507, 0.644]$	$0.550\ [0.479, 0.620]$	$0.601\ [0.534, 0.668]$
Roughness 0.7	$0.504\ [0.430, 0.581]$	$0.479\ [0.411, 0.556]$	$0.569\ [0.499, 0.639]$
Roughness 0.5	$0.606\ [0.538, 0.674]$	$0.511\ [0.440, 0.583]$	$0.631\ [0.561, 0.697]$
Output scale $\approx 3\times$	$0.567\ [0.497, 0.635]$	$0.574\ [0.510, 0.645]$	$0.554\ [0.485, 0.628]$
Noise $\sigma_\varepsilon = 0.2$	$1.000\ [0.999, 1.000]$	$0.994\ [0.988, 0.999]$	$1.000\ [1.000, 1.000]$
Noise $\sigma_\varepsilon = 0.4$	$0.999\ [0.997, 1.000]$	$0.966\ [0.942, 0.984]$	$1.000\ [0.999, 1.000]$
Known nonlinearity	$0.471\ [0.396, 0.544]$	$0.451\ [0.374, 0.525]$	$0.523\ [0.445, 0.593]$
Unseen nonlinearity	$0.896\ [0.854, 0.932]$	$0.931\ [0.896, 0.960]$	$0.887\ [0.847, 0.924]$
Burgers	$0.964\ [0.940, 0.984]$	$0.965\ [0.943, 0.983]$	$0.987\ [0.974, 0.997]$

Scope of the diagnostic. The two noise and two structural conditions give AUROC between 0.887 and 0.9999 across the three checkpoints. Roughness and output-scale conditions remain weakly separated from the in-distribution episodes, with AUROC between 0.479 and 0.631. Neither high AUROC nor a small diagnostic score certifies a reliable posterior or corrects prior mismatch. The three-checkpoint comparison covers this diagnostic only: the error, predictive spread, and coverage in the first table still come from one checkpoint. The shared episode bank does not triple the sample size, and three realized training runs do not quantify uncertainty over a population of training seeds. These results also do not establish detection power for shifts outside the listed conditions.

F UNIFORM CONTEXT REPLICATION AND ITS BAYESIAN SCOPE

The inter-function attention in Eq. equation 8 uses only the m demonstrations as keys and values. Its normalized weights explain the uniform-context-replication property mentioned in the main text. Here, replication means copying every existing demonstration, including its already observed noise, the same number of times while keeping X, w, $f_\star(X)$, and Y fixed. It does not mean making new measurements of the response.

Why the prediction is unchanged. Identical copies have identical input tokens and therefore identical attention scores and values. For example, duplicating every demonstration doubles both the total contribution associated with each original demonstration and the normalization denominator. Their ratio, and hence the inter-function update, is unchanged. The same cancellation holds when every demonstration is copied any common number of times. The query is excluded from keys and values, so there is no additional query term in that denominator.

The remaining encoder operations act in the same way on each copy: intra-function attention, layer normalization, residual updates, and the pointwise feed-forward layers preserve identical representations. Applying this argument through the encoder leaves the query representation unchanged. The decoder then produces the same μ and Σ, and thus the same q_θ. This statement concerns deterministic inference in exact arithmetic; floating-point reductions can introduce small numerical differences. Copying only some demonstrations can change their relative attention weights and is not covered by this identity.

What this means for Bayesian inference. Copies of an already observed noisy response provide no additional measurement information when they are known to be copies. Independent measurements at the same input are different: their observation noise is drawn again, and conditioning on these measurements can further contract the Bayesian posterior predictive. The future measurement still has the observation-noise variance σ_ε^2.

The tokens specified in the main text contain function values, an observed response, a demonstration flag, and spatial position features. They do not separately encode whether equal observed values came from copied records or independent measurements. Consequently, the replication identity limits the architecture's ability to represent $P^\star(\cdot \mid \mathcal{C})$ for every possible independently measured context. It does not imply that new measurements are uninformative, and it does not establish the cause of the beyond-training-range posterior-contraction errors reported in Section 5.1.

G ARCHITECTURE OVERVIEW

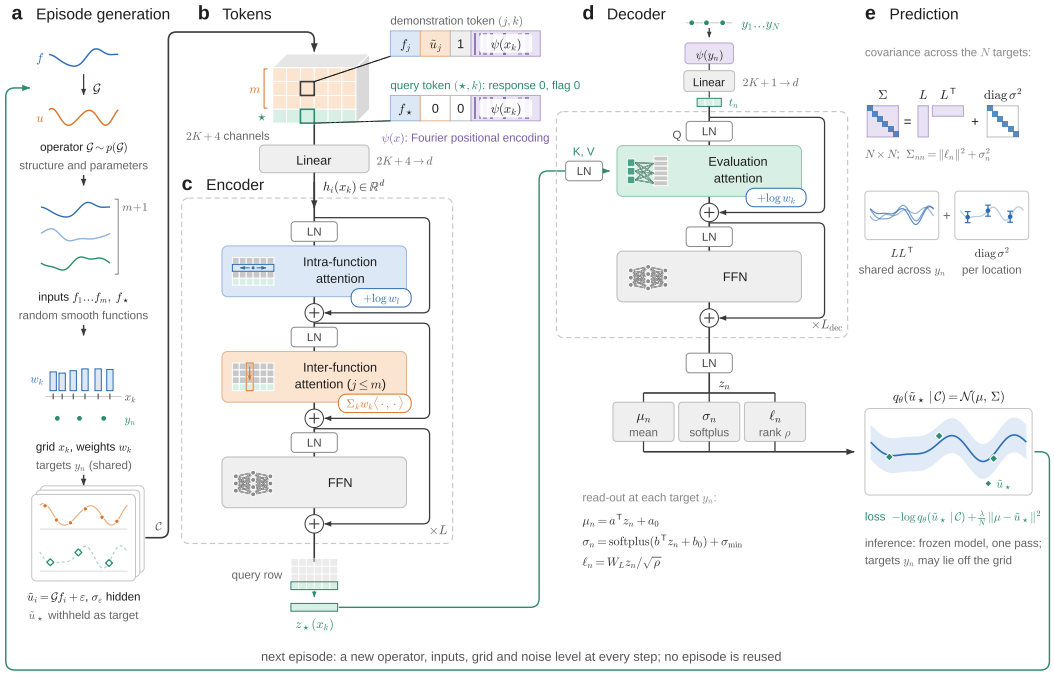

Figure 6: **The ORIA architecture.**