NEURAL OPERATOR PROCESSES FOR PROBABILISTIC OPERATOR LEARNING UNDER PARTIAL OBSERVATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural operators learn mappings between function spaces, but are typically developed with dense input–output training fields and fully observed inputs at inference. Many scientific problems require instead predicting solution fields from sparse, irregular, or partial observations under uncertainty. We introduce Neural Operator Processes (NOPs), a framework that unifies neural-process conditioning with neural-operator decoding to predict full output fields from limited context. NOPs condition on sparse joint input–output observations and support deterministic and probabilistic prediction within a shared encoder–decoder architecture. We study two conditioning strategies, convolutional pooled summaries and query-aligned attention, and analyze how their interaction with latent stochastic variables depends on PDE geometry. Across function regression and three PDE benchmarks, we find that sparse conditional operator learning is viable and can match dense-grid behavior in several regimes, that preserving local context-query geometry is essential in non-periodic settings but less so in spectrally smooth periodic regimes, and that uncertainty-aware operator learning succeeds when latent conditioning complements rather than overwrites the local geometric pathway. These results provide a basis for probabilistic operator learning under partial observations and help bridge operator learning and probabilistic meta-learning in function space. [1]

1 INTRODUCTION

Learning mappings between function spaces has become a central problem in scientific machine learning. In many applications, the goal is not to learn a finite-dimensional input–output map, but an operator that maps one field to another, such as a coefficient field to the solution of a partial differential equation (Kovachki et al., 2024; Li et al., 2020). Neural operators (NOs), particularly architectures such as the Fourier Neural Operator (FNO), have shown a strong performance in this regime and established operator learning as an effective paradigm for surrogate modeling, simulation acceleration, and PDE-based inference (Li et al., 2021; Lu et al., 2021). Many scientific settings provide both a structured input field and sparse measurements of the corresponding physical response. Full-field prediction must then integrate these irregular, instance-specific observations with the underlying operator input, often under uncertainty. This setting arises naturally in applications such as data assimilation and sparse sensing (Cheng et al., 2023; Blanke et al., 2024).

Uncertainty quantification remains comparatively underdeveloped in operator learning, particularly under sparse response conditioning. Many scientific settings require full-field prediction from a known operator input and only a small, irregular set of response measurements, where uncertainty reflects not just pointwise noise but task ambiguity induced by limited context. Existing probabilistic operator-learning approaches typically assume fully observed fields and do not directly address variable sparse response conditioning (Bülte et al., 2025; Magnani et al., 2022).

Neural Processes (NPs) offer a complementary perspective by learning distributions over functions from sparse context sets via amortized latent-variable inference, enabling rapid adaptation to new tasks without per-task optimization and providing predictive uncertainty (Garnelo et al., 2018b; Lara-Rangel et al., 2025; Jha et al., 2022). However, standard NPs are designed primarily for pointwise function regression and have neither been adapted nor widely applied to structured function-to-function mappings that demand dense PDE-scale decoding.

[1] Code release is planned upon publication.

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.

In this work, we introduce the *Neural Operator Process* (NOP), a framework for probabilistic operator learning under partial response observations that combines the set-conditioned inference of NPs with the structured operator decoding of NOs. NOPs condition predictions on a sparse context set of joint input–response observations and predict the solution field over a dense query grid. The query-side operator input is available over the prediction domain, while the number and geometry of response observations may vary across tasks. This yields an episodic formulation of operator regression with variable context availability. The framework naturally supports both deterministic and probabilistic variants, enabling a unified study of reconstruction accuracy, latent conditioning, and uncertainty quantification.

We study two complementary conditioning mechanisms: a SetConv pathway (Gordon et al., 2020), which lifts sparse context into a continuous representation, and a query-aligned attention pathway (Kim et al., 2019), which preserves local geometric structure. We extend both to probabilistic variants by introducing a latent variable that captures task-level uncertainty while retaining the structured operator decoder. We examine how these pathways behave across PDE regimes and between deterministic and uncertainty-aware prediction.

Our experiments span four complementary benchmark families. Gaussian-process regression serves as a probabilistic diagnostic for sparse conditioning, latent inference, and uncertainty quantification. PDE experiments cover Burgers as a smooth periodic setting suited to Fourier decoding, Darcy as a non-periodic and geometry-sensitive elliptic problem, and Navier–Stokes as a dynamically richer periodic 2D regime. Our results show that sparse conditional operator prediction is viable when conditioning on only a small fraction of response observations, that preserving local context–query geometry is regime-dependent and can be decisive in geometry-sensitive settings, that the framework provides informative uncertainty quantification, and that probabilistic operator learning is most reliable when latent conditioning preserves and complements the local geometric pathway. Controlled ablations further examine observation geometry, test-time context scaling, train–evaluation mismatch, cross-resolution transfer, local-conditioning design, and computational cost.

More broadly, our results suggest that operator learning and probabilistic meta-learning can be unified within a common framework that accounts jointly for structure, sparsity, uncertainty, adaptation, and probabilistic prediction in function space. Our main contributions are:

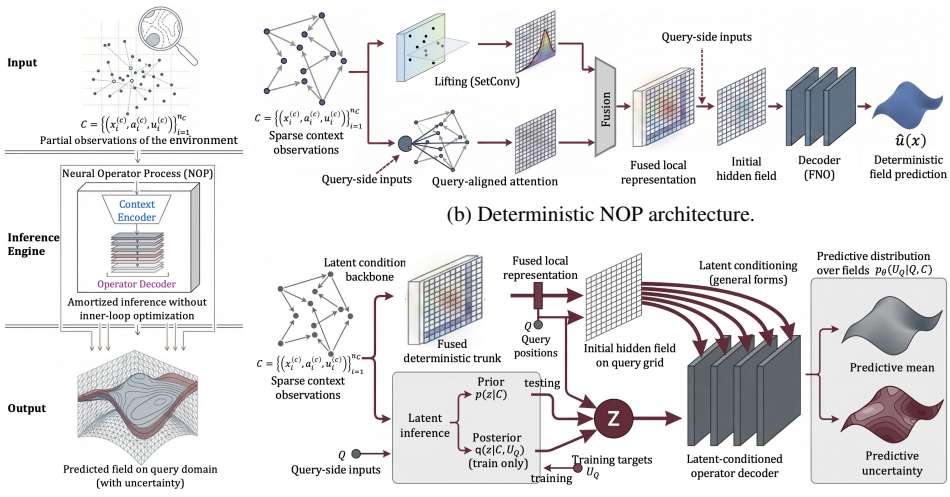
- We formulate operator learning with partial response observations as context-conditioned prediction in function space and propose NOPs as a corresponding framework.
- We introduce deterministic and probabilistic NOP architectures sharing an encoder–decoder pipeline, with different mechanisms for context aggregation and latent uncertainty.
- We identify empirical distinctions between set-based and query-aligned conditioning, with attention being especially valuable in geometry-sensitive settings, while the preferred pathway depends on the inference objective.
- Our experiments show that NOPs can match dense-grid Neural Operators using sparse response context, provide informative full-field uncertainty, and remain computationally competitive at inference.

2 Background and Related Work

Meta-Learning and Neural Processes. Meta-learning studies how to use experience from related tasks to adapt quickly to a new task from limited data (Finn et al., 2017). Given a context dataset C, rather than fitting a single predictor over pooled data, meta-learning trains over related tasks to learn the map $C \mapsto f(\cdot; C)$, that is, an inference mechanism that adapts from a small context set (Garnelo et al., 2018a).

Neural Processes instantiate this view by learning a direct mapping from a sparse context set $C = \{(x_i, y_i)\}_{i=1}^{n_C}$ to a predictive distribution $p_\theta(Y_T \mid X_T, C)$ over function values at target inputs (Jha et al., 2022). The context is treated as a *set* with variable cardinality and no canonical ordering, requiring permutation-invariant processing (Zaheer et al., 2017; Garnelo et al., 2018a).

In conditional NPs, context pairs are aggregated through a commutative operator to form a global summary, and a decoder conditioned on it outputs pointwise Gaussian predictions (Garnelo et al., 2018a). Latent NPs extend this framework with a global stochastic latent variable z, yielding



(a) NOP framework.

(b) Deterministic NOP architecture.

(c) Probabilistic NOP architecture.

Figure 1: **NOP Pipeline.** (a) NOPs combine sparse context with query information to predict the full solution field at query locations. (b) Deterministic NOPs combine a local context-conditioned representation with query information and pass the resulting hidden field through an NO decoder. (c) Probabilistic NOPs additionally infer a shared latent variable, using the target-informed posterior during training and the context-conditioned prior at test time, and marginalize it to obtain a predictive distribution over output fields. Arrows depict general decoder-wide latent conditioning, our main models inject the latent only through the initial hidden field and B.3 compares both strategies.

predictions by integrating over its context-conditioned distribution. This allows the model to represent coherent task-level uncertainty shared across targets (Garnelo et al., 2018b).

A known limitation is that aggressive global aggregation can discard local query-relevant structure (Jha et al., 2022). Attention-based variants address it by allowing each target to attend directly to the context set (Kim et al., 2019), while convolutional variants lift sparse contexts into continuous representations via kernelized set convolutions before convolutional processing (Gordon et al., 2020).

Neural Operators. Neural Operators (NOs) aim to learn mappings between function spaces, rather than between finite-dimensional vectors (Kovachki et al., 2024). By targeting the underlying operator itself, instead of a single discretization, they can in principle generalize across resolutions (Li et al., 2020). NOs are typically trained on fully paired dense input–output fields. Particularly, the FNO (Li et al., 2021) parameterizes the operator through spectral convolution layers:

$$v_{\ell+1}(x) = \sigma\big(W_\ell v_\ell(x) + \mathcal{F}^{-1}\big(R_\ell(k)\,\mathcal{F}(v_\ell)(k)\big)(x)\big)\,,$$

where $\mathcal{F}$ is the Fourier transform, $R_\ell(k)$ is a learned spectral multiplier truncated to finite modes, W_ℓ is a pointwise linear map, and σ is a nonlinearity. This gives FNO a strong inductive bias toward globally coupled operators and is especially effective in periodic or spectrally smooth settings.

Recent work has introduced probabilistic and meta-learned extensions of NOs (Zhang et al., 2023; Magnani et al., 2022; Bülte et al., 2025) generally under fully observed-field settings. We consider a complementary setting where sparse response observations are available as context and their number and locations may vary across episodes. NPs provide set-conditioned inference from these observations, while NOs provide structured decoding over the full query field. The challenge is to combine both mechanisms while retaining efficient dense prediction and probabilistic inference.

3 Neural Operator Processes

3.1 Problem Setup

Let $\Omega \subset \mathbb{R}^{d_x}$ be the spatial domain and $\mathcal{G} : a \mapsto u$ an operator mapping an input field a to an output field u. Each task corresponds to an underlying input–output field pair $(a(\cdot), u(\cdot))$, for example a coefficient field and its PDE solution, or an initial condition and its evolved state.

For a given task, we sample a sparse context set $C = \{(x_i^{(c)}, a_i^{(c)}, u_i^{(c)})\}_{i=1}^{n_C}$, $x_i^{(c)} \in \Omega$, and define a query set $Q = \{(x_j^{(q)}, a_j^{(q)})\}_{j=1}^{n_Q}$ with target outputs $U_Q = \{u_j^{(q)}\}_{j=1}^{n_Q}$. The context C provides sparse joint input–response observations for conditioning, while Q specifies the prediction locations and corresponding query-side inputs. In our PDE experiments, Q is the full structured grid and the context locations form a sparse subset of it. The target field U_Q provides training supervision and is unavailable for test-time conditioning.

This defines a conditional operator-prediction problem from sparse response observations. The model uses the context C together with the query-side information Q to predict the response over the full query grid. In our episodic construction, conditioning is sparse while supervision covers the full query field. In the deterministic setting, the goal is $\hat{U}_Q = f_\theta(Q, C)$; in the probabilistic setting, $p_\theta(U_Q \mid Q, C)$.

3.2 Model Overview

As shown in Figure 1a, NOP proceeds by encoding the sparse context set into a context-conditioned representation $R_C(x)$ evaluated at query locations, combining it with query-side input information to form an initial hidden field $v_0(x)$, and then decoding it with a NO decoder $\mathcal{D}_\theta$ to obtain the final prediction $\hat{u}(x)$. This general pipeline can be instantiated in a deterministic or probabilistic version.

3.2.1 Deterministic Neural Operator Process

The deterministic NOP in Figure 1b defines the core conditional architecture. A key design choice is how sparse context information is represented before decoding. To bridge the set-based view of NPs with the continuous-domain view of NOs, we use a set-to-function lifting step (Gordon et al., 2020). We consider two complementary pathways, which may be used separately or fused before decoding.

SetConv Conditioning. We construct a continuous representation field $R_C(x) = \mathrm{SetConv}_\phi(C; x)$, evaluable at arbitrary query locations $x \in \Omega$. Let w_ϕ be a kernel and $\epsilon > 0$ a stabilizer:

$$R_C(x) = \left[\sum_{i=1}^{n_C} w_\phi(x - x_i^{(c)}), \quad \frac{\sum_{i=1}^{n_C} [a_i^{(c)}, u_i^{(c)}]\, w_\phi(x - x_i^{(c)})}{\sum_{i=1}^{n_C} w_\phi(x - x_i^{(c)}) + \epsilon} \right].$$

This lifts an irregular sparse context set into a continuous field compatible with the query domain for the operator decoder, while separately encoding where context support is concentrated (density) and what local field values are observed (signal). We implement w_ϕ as an RBF kernel. In practice, additional positional or geometric features can be included to better account for domain geometry.

Query-Aligned Attention Conditioning. To capture context–query interactions beyond kernel proximity, we introduce a query-aligned attention pathway that models which context points are most relevant to each query (Kim et al., 2019; Lee et al., 2019). Let $h_i^{(c)} = \phi_{\mathrm{ctx}}(x_i^{(c)}, a_i^{(c)}, u_i^{(c)})$ denote a learned context embedding and $q(x) = \phi_{\mathrm{qry}}(x, a(x))$ a query embedding:

$$r_{\mathrm{att}}(x) = \sum_{i=1}^{n_C} \alpha_i(x)\, h_i^{(c)}, \qquad \alpha_i(x) = \frac{\exp(\omega_\theta(h_i^{(c)}, q(x)))}{\sum_{j=1}^{n_C} \exp(\omega_\theta(h_j^{(c)}, q(x)))},$$

where ω_θ is a learned compatibility function. This yields a target-specific context summary that preserves local geometric structure.

Fused Local Representation. The two pathways can be combined into a fused representation $r_{\mathrm{fuse}}(x) = \phi_{\mathrm{fuse}}(R_C(x), r_{\mathrm{att}}(x))$, implemented in this work as a pointwise MLP on concatenated inputs. This is combined with query-side information to form the initial hidden field $v_0(x) = P_\theta\big([x,\ a(x),\ r_{\mathrm{fuse}}(x),\ b(x)]\big)$, where P_θ is a pointwise projection and $b(x)$ denotes optional positional features (e.g., boundary-distance encodings). Both conditioning pathways cost $O(n_C n_Q)$ for context–query interactions, while the subsequent L-layer FNO decoder scales as $O(L n_Q \log n_Q)$, up to memory and channel constants, with $n_C \ll n_Q$ in our sparse regime.

Operator Decoder. The resulting hidden field is passed through an NO decoder; we use an FNO-style decoder. After L operator layers, a pointwise output head maps the final representation to the

prediction, yielding $\hat{u}(x) = \phi_{\text{out}}(v_L(x))$. To encourage accurate prediction of the full query field while remaining robust to scale differences across tasks, we train using a relative L^2 loss.

3.3 Probabilistic Neural Operator Process

As depicted in Figure 1c, we extend the deterministic architecture with a latent-variable pathway to learn the predictive distribution $p_\theta(U_Q \mid Q, C)$.

Latent Prior and Posterior. Let $z \in \mathbb{R}^{d_z}$ be a global latent variable capturing uncertainty shared across the response field. We infer z from a global summary of the context-induced representation, obtained by permutation-invariant aggregation over the structured query grid Q_{ref}, typically corresponding to the decoder query grid, $r_C = \text{Agg}(\{R_C(x)\}_{x \in Q_{\text{ref}}})$. The context-conditioned prior is $p_\theta(z \mid C) = \mathcal{N}(z; \mu_p(r_C), \text{diag}(\sigma_p^2(r_C)))$. During training, an approximate posterior $q_\phi(z \mid C, Q, U_Q)$ additionally conditions on the supervised query field, using the same set-to-function construction to obtain a target-informed representation r_T.

Latent Conditioning and Predictive Distribution. The latent z modulates the deterministic backbone via broadcast concatenation with $v_0(x)$. Conditioned on z, Q, and C, a probabilistic head outputs a predictive mean and conditional variance at each query location, $\mu_u(x_j^{(q)}; z, Q, C)$ and $\sigma_u^2(x_j^{(q)}; z, Q, C)$. The predictive distribution is obtained by marginalizing the conditionally factorized likelihood:

$$p_\theta(U_Q \mid Q, C) = \int \left[\prod_{j=1}^{n_Q} \mathcal{N}\Big(u_j^{(q)}; \mu_u(x_j^{(q)}; z, Q, C), \sigma_u^2(x_j^{(q)}; z, Q, C)\Big) \right] p_\theta(z \mid C)\, dz.$$

Target outputs are conditionally independent given z, Q, and C, while marginalizing the shared latent variable generally induces dependence. Predictive uncertainty thus combines a global latent component from variation across latent draws with a local conditional component from the output likelihood. Our probabilistic models inject z only through v_0, which yielded more stable prior-conditioned inference than decoder-wide modulation in our experiments (Appendix B.3).

Training and Inference. During training, the target-informed approximate posterior $q_\phi(z \mid C, Q, U_Q)$ is matched to the context-conditioned prior $p_\theta(z \mid C)$:

$$\mathcal{L}_\beta = \mathbb{E}_{q_\phi(z \mid C, Q, U_Q)} \left[\log p_\theta(U_Q \mid z, Q, C)\right] - \beta\, \text{KL}\left(q_\phi(z \mid C, Q, U_Q) \,\|\, p_\theta(z \mid C)\right).$$

At test time, U_Q is unavailable and predictions instead use $z \sim p_\theta(z \mid C)$, with the predictive distribution approximated by Monte Carlo sampling. We use linear β-scheduling during training.

4 Experimental Setup

4.1 Benchmarks and Task Families

We evaluate four benchmark families in a controlled progression to assess probabilistic coherence, sparse context-conditioned PDE prediction, conditioning effectiveness, uncertainty quality without substantial accuracy loss, and sensitivity to different regimes and geometries. We focus on representative regimes, with extended ablations in Appendix examining conditioning mechanisms, observation geometry, uncertainty quantification, distribution shift, resolution transfer, and computational cost.

GP 1D regression. Tasks sampled from GPs with RBF, Matérn-5/2, and mixed kernels. This serves to verify that sparse conditioning, latent inference, and uncertainty quantification behave coherently. GP tasks are exposed through the same NOP interface used for PDEs.

Burgers 1D equation. Viscous Burgers with periodic boundary conditions, mapping initial condition $u_0(x) \mapsto u_T(x)$. Smooth, periodic, and spectrally aligned with FNO.

Darcy 2D flow. Elliptic operator on a non-periodic bounded domain, mapping permeability $a(x)$ to pressure $u(x)$. Boundary-sensitive and geometry-demanding to test local geometry preservation.

Table 1: Deterministic and probabilistic heteroscedastic variants rel-L^2 on GP benchmark with different latent inference modes. Mean values over 4 seeds with standard deviation. Lower is better.

	Deterministic		Prior-mean		Posterior-mean		Zero-latent	
GP	**DNOP**	**DANOP**	**PNOP-HE**	**PANOP-HE**	**PNOP-HE**	**PANOP-HE**	**PNOP-HE**	**PANOP-HE**
RBF	0.248±0.002	0.243±0.002	0.264±0.001	0.265±0.000	0.197±0.001	0.197±0.002	0.543±0.012	0.568±0.013
Mixed	0.352±0.001	0.340±0.001	0.359±0.000	0.363±0.001	0.295±0.001	0.295±0.006	0.648±0.001	0.648±0.008
Matérn	0.411±0.001	0.395±0.001	0.419±0.001	0.420±0.001	0.352±0.001	0.351±0.005	0.710±0.008	0.711±0.020

Table 2: Deterministic models mean rel-L^2 error over 4 seeds across PDEs, summarizing uniform-context, clustered-context, and dedicated paired Coarse-to-Fine evaluations.

	Reference	Uniform Context		Clustered Context		Paired Zero-Shot	
PDE	**FNO**	**DNOP**	**DANOP**	**DNOP**	**DANOP**	**DNOP**	**DANOP**
Burgers	0.0050±0.0001	0.0058±0.0001	0.0047±0.0001	0.0115±0.0003	0.0059±0.0003	0.0057±0.0001	0.0048±0.0001
Darcy	0.0254±0.0003	0.0242±0.0002	0.0226±0.0008	0.0176±0.0013	0.0232±0.0008	0.3316±0.0131	0.3196±0.0167
NS	0.0343±0.0004	0.0293±0.0005	0.0326±0.0016	0.0354±0.0002	0.0336±0.0007	0.2195±0.0033	0.2031±0.0045

Navier–Stokes 2D. Incompressible Navier–Stokes in vorticity form on a periodic domain ($\nu = 10^-3, T = 5.0$), mapping $\omega_0(x) \mapsto \omega_T(x)$. This horizon captures substantial dynamical evolution, with an approximately 20% reduction in field standard deviation and visible structural changes in our trajectories. It provides a spatially complex periodic 2D regime between smooth Burgers and Darcy. Strongly turbulent, lower-viscosity regimes remain outside our scope and are left for future work.

Each task is one episode, with prediction over the full query grid and sparse context sampled from the corresponding response field. At evaluation, NOPs condition on held-out response fields through sparse context without test-time optimization. The task split is 1000/200/200 across all datasets. Context sizes range from $16-64$ points on Burgers ($6.25\%-25\%$ of 256 points), $32-256$ on Darcy ($\approx 0.78\%-6.25\%$ of $64^2 = 4096$ points), and $64-256$ on Navier-Stokes ($\approx 1.56\%-6.25\%$ of 4096 points). Appendix studies additional context-sampling strategies, train-evaluation combinations, and cross-resolution.

4.2 Models

We evaluate four NOP variants. **DNOP** is the deterministic baseline with SetConv conditioning, and **PNOP** its probabilistic extension with input-stage latent injection. **DANOP** combines SetConv and attention conditioning, with **PANOP** as its probabilistic extension. We also compare heteroscedastic (HE) and homoscedastic (HO) output heads. We fix the FNO decoder across the variants to isolate the effects of sparse conditioning and probabilistic inference from changes in the operator backbone.

As a standard NO reference, we adapt an FNO to the same tasks and output targets. It receives the complete input field a_Q and learns $a_Q \mapsto U_Q$ without response context. This provides a conventional operator-learning reference for comparison with sparse-context NOPs. The Appendix includes matched sparse-conditioning controls, architectural variants, and non-spectral NO baselines.

5 Results

NOP Behavior on GP regression. Table 1 shows that both architectures learn the underlying functions, with DANOP consistently outperforming DNOP across GP families, particularly on Mixed and Matérn. Preserving local context–query geometry is already useful in sparse function regression. For probabilistic variants, prior-mean performance remains substantially closer to posterior-mean than zero-latent performance, indicating an active latent pathway and prior–posterior alignment. As shown in Figure 2, uncertainty widens in weakly constrained regions, consistent with coherent probabilistic behavior under sparse conditioning.

Appendix B.5 further show that PNOP-HE and PANOP-HE achieve significantly lower NLL and CRPS than a calibrated deterministic ensemble across all GP families, with comparable point accuracy on RBF and Mixed. These gains primarily reflect improved predictive-distribution quality.

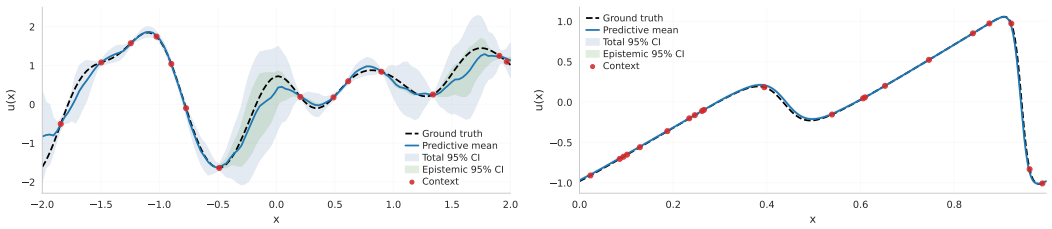


Figure 2: **PNOP-HE behavior in 1D GP regression and Burgers.** On GP-RBF (left), the model tracks the function, with uncertainty increasing in weakly constrained regions. Here, epistemic intervals denote variation in predictive means across latent samples $z \sim p_\theta(z \mid C)$, and total intervals also include the conditional likelihood variance. On Burgers (right), the predictive mean is accurate and intervals are small relative to the solution scale. Figure 3 shows the variance decomposition.

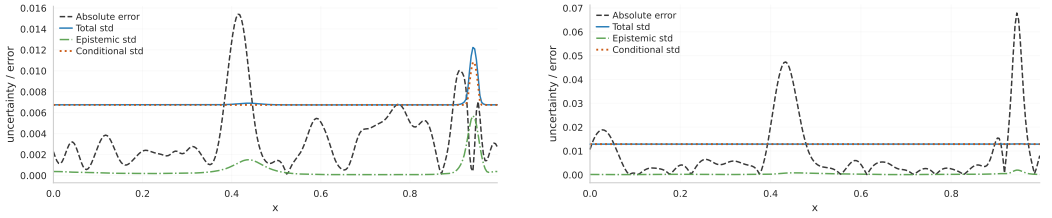


Figure 3: **PANOP variance decomposition in Burgers.** Absolute prediction error and total, epistemic, and conditional predictive standard deviations are shown. The HE head (left) produces more spatially adaptive conditional uncertainty, while HO (right) yields a more uniform profile. Comparing error and uncertainty qualitatively shows whether uncertainty increases in locally difficult regions.

Deterministic Sparse Operator Learning. Table 2 shows that deterministic prediction with sparse response context is effective. Both NOP variants achieve errors comparable to or slightly below the FNO reference. Preserving local context–query through attention often reduces relative L^2 error, with regime-dependent gains. On Burgers, whose periodic dynamics suit the spectral decoder, attention improves accuracy even under standard evaluation. On Navier–Stokes, DNOP remains competitive under uniform evaluation, while attention mainly improves robustness to changes in context sampling strategy and budget. Furthermore, attention is especially valuable when local spatial topology is irregular and task-critical, improving robustness to context size and spatial distribution (Appendix C.1 and C.3). Appendix C.6 shows that the benefits of learned conditioning and decoder choice are PDE-dependent, with no single mechanism dominating across all benchmarks.

Probabilistic Sparse Operator Learning. Table 3 reports fixed-latent evaluations at the context-conditioned prior mean and target-informed posterior mean. Appendix B.2 evaluates the full test-time predictive distribution under prior sampling. Latent conditioning retains good predictive accuracy while providing informative predictive uncertainty. Prior and posterior performance remain close across all PDE families and models, indicating good predictive prior-posterior alignment and limited information loss at test time. Predictive distributions remain informative, with benchmark-dependent calibration, including conservative coverage in some settings and mild under-coverage in Navier–Stokes. Zero-latent evaluation degrades substantially, mainly for PNOP, suggesting that geometry-aware conditioning better supports latent inference. Appendix B.2 also shows that prior-sampled performance remains close to prior-mean for the strongest variants.

On Burgers, probabilistic performance remains close to the deterministic setting, with negligible posterior–prior gaps. Thus, uncertainty can be added with little effect on the predictive mean, making Burgers the closest PDE analogue to the GP diagnostic regime.

Across benchmarks, PANOP is geneally the strongest probabilistic model, but the preferred variance parameterization depends on the benchmark. On Burgers, PANOP-HE performs best overall, while PANOP-HO achieves more conservative empirical coverage at the cost of greater error, worse likelihood, and substantially wider intervals. Figure 3 shows that HO produces more uniform uncertainty, while HE is more expressive and provides useful spatial adaptation near sharper transitions.

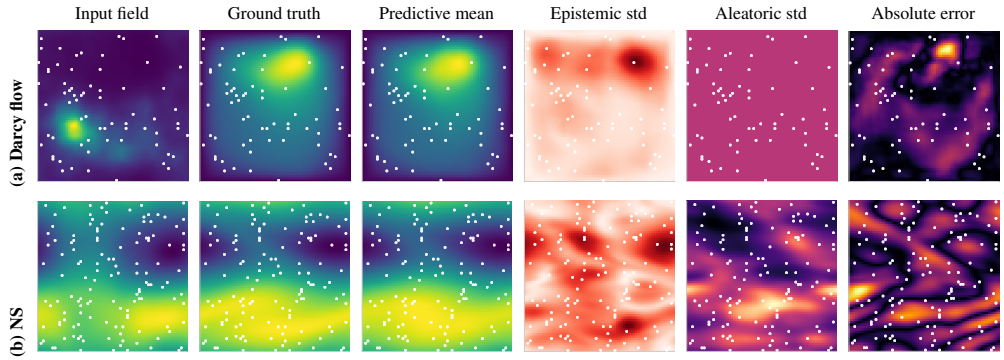


Figure 4: **PANOP-HE behaviour on Darcy and Navier–Stokes.** The panels show ground truth, predictive mean, epistemic and aleatoric uncertainty, and absolute error; white markers indicate context locations. Here, epistemic uncertainty denotes variation across latent samples, while aleatoric std denotes the output head's conditional likelihood std. The model maintains accurate predictive means, with uncertainty concentrating around Darcy's difficult high-response region and showing a broader, more spatially structured pattern in Navier-Stokes. Error-uncertainty correspondence is qualitative; calibration is assessed separately through coverage and likelihood metrics.

On Darcy, variance parameterization plays a particularly important role. PANOP-HO is the strongest probabilistic model under both posterior and prior evaluation. With similar interval widths across models, its improvement reflects better predictive means rather than over-dispersion. Darcy ablations in Appendices C.5 and B.2 suggest that the more constrained HO head can stabilize probabilistic inference. Figure 4 separately illustrates accurate predictive means and uncertainty concentrated in difficult regions rather than broadly diffused across the field.

Navier–Stokes is the most sensitive to variance parameterization. Both architectures deteriorate sharply with HO, consistent with a single global noise scale being too restrictive for this heterogeneous dynamical regime. PANOP-HE performs best under both posterior and prior evaluation, with well-behaved uncertainty under prior conditioning. HO variants produce substantially wider intervals yet perform worse than their HE counterparts. Figure 4 shows a broader and more spatially structured uncertainty pattern in Navier–Stokes than in Darcy, consistent with its more complex dynamics.

As in the deterministic setting, observation geometry has regime-dependent effects, particularly in 2D, where geometry-aware conditioning and diverse training observation patterns improve robustness. Appendix C.4 further shows that sparse cross-resolution transfer remains limited, with meaningful zero-shot transfer on Burgers but greater difficulty in the 2D regimes.

Overall, learned priors are often well aligned with posteriors, and PANOP generally achieves stronger probabilistic performance. Variance parameterization remains task-dependent, with Burgers and Navier–Stokes favoring HE and Darcy benefiting from the more constrained HO parameterization. HE enables spatially adaptive conditional uncertainty but can complicate optimization of the joint probabilistic objective, while HO provides stronger regularization that benefits some regimes.

6 Discussion and Further Work

Using sparse response context, NOPs attain predictive performance comparable to FNO and in some cases improve upon it across the PDE benchmarks considered. The framework provides informative full-field uncertainty estimates, decomposed into a global latent component and local conditional uncertainty from the output likelihood. The small prior-posterior mean gaps in the main evaluations, with the stable prior-sampled behavior for the strongest variants, show that amortized inference can extract instance-specific representations from sparse context that remain effective at test time.

Architectural choices depend on the regime and objective. Smooth periodic regimes are often well served by coarse set-based conditioning, while geometry-sensitive 2D regimes require stronger preservation of local geometry. The preferred conditioning pathway also depends on whether the task involves deterministic reconstruction or uncertainty-aware inference. Variance parameterization is similarly regime-dependent, with some settings benefiting from flexible heteroscedastic heads

Table 3: Probabilistic models under fixed-latent posterior-mean and prior-mean evaluation. Mean over 4 seeds. Lower rel-L^2, MC-NLL, and width are better; coverage is best near the 0.95 level.

		Posterior-mean				Prior-mean			
PDE	**Model**	**Rel-L^2**	**MC-NLL**	**Cov.**	**Width**	**Rel-L^2**	**MC-NLL**	**Cov.**	**Width**
Burgers	PNOP-HE	0.0096 ± 0.0007	-989	0.971	0.025	0.0096 ± 0.0006	-989	0.972	0.025
	PNOP-HO	0.0098 ± 0.0006	-991	0.971	0.025	0.0098 ± 0.0005	-991	0.970	0.025
	PANOP-HE	0.0089 ± 0.0002	-1001	0.974	0.025	0.0089 ± 0.0003	-1001	0.974	0.025
	PANOP-HO	0.0110 ± 0.0006	-862	0.988	0.045	0.0110 ± 0.0006	-862	0.988	0.045
Darcy	PNOP-HE	0.0727 ± 0.0023	-16601	0.996	0.024	0.0730 ± 0.0020	-16599	0.996	0.024
	PNOP-HO	0.0704 ± 0.0033	-16625	0.996	0.024	0.0703 ± 0.0033	-16624	0.997	0.024
	PANOP-HE	0.0712 ± 0.0036	-16617	0.996	0.024	0.0712 ± 0.0036	-16617	0.996	0.024
	PANOP-HO	0.0657 ± 0.0001	-16672	0.998	0.024	0.0657 ± 0.0002	-16672	0.998	0.024
NS	PNOP-HE	0.0511 ± 0.0074	-12865	0.907	0.062	0.0543 ± 0.0118	-12392	0.890	0.062
	PNOP-HO	0.1246 ± 0.0396	-8352	0.892	0.108	0.1258 ± 0.0448	-8058	0.889	0.108
	PANOP-HE	0.0496 ± 0.0048	-13001	0.903	0.058	0.0495 ± 0.0105	-12951	0.902	0.058
	PANOP-HO	0.1180 ± 0.0378	-8508	0.898	0.107	0.1185 ± 0.0381	-8499	0.897	0.107

and others, such as Darcy, from the regularization of more constrained parameterizations. The conditioning and decoder controls in Appendix C.6 further support these conclusions.

Failure Modes and Design Principles. The success of sparse conditional operator learning depends on how geometric information is represented and propagated through the operator decoder. Overcompressing local context into global summaries or strongly modulating it with a global latent can weaken spatial structure. This effect is limited in smooth periodic settings such as Burgers, but more pronounced in boundary-sensitive regimes such as Darcy, where we observe diffuse predictive means and increased uncertainty. The results suggest that an important bottleneck is the stability of the conditioning interface between sparse context, latent modulation, and operator decoding. Geometry-sensitive regimes are particularly vulnerable to blurred local structure and uncertainty mechanisms that absorb prediction difficulty. These distinct bottlenecks call for explicitly preserving local geometry and ensuring that latent mechanisms and variance heads complement the deterministic trunk's spatial inductive biases.

Limitations and future work. Our benchmarks do not cover strongly chaotic dynamics, highly multiscale systems, or irregular geometries. The probabilistic models use a single global Gaussian latent and diagonal likelihoods, limiting localized latent uncertainty and explicit spatial covariance. A trade-off between deterministic accuracy and uncertainty-aware prediction appears, especially in the harder 2D regimes where robust sparse cross-resolution transfer remains challenging. Probabilistic extensions increase training-time memory and runtime cost, although inference remains modest (Appendix C.7). We fix the FNO decoder in the main study to isolate sparse conditioning and probabilistic inference from backbone changes. Thus, Darcy's geometry sensitivity and stronger 2D cross-resolution degradation may partly reflect interactions between sparse conditioning and the spectral decoder's bias toward regular periodic structure. These limitations motivate spatially structured latents, likelihoods with spatial covariance, and alternative decoders suited to non-periodic boundaries and irregular geometries, extending NOPs to richer properties and broader applications.

7 Conclusion

We introduce Neural Operator Processes (NOPs), a framework for operator learning with partial response observations that combines the set-conditioned inference of NPs with the continuous-space mappings of NOs. Across benchmarks, NOPs match or outperform FNO baselines while conditioning on response observations at only 0.78–25% of grid locations. Crucially, their probabilistic extensions provide meaningful uncertainty estimates for full-field predictions. Our results further show that preserving local geometry has regime-dependent benefits and that conditioning and decoder inductive biases contribute differently across operator families. Reliable probabilistic inference also depends on latent conditioning and variance parameterizations that complement the deterministic trunk's spatial inductive biases. Overall, NOPs provide a practical, explicit, and extensible framework for combining sparse conditional inference, structured operator decoding, and uncertainty quantification, identifying key architectural factors governing its performance.

Reproducibility Statement

The NOP architecture, probabilistic formulation, training objectives, and evaluation protocols are specified in the main text and appendices. Reproducibility details, including context budgets and sampling strategies, model and latent configurations, software environment, computational resources, and computational profiling, are documented in Appendix A.1 and C.7. Unless otherwise stated, reported results are aggregated over four independent seeds, with uncertainty across runs provided for the primary comparison metrics. The extended appendices further document the observation-geometry, train–evaluation shift, cross-resolution, latent-inference, conditioning, and decoder ablations used to support the main conclusions. The source code and experiment configurations required to reproduce the reported experiments will be released publicly upon publication.

AI Use Statement

Generative AI tools were used to assist with code debugging, experimental workflow support, literature exploration and related-work positioning, manuscript editing, and LaTeX preparation. All AI-assisted code modifications were reviewed and tested by the authors before being used in reported experiments. AI-assisted text and literature-related outputs were reviewed and verified by the authors, and all reported numerical results, analyses, references, and scientific claims were checked against the corresponding experimental outputs and source materials. The authors take responsibility for the methodology, interpretation of results, and final content of the paper.

REFERENCES

Matthieu Blanke, Ronan Fablet, and Marc Lelarge. Neural incremental data assimilation. *arXiv preprint arXiv:2406.15076*, 2024.

Christopher Bülte, Philipp Scholl, and Gitta Kutyniok. Probabilistic neural operators for functional uncertainty quantification. *arXiv preprint arXiv:2502.12902*, 2025.

Sibo Cheng, Che Liu, Yike Guo, and Rossella Arcucci. Efficient deep data assimilation with sparse observations and time-varying sensors. *arXiv preprint arXiv:2310.16187*, 2023. URL https://arxiv.org/abs/2310.16187.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1126–1135. PMLR, 2017. URL https://proceedings.mlr.press/v70/finn17a.html.

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*, pp. 1704–1713. PMLR, 2018a.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J. Rezende, S. M. Ali Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018b. URL https://arxiv.org/abs/1807.01622.

Jonathan Gordon, Wessel P. Bruinsma, Andrew Y. K. Foong, James Requeima, Yann Dubois, and Richard E. Turner. Convolutional conditional neural processes. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Skey4eBYPS.

Saurav Jha, Dong Gong, Xuesong Wang, Richard E Turner, and Lina Yao. The neural process family: Survey, applications and perspectives. *arXiv preprint arXiv:2209.00517*, 2022.

Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkE6PjC9KX.

Nikola B Kovachki, Samuel Lanthaler, and Andrew M Stuart. Operator learning: Algorithms and analysis. *Handbook of Numerical Analysis*, 25:419–467, 2024.

Jose Lara-Rangel, Nanze Chen, and Fengzhe Zhang. Exploring pseudo-token approaches in transformer neural processes. *arXiv preprint arXiv:2504.14416*, 2025. URL https://arxiv.org/abs/2504.14416.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Graph kernel network for partial differential equations. *arXiv preprint arXiv:2003.03485*, 2020. URL https://arxiv.org/abs/2003.03485.

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. URL https://openreview.net/forum?id=c8P9NQVtmnO.

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. doi: 10.1038/s42256-021-00302-5.

Emilia Magnani, Nicholas Krämer, Runa Eschenhagen, Lorenzo Rosasco, and Philipp Hennig. Approximate bayesian neural operators: Uncertainty quantification for parametric pdes. *arXiv preprint arXiv:2208.01565*, 2022. URL https://arxiv.org/abs/2208.01565.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Omry Yadan. Hydra: A framework for elegantly configuring complex applications. GitHub, 2019. URL https://github.com/facebookresearch/hydra.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/6931-deep-sets.

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

A Overview of Extended Analysis and Ablations

This appendix complements the main text with a broader set of diagnostics, robustness analyses, and practical profiling results. Its purpose is twofold: first, to clarify how the probabilistic latent mechanisms behave beyond the main summary tables; and second, to stress-test the NOP framework under changes in observation geometry, context availability, discretization, architectural sub-components, and computational budget. Some appendix ablations use separately trained checkpoint suites, so matching labels across tables do not always denote repeated evaluations of the same checkpoint. Taken together, these studies are intended to distinguish which empirical gains arise from the general NOP formulation and which depend more specifically on particular conditioning mechanisms or evaluation regimes.

Unless otherwise stated, results are reported as means over four independent seeds. Standard deviations are shown for the primary comparison metrics, especially relative-L^2 errors and the main quantities supporting model comparisons. In wider diagnostic tables, some secondary metrics are reported as means only for readability, using the same set of runs.

For uncertainty evaluation, MC-NLL uses a Monte Carlo predictive mixture under latent sampling and reduces to the corresponding conditional NLL for fixed-latent evaluation modes; coverage is the fraction of targets within the nominal 95% predictive interval, and width is its average size. MC-NLL is normalized by the number of query points only when explicitly reported as average MC-NLL.

For clarity, the appendix is organized into two parts:

Probabilistic diagnostics, auxiliary baselines, and latent analyses (Section B). We begin with extended GP diagnostics, quantitative uncertainty comparisons against deterministic ensembles and an exact-GP oracle reference, latent robustness tests, and the latent-injection ablation. These diagnostics clarify how posterior-based, prior-based, and latent-ablated evaluations should be interpreted, and support the probabilistic design choices beyond the condensed comparisons in the main text. In particular, they separate decoder quality under target-informed latent inference from deployment-relevant prior behavior, and show why stable prior-conditioned inference is more informative than posterior reconstruction alone. We also include a supplementary deterministic comparison with additional DeepONet-based neural operators, considered here as classical branch–trunk baselines, to position NOP relative to another canonical non-spectral operator-learning family with architectural properties different from those of FNO.

Extended architectural and generalization ablations (Section C). The second part collects the broader ablation suite on observation geometry, test-time context scaling, train-evaluation strategy mismatch, cross-resolution transfer, local conditioning decomposition, conditioning and decoder controls, and computational cost. The sequence is structured to move from data-distribution sensitivity, to out-of-distribution robustness and resolution transfer, to increasingly controlled architectural dissection, and finally to practical cost profiling.

Context strategy and context budget ablations (Section C.1). We first establish the in-distribution sensitivity of the models to the geometry and cardinality of the observation set. This provides the empirical baseline for the appendix: before testing transfer or robustness, we characterize how strongly sparse conditional operator learning depends on where observations are placed and how many are available.

Test-time context scaling (Section C.2). We then test whether the limitations identified in the previous section can be alleviated simply by supplying more context at inference time while keeping the checkpoint fixed. This serves as a direct control for the context-budget study, separating representational limitations from the trivial effect of additional observations at deployment.

Train–evaluation strategy ablation (Section C.3). Having characterized in-distribution behavior, we next move to out-of-distribution geometry shifts by evaluating checkpoints trained under one observation strategy and tested under another. This section closes the geometry-focused part of the appendix by showing when robustness comes from architecture, when it comes from training distribution, and how much specialization to sensor layout remains.

Cross-resolution transfer (Section C.4). We then shift from observation-space variation to discretization-space variation, testing whether coarse-trained models transfer zero-shot to finer grids. This section connects sparse response conditioning to resolution transfer and identifies where discretization invariance remains limited.

Local conditioning decomposition (Section C.5). After documenting the main robustness and transfer phenomena, we dissect the local conditioning trunk itself by isolating the SetConv and query-aligned attention pathways. This provides a more mechanistic view of which component contributes most strongly under different PDE families and inference objectives, and helps justify the combined design used in the main-text models.

Conditioning and decoder controls (Section C.6). We then extend the architectural analysis beyond within-family conditioning ablations through two complementary controls. A masked-channel FNO tests whether sparse observations can be injected directly on the query grid without a learned set-conditioning mechanism, including a geometry-matched variant for Darcy. Conversely, a ConvCNP-style control preserves the SetConv conditioning path while replacing the FNO decoder with compact and parameter-matched local convolutional decoders. These experiments separate the roles of sparse conditioning, geometric representation, and decoder inductive bias across PDEs.

Computational cost and complexity (Section C.7). Finally, we profile parameter count, memory footprint, and training/inference cost. This concludes the appendix with a practical perspective, showing the computational trade-offs associated with the geometry-aware and probabilistic extensions.

Scope and architectural limitation. One important scope condition applies across all of these studies: throughout the main experiments we keep the neural operator decoder backbone fixed to an FNO. This allows the ablations to isolate the effects of conditional representation, latent modeling, and geometry-aware conditioning without conflating them with backbone changes. At the same time, it also limits the scope of the conclusions. In particular, some of the weaknesses observed in the more difficult transfer settings, especially in non-periodic or dynamically complex 2D regimes, may reflect not only limitations of the sparse conditional representation but also the interaction between that representation and the spectral FNO decoder. The appendix therefore should not be read as claiming that all remaining failure modes are intrinsic to the NOP framework itself. Rather, the results establish that the NOP framework yields clear gains in robustness, probabilistic stability, and geometry-aware conditioning under a fixed and widely used operator-decoder family, while also identifying decoder-sensitive regimes that remain promising directions for future extensions. Task-matched comparisons with probabilistic neural-operator baselines remain future work, as many require adapting dense-input uncertainty estimators to sparse response-conditioned inference.

A.1 Reproducibility and Ablation Setup

Compute Environment All models and dataloaders are implemented in Python 3.11.7 using PyTorch 2.5.1 with CUDA 12.7 (Paszke et al., 2019). Experiment configuration is managed with Hydra and OmegaConf(Yadan, 2019). The neural-operator decoder uses an FNO backbone based on the Fourier Neural Operator architecture of Li et al. (2021), following the publicly available Fourier Neural Operator implementation. Experiments were run on a single NVIDIA GH200 GPU node with 96 GB of GPU memory, an ARM Neoverse-V2 CPU with 72 cores (up to 3.47 GHz), and 477 GB of system RAM, under Linux (aarch64). Computational profiling results are reported in Section C.7. All datasets were generated by the authors. External software dependencies and referenced public codebases are used under their respective licenses; the primary assets include PyTorch (BSD-3-Clause), Hydra (MIT), OmegaConf (BSD-3-Clause), and the referenced FNO codebase (MIT).

Context Budgets. For some ablations, models are evaluated across three context budget regimes: low, base, and high. The absolute number of context points defining these regimes varies by benchmark to account for differences in domain dimensionality. Specifically, the low, base, and high budgets correspond to 8–16, 16–64, and 64–128 points for Burgers 1D; 16–32, 32–256, and 256–512 points for Darcy 2D; and 32–64, 64–256, and 256–512 points for Navier–Stokes 2D.

Table 4: Homoscedastic variants rel-L^2 on GP benchmark with different latent inference modes. Mean values over 4 seeds. Lower is better.

	Prior-mean		Posterior-mean		Zero-latent	
GP	**PNOP-HO**	**PANOP-HO**	**PNOP-HO**	**PANOP-HO**	**PNOP-HO**	**PANOP-HO**
RBF	0.301±0.001	0.331±0.026	0.221±0.001	0.145±0.064	0.735±0.016	1.014±0.230
Mixed	0.393±0.001	0.427±0.015	0.322±0.000	0.225±0.044	0.769±0.013	1.041±0.145
Matérn	0.444±0.001	0.465±0.019	0.375±0.001	0.285±0.078	0.745±0.007	0.977±0.193

Context sampling strategies. Uniform contexts sample indices uniformly without replacement. Clustered contexts draw a random center in the domain and select the n_C nearest grid points, using normalized absolute distance in 1D and normalized elliptical distance in 2D, with each coordinate normalized by 20% of its domain range. Mixed contexts use clustered sampling with probability 0.5 and uniform sampling otherwise.

Model Configuration. Model configurations are kept largely consistent across experiments, using hidden dimension 64, two-layer input/output projections, four FNO layers, coordinate features, and forward FFT normalization. Fourier modes are set to 24 in 1D and 16 per spatial dimension in 2D. Probabilistic variants use a 16-dimensional latent variable, two-layer latent encoders with hidden dimension 128, linear β-warmup to 1 over 50 epochs, and 8 test-time Monte Carlo samples; attention-based variants use four attention heads. Dataset-specific changes are mainly limited to boundary features and SetConv lengthscales. Additional configuration and hyperparameter details can be found in the accompanying code.

B Probabilistic diagnostics and latent analyses

B.1 Additional GP Diagnostics

We provide additional GP diagnostics to disentangle the deterministic sparse reconstruction quality, posterior-informed decoder quality, and deployment-relevant prior-based probabilistic behavior.

Latent behavior and variance parameterization. As shown in Table 4, the HO variants follow the same basic pattern in the GPs as the HE results in Table 1. Posterior-mean evaluation is much stronger than zero-latent evaluation, confirming that the latent pathway is active, while prior-mean remains reasonably close to posterior-mean in the stable regimes. At the same time, PANOP-HO configuration exhibits a larger prior–posterior gap and substantially higher variability across seeds than the HE variants in Table 1, despite occasionally very low posterior-mean errors.

Figure 5 illustrates the transition from deterministic to probabilistic prediction on GP-Matérn. DNOP reconstructs the function well near observed regions, while the error increases across longer context gaps and around sharper local variations. The probabilistic extension (PNOP) preserves a similar predictive mean in well constrained regions, but in sparsely observed intervals the mean becomes smoother than the ground truth and the predictive uncertainty expands substantially.

Figure 6 shows that with an HO head the conditional term remains nearly constant and the spatial variation of total uncertainty is driven mainly by the epistemic component, yielding a cleaner and more regular uncertainty profile. With an HE head, the uncertainty becomes more spatially adaptive, with a larger share of predictive spread assigned to the conditional variance term. In both cases, uncertainty is largest in the least constrained regions, and the Matérn example makes visible that the main source of ambiguity comes from sparse conditioning, while the choice of variance head changes how that ambiguity is distributed across the domain.

Posterior-informed decoder quality. Considering Tables 1 and 4, all probabilistic variants improve the posterior-informed decoder quality over the deterministic baselines in this regime, with gains that are most pronounced on RBF and more moderate on Matérn-5/2. Once the latent is well informed by the target, the decoder has sufficient capacity to reconstruct the function accurately. At the same time, PANOP-HO remains stronger than PNOP-HO, reflecting not only latent inference quality

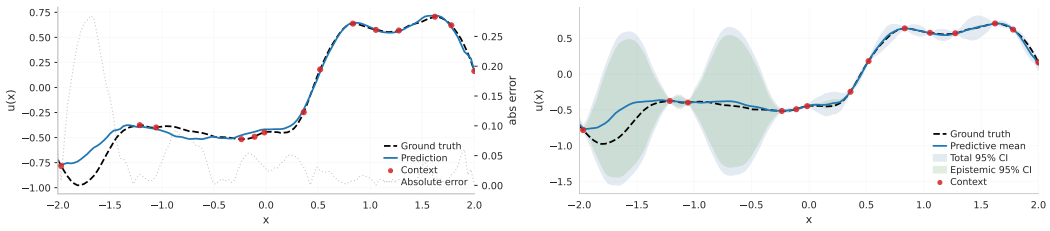


Figure 5: **Deterministic NOP and probabilistic PNOP behavior on 1D GP-Matérn**. Left: deterministic prediction from sparse context observations, with larger errors across wider context gaps and sharper local variations. Right: probabilistic prediction with a HE head. The predictive mean remains accurate near context points, while uncertainty expands in underconstrained intervals. The epistemic band is computed from variance across latent samples from the learned conditional prior $p_\theta(z \mid C)$, and the total band additionally includes the predicted conditional likelihood variance.

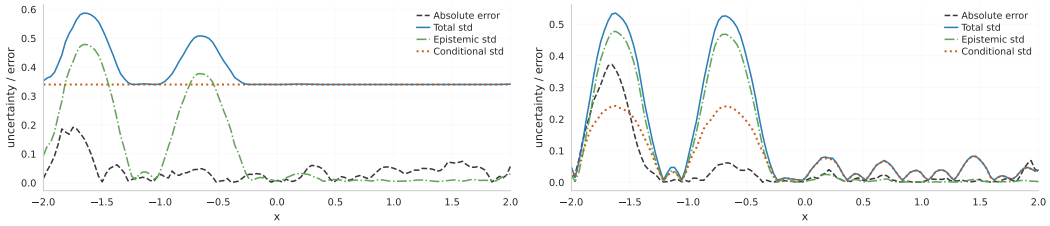


Figure 6: **PNOP variance decomposition for the GP-Matérn example in Figure 5.** In the HO case (left), the conditional standard deviation remains nearly constant and the total uncertainty is driven mainly by the epistemic component. In the HE case (right), both epistemic and conditional terms vary across the domain, yielding a more spatially adaptive uncertainty profile. In both parameterizations, uncertainty is largest in the least constrained regions.

and robustness but also how effectively each architecture supports the interaction between latent conditioning and decoder reconstruction.

Prior-based predictive behavior. Table 5 reports predictive quality and uncertainty metrics under prior sampling. Across the stable variants, relative-L^2 and interval width increase from RBF to Mixed to Matérn, while coverage remains broadly near nominal. This is consistent with the increasing roughness of the task distribution. The learned priors remain well behaved at test time and capture global uncertainty induced by sparse conditioning, yielding meaningful stochastic predictions rather than pathological or trivially over-dispersed behavior.

B.2 Latent Probabilistic Robustness

Tables 6 and 7 report two complementary probabilistic robustness analyses. Prior-sample evaluation probes the stochastic behavior of the learned prior more directly than prior-mean, while zero-latent evaluation acts as a stress test for how much useful information is carried by the latent pathway.

The prior-sample results in Table 6 broadly reinforce the conclusions from the main probabilistic results. In particular, PANOP remains very stable under stochastic prior draws. Across all PDE benchmarks, PANOP-HE and PANOP-HO are essentially unchanged relative to prior-mean evaluation, indicating that the learned prior is not only well aligned on average, but also stable under sampling. This is especially clear for PANOP-HE, whose prior-sample results remain nearly identical to its posterior-mean and prior-mean performance.

By contrast, PNOP shows mild to moderate sensitivity to prior sampling, which is mostly visible in Burgers under the HO parameterization. There, PNOP-HO exhibits a noticeable degradation together with a large increase in interval width and seed variability. Since the same configuration is much more stable under posterior-mean and prior-mean evaluation, this suggests that the issue is not simply poor mean prediction, but rather instability in the stochastic behavior of the learned prior under sampling. Outside of this case, PNOP remains reasonably stable, but still less consistent than PANOP.

Table 5: GP prior-sampled predictive performance. Lower is better for rel-L^2, MC-NLL, and width. Coverage is best when close to nominal 0.95 level. Mean values over 4 seeds.

GP	Model	Rel-L^2	MC-NLL	Coverage	Avg. width
RBF	PNOP-HE	0.278±0.002	−238.02	0.920	0.699
	PNOP-HO	0.301±0.001	−237.91	0.920	0.698
	PANOP-HE	0.283±0.004	−237.25	0.921	0.709
	PANOP-HO	0.366±0.047	460.86	0.931	1.070
Mixed	PNOP-HE	0.375±0.002	−130.50	0.897	0.966
	PNOP-HO	0.393±0.001	−130.44	0.902	0.983
	PANOP-HE	0.379±0.002	−130.65	0.902	0.993
	PANOP-HO	0.457±0.022	183.37	0.920	1.391
Matérn	PNOP-HE	0.432±0.002	−80.42	0.899	1.162
	PNOP-HO	0.444±0.001	−79.98	0.896	1.154
	PANOP-HE	0.433±0.002	−81.27	0.899	1.165
	PANOP-HO	0.491±0.033	141.78	0.913	1.527

Table 6: Probabilistic models under prior-sample evaluation. Mean values over 4 seeds. Lower rel-L^2 and MC-NLL are better.

Benchmark	Model	Rel-L^2	MC-NLL	Coverage	Width
Burgers	PNOP-HE	0.0097±0.0007	−997.4	0.971	0.025
	PNOP-HO	0.0403±0.0522	−995.0	0.974	0.166
	PANOP-HE	0.0088±0.0002	−1002.3	0.974	0.025
	PANOP-HO	0.0110±0.0005	−861.8	0.988	0.045
Darcy	PNOP-HE	0.0732±0.0018	−16606.2	0.996	0.024
	PNOP-HO	0.0705±0.0033	−16627.3	0.997	0.024
	PANOP-HE	0.0712±0.0037	−16619.3	0.996	0.024
	PANOP-HO	0.0657±0.0001	−16671.9	0.998	0.024
NS	PNOP-HE	0.0551±0.0028	−12639.7	0.904	0.066
	PNOP-HO	0.1260±0.0122	−8034.1	0.894	0.110
	PANOP-HE	0.0496±0.0024	−12800.9	0.902	0.058
	PANOP-HO	0.1187±0.0157	−8261.7	0.899	0.107

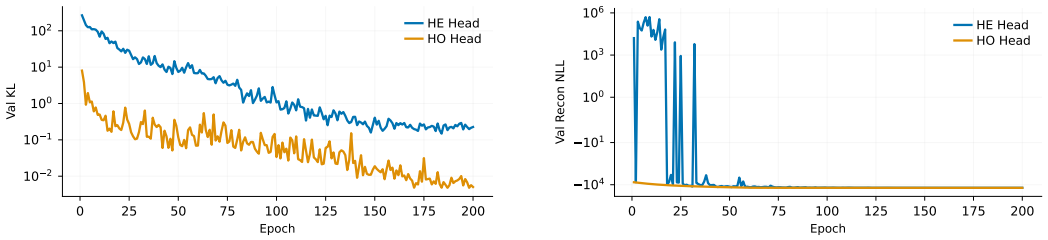
The zero-latent results in Table 7 show that all models degrade, often sharply, indicating that the latent pathway carries useful predictive information. On Burgers, removing the latent causes a severe collapse for PNOP, whereas PANOP degrades much less and remains far more accurate. The same qualitative trend appears on Navier–Stokes, where PANOP again deteriorates less severely than PNOP, especially in the HE setting. Darcy is less extreme, but the same ordering remains visible. Zero-latent performance is worse than posterior- and prior-based inference for all models, yet PANOP still retains a clear advantage over PNOP. Thus, the architecture is not merely achieving stronger results through a fragile latent dependence, but combining a more robust deterministic structure with a better behaved latent mechanism.

Furthermore, the robustness of the latent pathway is tightly coupled to the choice of variance parameterization. The training dynamics on Darcy in Figure 7 illustrates how the HO model enters a low-KL regime quickly with smooth optimization, effectively regularizing the latent space. In contrast, the HE model shows erratic early-phase KL spikes before stabilizing, reflecting the optimization difficulty of jointly learning a flexible local variance and a global latent prior in geometry-sensitive regimes. Thus, obtaining spatially adaptive predictive uncertainty introduces an additional optimization challenge, and its effectiveness depends in part on the stability of the underlying conditional representation.

These analyses support two additional conclusions. First, PANOP exhibits more stable prior behavior under sampling, supporting stronger alignment with test-time prior inference. Second, although all models rely on latent conditioning to some extent, PANOP degrades more gracefully when the latent is removed, reflecting a more stable interaction between latent information and the underlying NO decoder.

Table 7: Probabilistic models under zero-latent evaluation. Mean values over 4 seeds. This is a stress test of latent dependence rather than a standard inference mode. Low rel-L^2 and MC-NLL are better.

Benchmark	Model	Rel-L^2	MC-NLL	Coverage	Width
Burgers	PNOP-HE	0.2158±0.0237	37221.0	0.101	0.025
	PNOP-HO	0.1913±0.0767	26157.3	0.142	0.027
	PANOP-HE	0.0194±0.0166	−686.2	0.860	0.025
	PANOP-HO	0.0226±0.0120	−701.0	0.915	0.045
Darcy	PNOP-HE	0.1268±0.0342	−15764.2	0.949	0.024
	PNOP-HO	0.1077±0.0298	−16086.9	0.968	0.024
	PANOP-HE	0.0865±0.0195	−16424.3	0.987	0.024
	PANOP-HO	0.0785±0.0108	−16532.4	0.993	0.024
NS	PNOP-HE	0.4459±0.0416	252007.2	0.191	0.051
	PNOP-HO	0.5350±0.0643	115069.3	0.409	0.108
	PANOP-HE	0.1900±0.0812	53719.4	0.400	0.048
	PANOP-HO	0.2696±0.0888	18053.3	0.632	0.107



(a) KL Divergence

(b) Reconstruction NLL

Figure 7: **Training dynamics of PANOP on Darcy.** The homoscedastic model maintains stable low-KL training; the heteroscedastic model exhibits early instability.

B.3 Latent Injection Mechanism Ablation

We examine the effect of how the global latent variable is injected into the probabilistic operator decoder. To this end, we compare PNOP, where the latent is injected once at the input stage before operator propagation, to a model variant named PNOP-m in which the latent is injected through decoder-wide FiLM modulation, so that it modulates each FNO block via feature-wise affine transformations. This ablation isolates the latent-to-decoder interface while keeping the set encoder, the global latent prior/posterior mechanism, and the Gaussian output head otherwise comparable.

Table 8 reports this comparison on Darcy, where the effect of the injection mechanism is clearer. The two variants are relatively close under posterior-mean evaluation, with PNOP-m slightly better in both variance settings. However, this advantage does not transfer to prior-conditioned inference. For PNOP-m the transition from posterior-mean to prior-mean leads to a substantial degradation in error and likelihood, with a further loss under prior sampling, especially in the HO case. This posterior test confirms that the decoder retains sufficient capacity to reconstruct the solution; the primary bottleneck is the prior inference. By contrast, PNOP remains highly stable across posterior-mean, prior-mean, and prior sample, with only negligible changes in all reported metrics. The posterior-prior gap is therefore much larger for the injection style in PNOP-m than for PNOP.

Similarly, removing the latent causes a severe degradation for PNOP-m, together with a marked increase in interval width, whereas PNOP deteriorates much less sharply. Indeed, the FiLM-based variant relies on a latent interaction that is less robust when prior conditioning is used directly or when latent information is weakened.

These results are consistent with stronger coupling between the global latent and iterative operator propagation under decoder-wide FiLM modulation. This may contribute to the poorer prior-conditioned generalization observed on Darcy. In contrast, input-stage injection appears to provide a cleaner interface between global stochastic conditioning and the operator decoder, yielding substantially more stable inference across evaluation modes.

Table 8: Latent injection ablation on Darcy for the probabilistic models. Mean values over 4 seeds. PNOP-m injects the global latent through decoder-wide FiLM modulation at each FNO block, whereas PNOP injects the latent once at the input stage before operator propagation.

Model	Eval Mode	Rel-L^2	MC-NLL	Coverage	Width
PNOP-m-HE	Posterior-mean	0.0677±0.0006	−16652.3	0.997	0.024
	Prior-mean	0.1295±0.0277	−15472.4	0.934	0.025
	Prior-sample	0.1571±0.0162	−16571.7	0.989	0.034
	Zero-latent	1.1086±0.4235	−6456.0	0.989	0.277
PNOP-m-HO	Posterior-mean	0.0700±0.0001	−16630.0	0.997	0.024
	Prior-mean	0.1855±0.0936	−14882.7	0.936	0.031
	Prior-sample	0.2327±0.1310	−16445.0	0.989	0.047
	Zero-latent	1.9424±0.3108	−3564.4	0.990	0.591
PNOP-HE	Posterior-mean	0.0727±0.0023	−16601.2	0.996	0.024
	Prior-mean	0.0730±0.0020	−16598.6	0.996	0.024
	Prior-sample	0.0732±0.0018	−16606.2	0.996	0.024
	Zero-latent	0.1268±0.0342	−15764.2	0.949	0.024
PNOP-HO	Posterior-mean	0.0704±0.0033	−16624.5	0.996	0.024
	Prior-mean	0.0703±0.0033	−16624.4	0.997	0.024
	Prior-sample	0.0705±0.0033	−16627.3	0.997	0.024
	Zero-latent	0.1077±0.0298	−16086.9	0.968	0.024

Table 9: Deterministic rel-L^2 error across PDE benchmarks for NOP and DeepONet variants. Mean values over 4 seeds.

	Uniform Sparse Context		Dense Grid References		
PDE	DNOP	DANOP	DeepONet-Flat	DeepONet-Attn	DeepONet-FiLM
Burgers	0.0058±0.0001	0.0047±0.0001	0.1959±0.0015	0.3339±0.0363	0.1025±0.0085
Darcy	0.0242±0.0002	0.0226±0.0008	0.1275±0.0045	0.1165±0.0053	0.0896±0.0017
NS	0.0293±0.0005	0.0326±0.0016	0.1364±0.0011	0.2087±0.0437	0.1121±0.0081

B.4 Supplementary Comparisons with Branch–Trunk Operator Parameterizations

As a supplementary dense-input reference for architectural context, we additionally evaluate three instantiations of the classical DeepONet branch–trunk parameterization (Lu et al., 2021). This comparison is not used to establish the main claims, but rather to place the proposed framework relative to a canonical non-spectral neural-operator family based on branch–trunk composition.

We keep the variants close to a standard DeepONet formulation while probing a small number of modest extensions. DeepONet-Flat follows the classical dense-input formulation, where the discretized input field is flattened and processed by a standard MLP. DeepONet-Attention uses learned query tokens that attend to the input point–value pairs through multi-head cross-attention (Vaswani et al., 2017) without changing the classical branch–trunk factorization. DeepONet-FiLM conditions the trunk through feature-wise linear modulation generated from the branch representation, allowing the basis functions to adapt while preserving the overall branch–trunk decomposition.

Table 9 shows a clear separation between the sparse conditional NOP models and the additional DeepONet references. Across all three PDE benchmarks, DNOP and DANOP remain substantially more accurate when using sparse response context under the standard uniform-context setting. The relative ordering within the DeepONet family varies across benchmarks, but all three variants remain clearly weaker. Indeed, in the present setting, the classical branch–trunk parameterization remains less effective than the NOP models operating under sparse context.

Table 10: Quantitative uncertainty evaluation on the GP benchmarks. The DNOP ensemble estimates uncertainty from independently trained deterministic models; Ens.+Val additionally includes a scalar residual variance fitted only on validation data. Exact GP uses the true task-specific generative kernel and is therefore an oracle reference. NLL is normalized per target. Lower is better for relative-L^2, NLL, and CRPS; coverage is best near the nominal 0.95 level, and higher is better for the variance–error rank correlation ρ.

GP	Method	Rel-L^2 ↓	NLL/target ↓	CRPS ↓	Cov.$_{95}$	$\rho_{\mathrm{var,error}}$ ↑
RBF	DNOP Ens.	0.2689	104.2300	0.1494	0.5179	0.6299
	DNOP Ens.+Val	0.2689	0.3760	0.1563	0.9133	0.6299
	PNOP-HE	0.2801	-1.6277	0.1151	0.9587	0.8892
	PANOP-HE	0.2791	-1.6220	0.1159	0.9611	0.8694
	Exact GP	0.2050	-2.6273	0.0822	0.8984	0.8910
Matérn-$5/2$	DNOP Ens.	0.4142	400.2413	0.2571	0.3350	0.2628
	DNOP Ens.+Val	0.4142	0.6389	0.2363	0.9136	0.2628
	PNOP-HE	0.4339	-0.3074	0.1994	0.9546	0.7689
	PANOP-HE	0.4301	-0.3206	0.1993	0.9535	0.7577
	Exact GP	0.4018	-0.5470	0.1799	0.9479	0.8310
Mixed	DNOP Ens.	0.3806	323.9728	0.2142	0.4107	0.3917
	DNOP Ens.+Val	0.3806	0.5228	0.2064	0.9353	0.3917
	PNOP-HE	0.3880	-0.4156	0.1713	0.9157	0.7779
	PANOP-HE	0.3855	-0.5103	0.1721	0.9253	0.7718
	Exact GP	0.3433	-1.1923	0.1408	0.9461	0.8738

B.5 Quantitative UQ Baselines and Exact-GP Reference

To complement the qualitative GP uncertainty diagnostics, we compare the probabilistic NOP variants against two explicit uncertainty references. First, we construct a deterministic DNOP ensemble from independently trained checkpoints and estimate epistemic uncertainty from the dispersion of their predictions. Since this ensemble alone captures only between-model variation, we additionally consider a stronger calibrated variant in which a single homoscedastic residual variance is fitted by minimizing NLL on the validation split. Second, we evaluate the exact GP posterior using the true task-specific generative kernel and hyperparameters, serving as an oracle reference.

All methods are evaluated using the shared 128-point grid and sparse noiseless contexts of 3–25 observations from the main GP benchmark, on the same frozen set of 64 held-out tasks for each of the RBF, Matérn-$5/2$, and Mixed GP families. The deterministic uncertainty baseline consists of a five-member DNOP ensemble. Because the number of target points varies across tasks, likelihood comparisons use NLL normalized per target rather than the raw task-summed NLL. Statistical comparisons use paired task-level bootstrap intervals with 1,000 resamples.

The uncalibrated DNOP ensemble is strongly underdispersed, with empirical 95% coverage ranging only from approximately 0.33 to 0.52 across the three GP families. Adding the validation-fitted residual variance substantially improves coverage to approximately 0.91–0.94, providing a considerably stronger deterministic uncertainty baseline. Nevertheless, PNOP-HE and PANOP-HE achieve significantly lower per-target NLL and CRPS than this calibrated ensemble on all three GP families, with the paired 95% bootstrap intervals for these differences excluding zero.

The improvement does not arise from uniformly superior predictive means. Relative-L^2 differences are not statistically distinguishable on RBF and Mixed, while the calibrated ensemble retains a modest but significant advantage on Matérn-$5/2$. The probabilistic NOP variants instead show substantially stronger alignment between predictive variance and realized error, particularly on Matérn-$5/2$ and Mixed. Since adding a scalar residual variance changes the scale but not the spatial ranking of ensemble uncertainty, this distinction indicates that the learned probabilistic models capture more task and location dependent predictive uncertainty.

The exact GP posterior remains the strongest overall reference in relative-L^2, NLL, and CRPS, as expected because it is evaluated with access to the true generative kernel. PNOP-HE and PANOP-HE

nevertheless reduce the predictive distribution gap to this oracle substantially relative to the calibrated deterministic ensemble. Thus, the probabilistic extension primarily improves predictive-distribution quality under sparse conditioning, while point-prediction gains are not universal.

C Extended architectural and generalization ablations

C.1 Context Strategy and Context Budget Ablations

Table 11: Deterministic context strategy and context budget ablation. Mean rel-L^2 error over 4 seeds, specific for this ablation suite.

PDE	Budget	Model	Uniform	Clustered	Mixed
Burgers	Low	DNOP	0.0054±0.0004	0.0106±0.0006	0.0067±0.0002
		DANOP	0.0047±0.0003	0.0057±0.0002	0.0047±0.0002
	Base	DNOP	0.0060±0.0003	0.0115±0.0003	0.0067±0.0004
		DANOP	0.0048±0.0002	0.0059±0.0003	0.0049±0.0001
	High	DNOP	0.0061±0.0005	0.0118±0.0002	0.0067±0.0002
		DANOP	0.0047±0.0002	0.0061±0.0002	0.0048±0.0002
Darcy	Low	DNOP	0.0242±0.0009	0.0217±0.0005	0.0199±0.0008
		DANOP	0.0231±0.0006	0.0233±0.0011	0.0234±0.0007
	Base	DNOP	0.0236±0.0003	0.0176±0.0013	0.0170±0.0010
		DANOP	0.0222±0.0009	0.0232±0.0008	0.0230±0.0004
	High	DNOP	0.0215±0.0004	0.0159±0.0003	0.0169±0.0006
		DANOP	0.0227±0.0005	0.0230±0.0006	0.0228±0.0009
NS	Low	DNOP	0.0324±0.0006	0.0347±0.0007	0.0353±0.0004
		DANOP	0.0340±0.0004	0.0336±0.0009	0.0335±0.0008
	Base	DNOP	0.0293±0.0006	0.0354±0.0002	0.0343±0.0004
		DANOP	0.0322±0.0003	0.0336±0.0007	0.0337±0.0003
	High	DNOP	0.0252±0.0006	0.0335±0.0006	0.0332±0.0002
		DANOP	0.0295±0.0008	0.0339±0.0008	0.0327±0.0007

To isolate the effects of architectural choices from the intrinsic properties of the observation distributions, we evaluate the NOP family across varying context sampling strategies and additional context budgets. This allows us to characterize how conditional operator learning under sparse partial observations depends on both the geometry and the amount of available context.

Deterministic Analysis and the Stabilizing Role of Attention. Table 11 summarizes the deterministic performance across benchmarks. We observe a geometric sensitivity that varies with the underlying physical problem. In Burgers 1D, the models exhibit a weak dependence on the context budget and consistently favor uniform coverage. The dynamics in Navier–Stokes reflect a similar preference for uniform distributions; increasing the context budget under non-uniform strategies (clustered or mixed) yields only limited benefits, suggesting that information propagation in this periodic domain relies more on broad spatial coverage than on local observation density. This is consistent with the interpretation that when the underlying PDE is smooth, periodic, and spectrally aligned with the FNO decoder, global observation coverage is sufficient to construct an accurate continuous representation.

Conversely, Darcy 2D exposes a much stronger dependence on observation topology. The deterministic baseline, DNOP, degrades under uniform sampling but improves under clustered or mixed observations. Increasing the context budget helps within each strategy but does not eliminate the performance gap induced by the sampling pattern. Thus, both context amount and geometry matter in Darcy, with sampling strategy having the larger effect.

The introduction of the query-aligned attention pathway in DANOP noticeably flattens the performance variation across strategies and budgets. In Darcy, DANOP mitigates much of the fragility

Table 12: Probabilistic context strategy ablation under prior-sample evaluation at the base context budget. Mean values specific for this ablation suite over 4 seeds.

PDE	Strategy	Model	Rel-L^2	MC-NLL	Coverage	Width
Burgers	Uniform	PNOP-HE	0.0145±0.0087	−1001.6	0.9745	0.0472
		PNOP-HO	0.0113±0.0004	−859.4	0.9869	0.0451
		PANOP-HE	0.0092±0.0005	−995.2	0.9720	0.0247
		PANOP-HO	0.0112±0.0001	−857.1	0.9873	0.0460
	Clustered	PNOP-HE	0.0098±0.0005	−994.7	0.9695	0.0248
		PNOP-HO	0.0153±0.0001	−766.3	0.9907	0.0674
		PANOP-HE	0.0089±0.0003	−998.0	0.9746	0.0248
		PANOP-HO	0.0167±0.0006	−739.3	0.9921	0.0755
	Mixed	PNOP-HE	0.0092±0.0001	−999.2	0.9730	0.0247
		PNOP-HO	0.0134±0.0001	−809.5	0.9889	0.0561
		PANOP-HE	0.0086±0.0003	−999.3	0.9736	0.0245
		PANOP-HO	0.0145±0.0002	−777.0	0.9917	0.0650
Darcy	Uniform	PNOP-HE	0.0754±0.0029	−16579.7	0.9948	0.0239
		PNOP-HO	0.0660±0.0003	−16669.4	0.9978	0.0238
		PANOP-HE	0.0754±0.0072	−16621.9	0.9964	0.0252
		PANOP-HO	0.0656±0.0001	−16672.1	0.9978	0.0238
	Clustered	PNOP-HE	0.0711±0.0045	−16654.7	0.9971	0.0242
		PNOP-HO	0.0640±0.0005	−16687.2	0.9982	0.0238
		PANOP-HE	0.0674±0.0002	−16656.4	0.9976	0.0238
		PANOP-HO	0.0654±0.0012	−16673.6	0.9979	0.0238
	Mixed	PNOP-HE	0.0696±0.0011	−16641.0	0.9969	0.0239
		PNOP-HO	0.0641±0.0003	−16686.4	0.9982	0.0238
		PANOP-HE	0.0671±0.0011	−16660.5	0.9976	0.0239
		PANOP-HO	0.0655±0.0002	−16672.8	0.9979	0.0238
NS	Uniform	PNOP-HE	0.0548±0.0028	−12694.2	0.9010	0.0634
		PNOP-HO	0.1271±0.0122	−8155.2	0.9094	0.1196
		PANOP-HE	0.0478±0.0016	−12818.9	0.9016	0.0573
		PANOP-HO	0.1128±0.0078	−8311.0	0.8944	0.1026
	Clustered	PNOP-HE	0.0729±0.0032	−11812.0	0.9041	0.0867
		PNOP-HO	0.2038±0.0144	−6509.3	0.9193	0.1997
		PANOP-HE	0.0563±0.0020	−12332.2	0.9056	0.0695
		PANOP-HO	0.1735±0.0304	−6917.0	0.9165	0.1676
	Mixed	PNOP-HE	0.0751±0.0141	−11676.2	0.8833	0.0782
		PNOP-HO	0.2186±0.0216	−6470.3	0.9106	0.1909
		PANOP-HE	0.0763±0.0319	−12377.7	0.9180	0.0955
		PANOP-HO	0.1872±0.0104	−6878.6	0.9240	0.1831

observed in DNOP under suboptimal sampling distributions. Importantly, DANOP does not surpass the absolute best result of DNOP, achieved under the favorable clustered, high-budget regime, but instead yields a more robust architecture whose performance varies much less with the spatial distribution of observations. This suggests that explicitly preserving local geometry reduces the model's reliance on favorable sampling configurations.

Probabilistic Analysis and Variance Regularization. As summarized in Table 12, on Burgers the PANOP variant with a HE head consolidates itself as the most accurate architecture. For Navier–Stokes, it similarly demonstrates strong performance under uniform context, yielding the best accuracy and the sharpest uncertainty estimates; across our experiments, this variant also showed the strongest prior–posterior alignment among the probabilistic models.

The Darcy scenario, however, reveals more nuanced behavior regarding the parameterization of predictive uncertainty. The HE variants maintain a residual sensitivity to the context strategy, continuing to benefit modestly from non-uniform distributions. This is consistent with previous results suggesting that the additional flexibility of spatially varying predictive variance can partially absorb observation-geometry effects rather than requiring all improvements to come from the predictive mean alone. By contrast, restricting the models to a HO predictive variance largely removes the dependence on sampling strategy. As also observed in previous sections, although a HO head sacrifices the localized adaptivity of HE uncertainty, it can act as an implicit regularizer, encouraging the probabilistic model to rely more consistently on the representational capacity of its geometry-preserving deterministic trunk (PANOP). In turn, this yields a stable predictive performance across uniform, clustered, and mixed strategies.

C.2 Test-Time Context Scaling

Table 13: Deterministic test-time context scaling across benchmarks. Mean values over 4 seeds. Models are evaluated by changing only the inference-time context budget while reusing checkpoints trained under the base context regime. Values reported represent rel-L^2 error and are specific to the fixed-checkpoint scaling evaluation in this subsection.

PDE	Model	Low Test Budget	Base Test Budget	High Test Budget
Burgers	DNOP	0.00722 ± 0.00053	0.00598 ± 0.00033	0.01042 ± 0.00082
	DANOP	0.00544 ± 0.00022	0.00476 ± 0.00017	0.00481 ± 0.00019
Darcy	DNOP	0.02391 ± 0.00032	0.02354 ± 0.00032	0.02474 ± 0.00047
	DANOP	0.02229 ± 0.00086	0.02222 ± 0.00088	0.02222 ± 0.00086
NS	DNOP	0.03097 ± 0.00088	0.02925 ± 0.00055	0.03412 ± 0.00030
	DANOP	0.03265 ± 0.00048	0.03224 ± 0.00033	0.03230 ± 0.00039

The previous ablations showed that the context distribution during training significantly impacts the learned representation. A natural follow-up question is whether simply providing more context observations at inference time, while reusing a fixed checkpoint trained under a specific context regime, yields systematic performance gains. To isolate this effect, we evaluate models trained at the base context budget across three test-time budgets to assess inference-time scaling behavior across varied physical dynamics.

Deterministic Behavior and Checkpoint Specialization. Table 13 summarizes the test-time context scaling results for the deterministic architectures. The most relevant observation across all three benchmarks is that increasing the context budget exclusively at test time does not yield monotonic performance improvements. For DNOP, the base budget used during training is typically best, or statistically indistinguishable from the best setting, while moving to the high test-time budget either leaves performance unchanged or degrades it, most visibly in Burgers and Navier–Stokes. This suggests that the baseline checkpoint remains adapted to the observation density seen during training and does not systematically benefit from denser test-time conditioning alone.

In contrast, DANOP architecture demonstrates substantially greater stability. Across Burgers, Darcy, and Navier–Stokes, the relative-L^2 error for DANOP remains nearly unchanged across the low, base, and high evaluation budgets. In some cases, the high-budget regime matches the base budget or yields

Table 14: Probabilistic test-time context scaling under prior-sample evaluation. All models were trained at the base context budget using a uniform context strategy. Mean values over 4 seeds.

PDE	Test Budget	Model	Rel-L^2	MC-NLL	Coverage	Width
Burgers	Low	PNOP-HE	0.0311 ± 0.0087	-966.5	0.9651	0.0421
		PNOP-HO	0.0119 ± 0.0004	-860.8	0.9883	0.0463
		PANOP-HE	0.0092 ± 0.0005	-995.4	0.9716	0.0250
		PANOP-HO	0.0114 ± 0.0001	-861.9	0.9873	0.0463
	Base	PNOP-HE	0.0165 ± 0.0005	-988.5	0.9689	0.0274
		PNOP-HO	0.0113 ± 0.0004	-863.9	0.9873	0.0462
		PANOP-HE	0.0092 ± 0.0005	-995.2	0.9722	0.0247
		PANOP-HO	0.0113 ± 0.0001	-862.1	0.9870	0.0462
	High	PNOP-HE	0.0205 ± 0.0005	-977.7	0.9669	0.0311
		PNOP-HO	0.0139 ± 0.0004	-855.3	0.9876	0.0461
		PANOP-HE	0.0100 ± 0.0005	-986.7	0.9663	0.0247
		PANOP-HO	0.0117 ± 0.0001	-860.4	0.9872	0.0461
Darcy	Low	PNOP-HE	0.0814 ± 0.0029	-16514.0	0.9950	0.0238
		PNOP-HO	0.0660 ± 0.0003	-16668.8	0.9980	0.0238
		PANOP-HE	0.0802 ± 0.0036	-16528.1	0.9948	0.0238
		PANOP-HO	0.0657 ± 0.0001	-16671.5	0.9980	0.0239
	Base	PNOP-HE	0.0755 ± 0.0029	-16575.7	0.9960	0.0238
		PNOP-HO	0.0658 ± 0.0003	-16670.0	0.9979	0.0238
		PANOP-HE	0.0721 ± 0.0036	-16606.8	0.9965	0.0238
		PANOP-HO	0.0657 ± 0.0001	-16671.8	0.9979	0.0239
	High	PNOP-HE	0.0770 ± 0.0029	-16560.6	0.9963	0.0238
		PNOP-HO	0.0661 ± 0.0003	-16667.1	0.9979	0.0238
		PANOP-HE	0.0716 ± 0.0036	-16611.8	0.9963	0.0238
		PANOP-HO	0.0657 ± 0.0001	-16671.4	0.9979	0.0239
NS	Low	PNOP-HE	0.0580 ± 0.0028	-11986.3	0.8936	0.0644
		PNOP-HO	0.1336 ± 0.0122	-7072.5	0.9067	0.1199
		PANOP-HE	0.0488 ± 0.0016	-12632.5	0.9046	0.0572
		PANOP-HO	0.1117 ± 0.0078	-7310.1	0.8960	0.1025
	Base	PNOP-HE	0.0548 ± 0.0028	-12131.0	0.9010	0.0634
		PNOP-HO	0.1271 ± 0.0122	-7128.9	0.9094	0.1196
		PANOP-HE	0.0478 ± 0.0016	-12608.4	0.9016	0.0573
		PANOP-HO	0.1128 ± 0.0078	-7368.9	0.8944	0.1026
	High	PNOP-HE	0.0681 ± 0.0028	-11616.1	0.8984	0.0637
		PNOP-HO	0.1769 ± 0.0122	-6527.9	0.9048	0.1196
		PANOP-HE	0.0495 ± 0.0022	-12513.9	0.9022	0.0579
		PANOP-HO	0.1116 ± 0.0078	-7345.7	0.8948	0.1029

only a slight improvement, but these effects are small and not systematic. This suggests that once a robust conditional representation has been learned, additional test-time context can be assimilated without destabilizing inference, yet does not reliably translate into further gains in predictive accuracy.

Probabilistic Behavior and Uncertainty Stability. Table 14 extends this analysis to the probabilistic models evaluated under prior-sampled inference. The results broadly reinforce the deterministic findings while highlighting differences in architectural robustness. The strongest probabilistic variants, particularly PANOP, remain stable across test-time budgets in most settings. In Burgers and Navier–Stokes, PANOP-HE shows only minor variation in predictive accuracy, empirical coverage, and average interval width as the inference budget changes.

A more nuanced pattern appears in Darcy. The HO variants of both PNOP and PANOP are almost completely invariant to the test-time context budget, with negligible differences across all reported metrics. By contrast, the HE models retain some sensitivity, and PANOP-HE shows a visible improvement when moving from the low-budget regime to the base or high-budget regimes. Thus, PANOP remains generally stable overall, with near invariance in the HO setting, while Darcy with HE variance still benefits modestly from moving away from the low-budget regime.

Conversely, the global-latent PNOP architecture with a HO head exhibits greater fragility. In Navier–Stokes, increasing the test-time context to the high-budget regime causes a clear degradation in predictive accuracy without correspondingly sharpening predictive intervals. This shows that additional test-time observations do not necessarily improve robustness in this configuration.

Overall, we observe that performance is governed primarily by the conditional representation learned during training, rather than improving automatically with additional test-time context. Architectures that preserve local geometry, such as DANOP or PANOP, tolerate moderate changes in context budget without substantial degradation but show no systematic gains from extra context alone.

C.3 Train-Evaluation Strategy

Table 15: Deterministic train–evaluation context strategy matrix across benchmarks. Each entry reports mean rel-L^2 error over 4 seeds evaluated at the base context budget for this ablation.

PDE	Model	Train Strategy	Eval: Uniform	Eval: Mixed	Eval: Clustered
Burgers	DNOP	Uniform	0.0060 ± 0.0003	0.0119 ± 0.0008	0.0175 ± 0.0008
		Mixed	0.0069 ± 0.0002	0.0069 ± 0.0002	0.0068 ± 0.0004
		Clustered	0.0140 ± 0.0008	0.0104 ± 0.0008	0.0073 ± 0.0004
	DANOP	Uniform	0.0048 ± 0.0002	0.0066 ± 0.0002	0.0092 ± 0.0002
		Mixed	0.0049 ± 0.0001	0.0049 ± 0.0001	0.0049 ± 0.0001
		Clustered	0.0050 ± 0.0004	0.0050 ± 0.0004	0.0050 ± 0.0004
Darcy	DNOP	Uniform	0.0235 ± 0.0003	0.0436 ± 0.0005	0.0582 ± 0.0005
		Mixed	0.0170 ± 0.0010	0.0170 ± 0.0010	0.0171 ± 0.0010
		Clustered	0.0178 ± 0.0013	0.0177 ± 0.0013	0.0176 ± 0.0013
	DANOP	Uniform	0.0222 ± 0.0009	0.0236 ± 0.0005	0.0244 ± 0.0009
		Mixed	0.0230 ± 0.0004	0.0230 ± 0.0004	0.0230 ± 0.0004
		Clustered	0.0232 ± 0.0008	0.0232 ± 0.0008	0.0232 ± 0.0008
NS	DNOP	Uniform	0.0292 ± 0.0005	0.0918 ± 0.0002	0.1337 ± 0.0002
		Mixed	0.0340 ± 0.0009	0.0343 ± 0.0004	0.0343 ± 0.0004
		Clustered	0.0366 ± 0.0002	0.0360 ± 0.0002	0.0354 ± 0.0002
	DANOP	Uniform	0.0322 ± 0.0003	0.0407 ± 0.0004	0.0466 ± 0.0004
		Mixed	0.0336 ± 0.0003	0.0337 ± 0.0003	0.0337 ± 0.0003
		Clustered	0.0335 ± 0.0007	0.0335 ± 0.0007	0.0335 ± 0.0007

To determine whether the models learn a robust conditional representation or specialize to the context geometry observed during training, we evaluate their performance under out-of-distribution (OOD) geometry shifts. For each benchmark, models are trained under three context sampling strategies and each resulting checkpoint is evaluated across all three strategies at a fixed base budget.

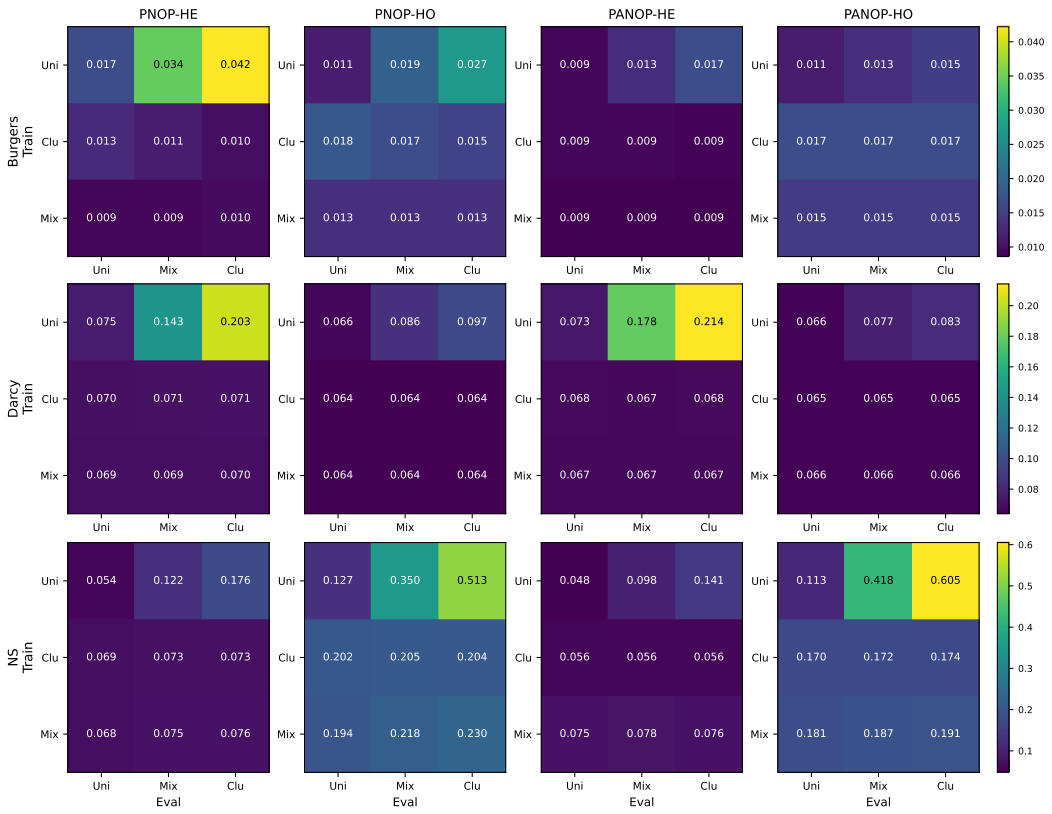


Figure 8: **Probabilistic train–evaluation rel-L^2 matrix under prior-sample evaluation.** Each cell shows performance for a train-eval strategy pair, with rows indicating training distribution and columns evaluation distribution. The diagonal corresponds to matched train/eval geometry. PNOP shows pronounced off-diagonal degradation under uniform training, particularly in Darcy and NS. PANOP-HE reduces this degradation in Burgers and NS, while non-uniform training yields flatter evaluation rows for both architectures in Darcy. Per-PDE normalization is used to make patterns visible across regimes of different scale.

Deterministic Robustness and Specialization. As shown in Table 15, the baseline DNOP exhibits train–eval specialization across all three physical domains. When trained on uniform contexts, DNOP achieves strong in-distribution results but degrades under geometry shift, with the largest penalties typically appearing when evaluation moves to clustered observations and already sizable degradation under mixed evaluation. This effect is most severe in Navier–Stokes and also pronounced in Darcy, indicating that the SetConv-only representation remains tightly coupled to the spatial coverage pattern seen during training.

By contrast, training with mixed or clustered contexts yields much flatter evaluation rows, indicating a trade-off between specialization and robustness in Burgers and Navier–Stokes. In Darcy, however, the pattern is even stronger showing that for DNOP training on mixed or clustered contexts not only improves robustness to evaluation-time geometry shifts, but also improves absolute performance relative to training on uniform contexts, even under uniform evaluation itself. Thus, in Darcy, non-uniform training geometries appear better aligned with the structure of the task.

The introduction of the attention-based local trunk in DANOP largely reduces this tension. Across Burgers and Darcy, DANOP substantially flattens the evaluation matrix, effectively removing the severe performance drop under strategy mismatch. In Navier–Stokes, DANOP similarly reduces the cross-strategy penalty, although some residual sensitivity remains when training and evaluation geometries differ sharply. These results support query-aligned geometry preservation as an effective architectural mechanism for reducing sensitivity to sensor-placement shift.

Probabilistic Robustness and Calibration Shift. Figure 8 extends this analysis to the probabilistic models, showing that train–eval mismatch affects not only the predictive mean but also uncertainty quality. For the PNOP architecture, HE variants are highly sensitive to mismatch. When trained on uniform contexts and evaluated on clustered data, PNOP suffers a severe degradation in rel-L^2 error alongside a drop in empirical coverage.

PANOP mitigates part of this fragility. In Darcy, its HE variant achieves stable performance once training uses mixed or clustered contexts, but does not eliminate the strong mismatch associated with uniform training. In Navier–Stokes, residual sensitivity remains under the strongest train–eval mismatches, particularly when transferring from uniform training to non-uniform evaluation. For instance, PANOP-HE trained on uniform contexts and evaluated on clustered observations reaches a significant relative-L^2 degradation compared to performance in-distribution, highlighting that geometry mismatch in this dynamically complex regime remains challenging even for the strongest architecture.

Constraining the predictive variance to be HO yields the most stable OOD behavior overall. In Darcy, the PANOP-HO is remarkably flat across all train–test pairs once training includes mixed or clustered contexts, and PNOP-HO shows a similar pattern. In Navier–Stokes, the same broad trade-off observed in the deterministic models persists: uniform-trained probabilistic models yield the best in-distribution accuracy and relatively sharp intervals, but remain fragile under evaluation-time geometry shifts, whereas models trained on non-uniform geometries, particularly mixed and clustered regimes, maintain steadier performance and more stable coverage across evaluation strategies.

This ablation shows that observation geometry acts as a strong inductive bias during training. Uniform sampling can be favorable when deployment-time sensing is known to remain somewhat dense and evenly distributed. Training on more diverse non-uniform geometries, particularly mixed and clustered regimes, yields substantially stronger robustness across sensor configurations. Architecturally, explicitly preserving local context–query interactions (DANOP and PANOP) is important for improving generalization across observation patterns and for reducing both deterministic error spikes and probabilistic calibration degradation under train–eval mismatch.

C.4 Cross-Resolution Transfer

Table 16: Deterministic cross-resolution transfer across benchmarks. Models trained on coarse grids are evaluated on paired instances to isolate resolution shift from dataset shift. Evaluation protocols include the paired coarse reference (Coarse → Coarse), zero-shot super-resolution (Coarse → Fine), and super-resolution with proportionally scaled context (sc). The paired coarse datasets are specific to this ablation and are distinct from the native coarse test sets used in the main benchmark. Mean rel-L^2 over 4 seeds.

PDE	Model	Coarse → Coarse	Coarse → Fine	Coarse → Fine (sc)
Burgers	DNOP	0.0058 ± 0.0001	0.0057 ± 0.0001	0.0085 ± 0.0013
	DANOP	0.0048 ± 0.0001	0.0048 ± 0.0001	0.0048 ± 0.0001
Darcy	DNOP	0.0250 ± 0.0006	0.3316 ± 0.0131	0.3287 ± 0.0081
	DANOP	0.0234 ± 0.0008	0.3196 ± 0.0167	0.3193 ± 0.0164
NS	DNOP	0.0187 ± 0.0008	0.2195 ± 0.0033	0.3329 ± 0.0291
	DANOP	0.0191 ± 0.0006	0.2031 ± 0.0045	0.2042 ± 0.0060

A core motivation for neural operator learning is the capacity to learn discretization-invariant mappings that can transfer zero-shot to higher resolutions. However, learning operators from sparse, ungridded context observations introduces a structural challenge as the model must construct a continuous representation from discrete points before evaluating the operator. To assess whether the learned conditional representations transfer robustly across spatial discretizations without retraining, we evaluate models trained on coarse grids under three inference protocols: paired coarse (same resolution as training), paired fine (zero-shot transfer), and paired fine scaled (where the number of context points is proportionally increased to maintain relative observation density). Coarse and fine grids are $256 \to 512$ for Burgers and $64^2 \to 128^2$ for Darcy and Navier–Stokes.

Table 17: Probabilistic cross-resolution transfer under prior-sample evaluation on paired datasets. Mean values over 4 seeds. Avg. MC-NLL is normalized by the number of query points for comparability across resolutions. All results refer to the paired transfer benchmark designed for this ablation.

PDE	Transfer Setting	Model	Rel-L^2	Avg. MC-NLL	Coverage	Width
Burgers	Coarse → Coarse	PNOP-HE	0.0097	−3.896	0.9711	0.0250
		PANOP-HE	0.0089	−3.915	0.9743	0.0249
		PANOP-HO	0.0110	−3.366	0.9879	0.0454
	Coarse → Fine	PNOP-HE	0.0097	−3.899	0.9716	0.0250
		PANOP-HE	0.0089	−3.913	0.9733	0.0249
		PANOP-HO	0.0110	−3.367	0.9878	0.0453
	Coarse → Fine (sc)	PNOP-HE	0.0126	−3.711	0.9504	0.0250
		PANOP-HE	0.0094	−3.895	0.9706	0.0248
		PANOP-HO	0.0114	−3.355	0.9861	0.0454
Darcy	Coarse → Coarse	PNOP-HE	0.0727	−4.055	0.9958	0.0239
		PANOP-HE	0.0707	−4.059	0.9963	0.0238
		PANOP-HO	0.0674	−4.066	0.9973	0.0238
	Coarse → Fine	PNOP-HE	0.3809	−1.624	0.5512	0.0239
		PANOP-HE	0.3568	−1.898	0.5713	0.0239
		PANOP-HO	0.3403	−2.069	0.6025	0.0238
	Coarse → Fine (sc)	PNOP-HE	0.3871	−1.512	0.5487	0.0239
		PANOP-HE	0.3616	−1.836	0.5658	0.0239
		PANOP-HO	0.3404	−2.064	0.6024	0.0238
NS	Coarse → Coarse	PNOP-HE	0.0982	−3.857	0.9595	0.0263
		PANOP-HE	0.0973	−3.869	0.9598	0.0249
		PNOP-HO	0.4047	−2.518	0.9898	0.1087
		PANOP-HO	0.3885	−2.546	0.9910	0.1067
	Coarse → Fine	PNOP-HE	0.2222	−0.367	0.7087	0.0372
		PANOP-HE	0.2585	3.344	0.6503	0.0337
		PNOP-HO	0.4481	−2.141	0.9240	0.1089
		PANOP-HO	0.4291	−2.048	0.9197	0.1067
	Coarse → Fine (sc)	PNOP-HE	0.4385	3.622	0.5636	0.0301
		PANOP-HE	0.3037	3.636	0.6188	0.0323
		PNOP-HO	1.3305	−0.885	0.7152	0.1092
		PANOP-HO	0.5493	−2.014	0.9026	0.1067

The cross-resolution experiments use dedicated paired datasets constructed for this ablation. The paired coarse field is obtained by downsampling an independently simulated fine-resolution field, so the Coarse $\rightarrow$ Coarse evaluation provides the within-suite reference for the corresponding paired Coarse $\rightarrow$ Fine experiment; it is not the native coarse test set used in the main benchmark. For Navier–Stokes, for example, the paired 64×64 fields are downsampled from independently simulated 128×128 fields, whereas the main-text benchmark uses its own natively simulated 64×64 test set.

Deterministic Transfer and the Boundaries of Mesh Invariance. As shown in Table 16, cross-resolution behavior is highly benchmark-dependent, clarifying the empirical limits of mesh invariance in conditional operator learning. On the 1D Burgers equation, zero-shot transfer is largely successful. The baseline DNOP maintains low relative error across resolutions, and the geometry-preserving DANOP is nearly invariant, achieving a similar rel-L^2 error under all three evaluation settings. Thus, in a smooth periodic setting well aligned with the FNO decoder, the learned conditional representation can transfer cleanly across resolutions from sparse observations.

By contrast, the non-periodic and dynamically complex 2D settings expose a limitation. In Darcy, both deterministic models degrade upon transfer, with relative error increasing from the paired coarse setting to the paired fine grid. Navier–Stokes shows a similarly strong degradation. Scaling the context budget proportionally at the fine resolution does not rescue performance in either 2D benchmark. Thus, the issue is not simply a lack of spatial point density at inference time, but that the learned conditional operator model does not transfer cleanly across discretizations in these settings. Robust discretization transfer under sparse response conditioning remains challenging in the 2D benchmarks.

Probabilistic Transfer and Calibration Shift. Results in Table 17 reinforce the deterministic findings and highlight the vulnerability of uncertainty estimation under resolution shift. For Burgers 1D, where the deterministic mapping transfers well, the probabilistic models also generalize reliably. PANOP with a HE likelihood retains strong accuracy, stable MC-NLL, and near-nominal empirical coverage across all coarse and fine evaluation protocols.

However, in Darcy and Navier–Stokes, zero-shot resolution transfer induces both an accuracy failure and a marked calibration shift. For instance, the PANOP-HE drops severely in coverage from the coarse grid to the fine grid. The average interval widths remain nearly unchanged in Darcy and change only modestly in several Navier–Stokes settings despite the increase in predictive error, indicating substantial overconfidence under resolution shift. Proportional context scaling does not mitigate this calibration degradation, and in some cases further destabilizes MC-NLL. These issues mirror the deterministic transfer bottleneck: when the conditional operator model does not transfer reliably across discretizations, the associated uncertainty calibration also fails to adapt appropriately to new resolution regimes.

C.5 Local Conditioning Decomposition

To isolate the contribution of each local conditioning mechanism within the geometry-aware architectures, we decompose the local trunk into three distinct configurations: a purely convolutional pathway (*SetConv only*), a purely query-aligned attention pathway (*Attention only*), and their combination. The objective is to determine whether the empirical gains of the final models arise primarily from continuous kernelized smoothing, dynamic query–context geometry preservation, or their fusion.

Deterministic Decomposition and the Role of Periodicity. Figure 9 indicates that the most effective local conditioning mechanism depends on the underlying PDE structure. On the Navier–Stokes benchmark, *SetConv only* is the strongest variant in this deterministic trunk decomposition, achieving lower errors than the attention-based alternatives. In this deterministic setting, the smoother translation-friendly representation provided by SetConv is already well aligned with the downstream spectral decoder, so adding attention does not improve pure point prediction here.

On the Burgers equation, the full *SetConv + Attention* trunk performs best, although all three variants remain highly accurate. The gain is modest, but it indicates that combining the two pathways can still be beneficial even in a smooth periodic regime. By contrast, the non-periodic Darcy flow exposes a stronger role for the attention pathway. On this boundary-sensitive elliptic problem, *Attention only* is the strongest deterministic configuration, slightly outperforming the fused model

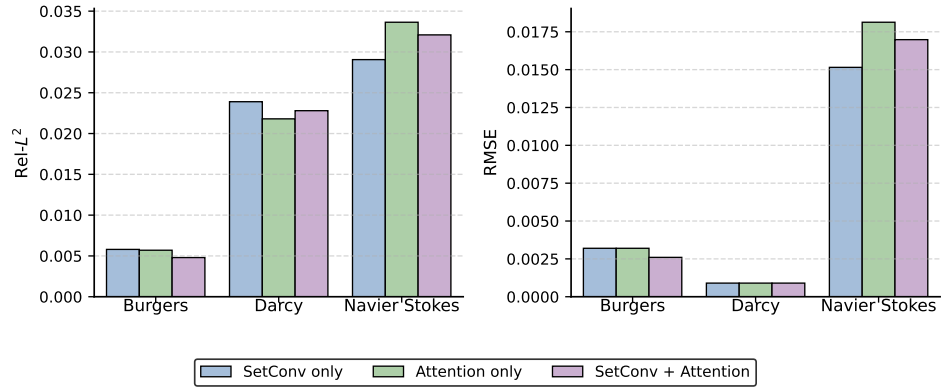


Figure 9: **Deterministic local conditioning decomposition across benchmarks.** The plot shows the isolated and combined effects of the SetConv and query-aligned attention pathways within the local conditioning trunk. Mean values over 4 seeds. Lower is better.

Table 18: Probabilistic local conditioning decomposition across benchmarks under prior-sample evaluation. The table reports predictive and calibration metrics for heteroscedastic and homoscedastic heads across all three local conditioning architectures. Mean values over 4 seeds. Lower is better for rel-L^2, RMSE, MC-NLL, and Width; coverage is best when close to the nominal 0.95 level.

PDE	Variance	Variant	Rel-L^2	RMSE	MC-NLL	Coverage	Width
Burgers	HE	SetConv only	0.0093	0.0051	-993.49	0.9712	0.0247
		Attention only	0.0082	0.0046	-1007.83	0.9760	0.0244
		SetConv + Attention	0.0094	0.0052	-994.16	0.9726	0.0248
	HO	SetConv only	0.0114	0.0062	-850.87	0.9862	0.0461
		Attention only	0.0116	0.0063	-847.02	0.9853	0.0460
		SetConv + Attention	0.0112	0.0061	-853.08	0.9870	0.0457
Darcy	HE	SetConv only	0.0756	0.0030	-16576.12	0.9941	0.0252
		Attention only	0.0678	0.0027	-16648.74	0.9970	0.0239
		SetConv + Attention	0.0686	0.0027	-16634.09	0.9967	0.0238
	HO	SetConv only	0.0662	0.0026	-16666.60	0.9975	0.0238
		Attention only	0.0655	0.0026	-16672.84	0.9979	0.0238
		SetConv + Attention	0.0659	0.0026	-16669.56	0.9978	0.0238
NS	HE	SetConv only	0.0570	0.0217	-12195.4	0.8982	0.0652
		Attention only	0.0560	0.0212	-12439.4	0.8971	0.0622
		SetConv + Attention	0.0471	0.0183	-12628.7	0.9017	0.0568
	HO	SetConv only	0.1456	0.0363	-7113.8	0.9159	0.1579
		Attention only	0.1164	0.0315	-6566.8	0.9195	0.1888
		SetConv + Attention	0.2307	0.0461	-6301.0	0.9245	0.2116

and clearly improving over *SetConv only*. Overall, the deterministic results suggest that attention is particularly valuable when local spatial topology is critical, whereas SetConv alone can already be highly competitive in smoother periodic settings.

Probabilistic Decomposition and Uncertainty-Aware Conditioning. Table 18 shows that the role of the local conditioning pathways becomes more nuanced once uncertainty estimation is introduced. On Darcy, *Attention only* remains marginally strongest, although the margin is much clearer in the HE case than in the HO. In particular, under HO variance the advantage over *SetConv + Attention* is small, so the results are better interpreted as indicating that attention is the main driver of improvement on Darcy, with the fused variant remaining competitive.

A different pattern emerges on Navier–Stokes. Under HE variance, the full *SetConv + Attention* trunk is clearly the best configuration, achieving the lowest rel-L^2 error, the best MC-NLL, and the narrowest predictive intervals. Neither pathway alone matches the combined model in this

Table 19: Direct masked-grid conditioning control. Values are mean relative-L^2 error $\pm$ standard deviation. All methods are re-evaluated on identical frozen tasks and context realizations. Masked-FNO-Geom is used only for Darcy and receives the same boundary-distance features as the NOP models.

PDE	Dense FNO	Masked-FNO	Masked-FNO-Geom	DNOP	DANOP
Burgers	0.005070 ± 0.000054	0.004919 ± 0.000299	–	0.005695 ± 0.000193	0.004954 ± 0.000087
Darcy	0.025408 ± 0.000290	0.025354 ± 0.000935	0.024294 ± 0.000197	0.024193 ± 0.000195	0.022549 ± 0.000758
Navier–Stokes	0.034341 ± 0.000468	0.033888 ± 0.001030	–	0.028395 ± 0.000476	0.032160 ± 0.000574

setting. This is consistent with the view that, when probabilistic prediction is required in a more demanding regime, the smoother SetConv signal and the query-aligned attention pathway can become complementary. However, this is not uniform across all settings, under HO variance on Navier–Stokes, the fused model is not the strongest configuration, and *Attention only* performs better in pointwise error. The advantage of the combined trunk should also be interpreted as regime-dependent rather than universal.

We can see that the value of local geometric conditioning depends both on the benchmark and on the inference objective. *SetConv only* is not uniformly sufficient, as it underperforms on Darcy and does not provide the strongest probabilistic results in the most demanding settings. *Attention only* is often very strong, particularly on Darcy, but it is not consistently best across all deterministic and probabilistic regimes. The combined *SetConv + Attention* design therefore remains a reasonably robust choice. Although it is not the strongest configuration in every setting, it performs well across multiple benchmarks and inference objectives, including deterministic Burgers and probabilistic Navier–Stokes with HE variance, offering a reliable overall trade-off across benchmarks and inference objectives.

C.6 Conditioning and Decoder Controls

The preceding ablations compare alternative conditioning mechanisms within the NOP family. We next perform two complementary controls that modify core components of the architecture to separate the roles of sparse conditioning and decoder inductive bias. First, we test whether sparse observations need to be encoded through a learned set-conditioning mechanism, or whether they can instead be rasterized directly as additional FNO input channels. Second, we preserve the SetConv conditioning pathway but replace the FNO decoder with a finite-receptive-field convolutional decoder.

Direct masked-grid conditioning control. We construct a Masked-FNO that receives $[a_Q,\ u_{\mathrm{obs}},\ m]$, where a_Q is the query-side input field, u_{obs} contains the observed context outputs at their corresponding grid locations and is zero elsewhere, and m is the associated binary observation mask. The model therefore receives the same sparse response observations and query-side input information as the NOP models, but replaces learned set-to-function conditioning with direct grid-channel injection.

For Darcy, where geometric information is important, we additionally evaluate a geometry-matched variant using the same boundary features as the NOP models,

$$b(x, y) = [x,\ 1 - x,\ y,\ 1 - y,\ \min(x, 1 - x, y, 1 - y)].$$

These features depend only on the query coordinates and contain no target information. The strict Masked-FNO therefore evaluates direct rasterization under the basic FNO input representation, while the geometry-matched variant controls for the additional geometric representation available to DNOP and DANOP.

All models are evaluated on identical frozen manifests containing 200 held-out tasks per PDE and the same context realizations across methods. Pairwise comparisons use 1,000 paired task-level bootstrap resamples.

When is direct rasterization sufficient? The effect of replacing learned conditioning by direct grid injection is strongly regime-dependent. On Burgers, Masked-FNO is statistically indistinguishable from both Dense FNO and DANOP under the paired evaluation, while outperforming DNOP ($\Delta = -7.76 \times 10^{-4}$, 95% CI $[-1.12, -0.47] \times 10^{-3}$, where Δ denotes Masked-FNO minus the comparator).

Table 20: Capacity matching for the local ConvCNP decoder. Matching is performed against the FNO decoder parameters only and leaves the four-block, kernel-3, receptive-field-9 local architecture unchanged.

PDE	Compact params	Matched width	Matched params	FNO params	Rel. diff.
Burgers	49,408	179	408,363	409,856	0.364%
Darcy	147,712	340	4,206,884	4,210,944	0.096%
Navier–Stokes	147,712	340	4,206,884	4,210,944	0.096%

Table 21: Decoder-family control under the same frozen evaluation manifests. ConvCNP-Compact and ConvCNP-Matched retain DNOP's SetConv conditioning pathway while replacing the FNO decoder with a local residual CNN. Values are mean relative-L^2 error $\pm$ standard deviation.

PDE	ConvCNP-Compact	ConvCNP-Matched	DNOP	DANOP
Burgers	0.125912 ± 0.003169	0.112334 ± 0.000749	0.005695 ± 0.000193	0.004954 ± 0.000087
Darcy	0.016389 ± 0.000260	0.018511 ± 0.000627	0.024193 ± 0.000195	0.022549 ± 0.000758
Navier–Stokes	0.045723 ± 0.001147	0.051971 ± 0.004323	0.028395 ± 0.000476	0.032160 ± 0.000574

Thus, in this smooth, periodic, and grid-aligned regime, exposing sparse observations directly to the spectral decoder is sufficient for strong point prediction, and a learned context-to-query representation provides little additional benefit under this protocol.

Darcy exposes a different effect. The strict Masked-FNO remains close to the dense reference, while adding the boundary representation reduces its mean error from 0.02535 to 0.02429. Once this geometric information is matched, Masked-FNO-Geom is statistically indistinguishable from DNOP ($\Delta = 1.02 \times 10^{-4}$, 95% CI $[-3.02, 5.18] \times 10^{-4}$). SetConv alone therefore does not provide a measurable advantage over geometry-matched rasterization in this setting. DANOP, however, remains significantly better than the geometry-matched control ($\Delta = 1.75 \times 10^{-3}$, 95% CI $[1.35, 2.19] \times 10^{-3}$), supporting an additional benefit from the query-aligned conditioning pathway.

In Navier-Stokes, Masked-FNO slightly improves over the dense FNO ($\Delta = -4.52 \times 10^{-4}$, 95% CI $[-9.11, -0.48] \times 10^{-4}$), but it has approximately 19% higher relative-L^2 error than DNOP ($\Delta = 5.49 \times 10^{-3}$, 95% CI $[4.14, 6.89] \times 10^{-3}$) and also significantly underperforms DANOP ($\Delta = 1.73 \times 10^{-3}$, 95% CI $[1.20, 2.36] \times 10^{-3}$). Thus, simply making the sparse observations available as grid channels does not reproduce the benefit of learned sparse conditioning in this more complex regime.

Local decoder control. The masked-grid experiment modifies the conditioning mechanism while retaining the FNO decoder. We next perform the complementary intervention: we preserve DNOP's SetConv conditioning path, input projection, query features, and output head, but replace the FNO decoder with a finite-receptive-field residual CNN. This produces a ConvCNP-style conditional model in which the encoder-side conditioning remains fixed while the operator decoder is changed.

The local decoder contains four residual convolutional blocks with kernel size three, giving a theoretical receptive field of nine. We evaluate two variants. The compact model uses a comparatively small local CNN, while the capacity-controlled model selects its internal width solely to approximately match the parameter count of the corresponding DNOP FNO decoder. The matched width is 179 for Burgers and 340 for Darcy and Navier–Stokes. Table 20 shows that the resulting decoder parameter counts differ from the FNO decoder by less than 0.4% in all cases. Parameter matching reduces the capacity confound, but does not eliminate other architectural differences between finite-range convolution and spectral/global mixing.

Decoder inductive bias is regime-dependent. Matching decoder capacity does not remove the strong dependence on PDE family. On Burgers, ConvCNP-Matched remains nearly $20\times$ worse in relative-L^2 than DNOP, with a paired difference of $\Delta = 0.1066$ (95% CI $[0.0952, 0.1181]$), where here Δ denotes ConvCNP-Matched minus DNOP. Increasing the local decoder capacity improves substantially over ConvCNP-Compact ($\Delta = -0.01358$, 95% CI $[-0.01698, -0.01010]$), but remains

far from the FNO-based models. The result therefore cannot be explained by the compact CNN simply having fewer parameters.

Navier–Stokes exhibits the same qualitative preference for the FNO decoder. ConvCNP-Matched significantly underperforms DNOP ($\Delta = 0.02358$, 95% CI $[0.02097, 0.02643]$), while the compact model also remains substantially worse. Moreover, increasing the local decoder capacity does not improve performance: the matched model has higher error than the compact variant ($\Delta = 0.00625$, 95% CI $[0.00579, 0.00675]$). This further indicates that parameter count alone does not explain the observed decoder gap.

Darcy reverses this pattern. ConvCNP-Matched significantly improves over DNOP ($\Delta = -0.00568$, 95% CI $[-0.00725, -0.00405]$), and the smaller compact decoder performs better still. The compact model also outperforms the matched variant ($\Delta = -0.00212$, equivalently Matched$-$Compact $= 0.00212$ with 95% CI $[0.00184, 0.00238]$).

The FNO decoder is therefore not uniformly advantageous within the NOP pipeline. These results are consistent with the earlier observation that a spectral decoder is particularly effective on the regular periodic Burgers regime, whereas the non-periodic elliptic Darcy problem can favor a more local convolutional representation. Since replacing the decoder also changes locality, spectral structure, and optimization properties, this experiment should be interpreted as evidence for decoder-family dependence rather than as isolating spectral mixing as the sole causal mechanism.

Joint interpretation. The two controls intervene on complementary components of the architecture. On Burgers, direct masked-grid conditioning is sufficient to match the strongest set-conditioned model, whereas replacing the FNO decoder with a local CNN is severely detrimental. Performance in this regime therefore depends more strongly on decoder inductive bias than on the particular sparse-context encoder.

On Darcy, matching the boundary representation closes the gap between direct rasterization and DNOP, while DANOP retains a significant advantage over the geometry-matched Masked-FNO. At the same time, both local convolutional decoders outperform the FNO-based models. This indicates sensitivity to both geometric representation and decoder choice, with the fixed FNO backbone becoming a limiting factor in this non-periodic elliptic setting.

Finally, on Navier–Stokes, both interventions degrade performance. Direct rasterization underperforms the learned SetConv representation, while replacing the FNO decoder with a local CNN produces a further substantial loss in accuracy. Learned sparse conditioning and FNO decoding therefore provide complementary benefits in this regime.

More generally, these controls do not identify a universally dominant conditioning or decoding mechanism. Instead, they show that the contribution of each component depends on the geometry and dynamics of the underlying operator. This supports interpreting NOP as a modular sparse conditional operator framework rather than tying its contribution to a particular conditioning mechanism or decoder family. The probabilistic NOP variants additionally model a predictive distribution over the output field, a capability not provided by either deterministic control considered here.

C.7 Computational Cost and Model Complexity

We note the main asymptotic distinction between the conditioning and decoding stages. For n_C context points and n_Q query locations, both SetConv and query-aligned attention require evaluating context–query interactions, yielding $O(n_C n_Q)$ computational cost and $O(n_C n_Q)$ memory when the corresponding interaction weights are materialized. The FNO decoder instead scales as $O(L n_Q \log n_Q)$ in the query grid size for L spectral layers, up to channel, mode, and implementation constants. Thus, the conditioning stage is inexpensive in the sparse regimes considered here, where $n_C \ll n_Q$, but can become a bottleneck as the context set grows.

We profile the models' resource requirements across all three benchmarks. The objective is to quantify the overhead introduced by the query-aligned attention pathway and the latent conditioning mechanisms relative to the SetConv-only deterministic baseline (DNOP). Table 22 summarizes the parameter counts, inference latency per batch, peak inference memory, mean epoch training time, and peak training memory for the final model configurations. For probabilistic models, inference

Table 22: Computational cost and complexity profiling across benchmarks. Mean values across 4 seeds, measured using evaluation batch size $B = 16$ for Burgers and $B = 8$ for Darcy/Navier–Stokes. Probabilistic inference uses 8 Monte Carlo samples. Training costs increase with the geometry-aware and probabilistic extensions, while deployment-time latency remains modest.

PDE	Model	Parameters	Infer time/batch (s) ↓	Peak infer mem (MB) ↓	Epoch time (s) ↓	Peak train mem (MB) ↓
Burgers	DNOP	418,625	0.00531	43.31	0.494	92.62
	DANOP	456,577	0.00568	48.90	0.718	110.37
	PNOP-HE	446,594	0.00506	43.45	0.694	102.21
	PNOP-HO	446,530	0.00532	43.44	0.699	102.19
	PANOP-HE	575,298	0.00590	46.36	1.030	153.28
	PANOP-HO	575,234	0.00597	46.35	1.040	153.22
Darcy	DNOP	4,220,097	0.00338	276.30	1.059	417.89
	DANOP	4,258,497	0.00434	341.74	1.514	749.78
	PNOP-HE	4,248,066	0.00387	276.66	2.127	4004.46
	PNOP-HO	4,248,002	0.00386	276.54	2.130	4004.33
	PANOP-HE	4,377,666	0.00468	277.16	4.331	4125.52
	PANOP-HO	4,377,602	0.00469	277.03	4.327	4125.40
NS	PNOP-HE	4,247,746	0.00390	281.29	2.135	4004.33
	PNOP-HO	4,247,682	0.00388	281.16	2.152	4004.20
	PANOP-HE	4,376,706	0.00501	281.78	4.365	4124.64
	PANOP-HO	4,376,642	0.00485	281.65	4.355	4124.51

is measured under the prior-sampled setting with a fixed Monte Carlo budget, reflecting the actual deployment-time evaluation.

Higher-dimensional scaling. The NOP formulation is not tied to a particular spatial dimension, and its computational requirements inherit the scaling of both the sparse conditioning mechanism and the chosen neural-operator backbone. Sparse context limits the additional conditioning overhead, while larger query grids primarily increase decoder and activation costs. Extending the empirical profiling to three-dimensional and larger-scale domains is left for future work.

The Training vs. Inference Cost Asymmetry. The main pattern emerging across benchmarks is that computational cost increases in a predictable order with architectural complexity, but this overhead is asymmetric; it is observed primarily during training rather than at inference. On Burgers, the inference latency of all models remains tightly clustered between 5.0 and 6.0 milliseconds per batch. Similarly, on Darcy, while inference time grows from the baseline DNOP (~ 3.38 ms) to the full probabilistic PANOP (~ 4.69 ms), it remains strictly on the order of milliseconds. Thus, deployment-time latency stays low across all profiled models, even as training cost increases.

By contrast, the training cost scales much more sharply. On Darcy, DANOP increases the mean epoch time by roughly 43% over DNOP, and the latent probabilistic models increase it by factors of roughly two to four. This trend is even more pronounced in peak training memory: while the deterministic models remained below 1 GB on Darcy, the probabilistic models required approximately 4 GB. This jump is much larger than what parameter count alone would suggest. The main practical cost of the probabilistic operator-process models may therefore lie in the optimization graph and training-time activations required by latent inference, rather than in deployment-time execution.

Computational Graph vs. Parameter Count. The Navier–Stokes profiling data isolates the cost differences between the probabilistic variants and helps clarify the source of the overhead. PANOP is only marginally larger in parameter count than PNOP (an increase of approximately 3%) and exhibits nearly identical peak memory usage during both inference and training. However, its runtime overhead is pronounced: mean epoch time roughly doubles and inference time per batch increases by approximately 25–28%. The computational premium of PANOP is not well explained by parameter count alone, but appears to be driven more by the computational graph induced by the architecture, specifically, the evaluation and fusion of both SetConv and query-aligned attention pathways before the operator decoder. The cost gap is therefore driven primarily by how the model computes, rather than by a large increase in raw parameter capacity.

Inference Memory Profiling. Peak inference memory is implementation-dependent, reflecting batching, activation materialization, CUDA allocator behavior, and evaluation scope. We therefore

treat it as a practical profiling statistic rather than a strict architectural ordering. The conclusion is that probabilistic variants mainly increase training-time cost, while inference latency remains modest.

These analyses support a favorable cost–performance trade-off for the proposed architectures, with an important qualification: the additional cost is concentrated mainly in training rather than in inference. While PANOP is substantially more expensive to optimize, its parameter count remains close to the simpler baselines, its inference latency stays within a small millisecond-scale range, and its deployment-time memory footprint remains manageable. Given the gains in geometric robustness and the improved probabilistic behavior observed in the main evaluations and related ablations, this additional optimization cost appears reasonable for applications requiring reliable operator inference with sparse response conditioning.