MF-SPECBO: MULTI-FIDELITY SPECTRAL BAYESIAN OPTIMIZATION FOR NEURAL OPERATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Under Direct Spectral Evaluation (DSE), a neural operator's cost is set by *which* Fourier modes it keeps, not just how many — yet whether informed mode selection can beat a plain isotropic truncation is untested. We give the question a screen: the gain of *any* selection rule over the matched-budget radial box is at most the spectrum's radial defect Δ, a one-pass statistic (Theorem 1). We then build the machinery to search — a Sobolev chordal set kernel, a downward-closed restriction small enough for *exact* acquisition maximization, a calibrated MF-GP-UCB fidelity rule — and run the comparison at exactly matched cardinality. On standard benchmarks Δ is tiny and, as predicted, no choice of *mode set* beats the box (a growth *curriculum* on the same set does — the win is in training, not selection); compression still yields 86.2–90.7% fewer modes, with 2.59–2.68$\times$ measured speedup on meshes above 10^3 nodes (the smallest mesh is launch-bound and gains none), but needs no search. On an anisotropic Darcy-flux family that passes the screen, MF-SpecBO improves on the box by up to 11.55 points and on a one-pass energy heuristic by up to 9.70 ($p < 0.001$ at κ=4), losing the advantage again at extreme anisotropy where the winning modes exceed what 800 samples can estimate. Mode selection pays in a regime Δ locates in advance; every tabulated number is machine-verified against its recorded measurement.

1 INTRODUCTION

Machine learning for scientific computing has witnessed rapid advancement through the development of *neural operators* (Kovachki et al., 2023; Lu et al., 2021; Li et al., 2021), which approximate non-linear solution operators between Banach spaces of partial differential equations (PDEs). Among these, the Fourier Neural Operator (FNO) (Li et al., 2021) computes non-local integral transforms via spectral convolution in the frequency domain.

Conventional FNOs restrict the representation to a dense, axis-aligned Cartesian box $\mathcal{K}_{\text{iso}} = \{\mathbf{k} \in \mathbb{Z}^d : \|\mathbf{k}\|_\infty \le k_{\max}\}$ (the full candidate lattice of Section 2; the matched-budget *radial* baseline B_M of Section 3 is instead an ℓ_2 disc, and we keep the two names distinct throughout). The standing argument against this is that physical spectra are directional — transonic shocks, shear boundary layers, anisotropic permeability — so an axis-agnostic box spends parameters on modes carrying negligible energy while under-resolving the directions that matter. Under the FFT this argument has little practical bite, since the transform computes all modes regardless. It acquires bite under *Direct Spectral Evaluation* (DSE) (Lingsch et al., 2024), which evaluates spectral convolutions on arbitrary point clouds $X = \{\mathbf{x}_n\}_{n=1}^N$ by non-uniform Vandermonde projection at $\mathcal{O}(N|\mathcal{K}|)$ cost: there, every excluded wave vector removes a column of the projection matrix and with it real FLOPs, memory and latency. Yet DSE implementations still use isotropic boxes, leaving the choice of $\mathcal{K}$ unexamined.

This paper asks whether examining it pays, and is organized so the answer can come out either way. The obstacle is cost rather than combinatorics — one candidate architecture costs a full training run — so we build the machinery to search efficiently, then run the comparison against the baselines that could displace it, including the one this argument is really about: a matched-budget isotropic box.

Treating spectral mode selection as an active search over binary indicators $\mathbf{z} \in \{0,1\}^D$ raises three difficulties: each evaluation costs a full operator training run; a kernel over mode *sets* must reflect

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.

harmonic decay rather than treating wave vectors exchangeably; and coarse evaluations are cheap but biased, with the bias varying over both spatial resolution ($\mathcal{O}(N_m^{-r/d})$) and training epochs ($\mathcal{O}(E_m^{-1})$).

Main Contributions.

1. **A screen with a theorem.** The gain of *any* selection rule over the matched-budget radial box is at most the radial defect Δ plus a computable slack (Theorem 1); both are one-pass statistics, while the popular second-moment anisotropy ratio does not control G (counterexample in Appendix A).
2. **An admissible space small enough to enumerate.** Downward closure cuts candidates from $2^{\Theta(k_{\max}^d)}$ to $\exp(\Theta(k_{\max}^{d-1}))$ (Theorem 2); at fixed cardinality the pool (42–9,825 sets) admits *exact* acquisition maximization, and under measured Gram decay $\beta = 3.42$, $\gamma_T = \mathcal{O}(T^{1/\beta}(\log T)^{1-1/\beta})$.
3. **A calibrated fidelity rule.** Stage 2 instantiates MF-GP-UCB's bias-thresholded escalation (Kandasamy et al., 2016) for a continuous resolution×epoch index; our addition is fitting ζ to measured anchors — with a plausible fixed constant the policy silently degenerates to single fidelity. Theorem 3 bounds target-fidelity queries by $\mathcal{O}(\beta_T^2\gamma_T/\varepsilon^2)$, closes no-starvation via pool finiteness, and quantifies the cross-fidelity transfer ceiling $\sigma_f^2(1-\rho^2)$.
4. **A falsifiable empirical study.** All rules at exactly matched $|\mathcal{K}|$ under one protocol, against the baselines that could displace the method (isotropic box, one-pass energy, iFNO-style growth, fine-tuned pruning), with equivalence framing and machine-checked numbers.

Closest work. DSE (Lingsch et al., 2024) makes mode identity a cost decision; iFNO (George et al., 2024) grows an *isotropic* threshold (compared in Table 1); Stage 2 instantiates MF-GP-UCB's escalation (Kandasamy et al., 2016); combinatorial BO follows BOCS/COMBO (Baptista & Poloczek, 2018; Oh et al., 2019), plus a PDE-dictated set kernel and an enumerable pool (extended survey: Appendix B).

2 Problem Formulation

Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain discretized by an arbitrary point cloud $X = \{\mathbf{x}_n\}_{n=1}^N$; we learn $\mathcal{G}^\dagger : L^2(\Omega;\mathbb{R}^{d_a}) \to L^2(\Omega;\mathbb{R}^{d_u})$ from pairs $\{(a_j,u_j)\}_{j=1}^{N_{\text{data}}}$. All benchmark operators use $L = 4$ spectral layers of width $d_v = 32$. **Parameter counting.** Each spectral layer stores one complex multiplier per active mode and channel pair — $d_v^2|\mathcal{K}|$ complex entries per layer, $Ld_v^2|\mathcal{K}|$ across the network — plus 8,705 real lifting/projection weights. Throughout, "Params" counts each complex multiplier as a single entry (e.g. 1,192,449 at $|\mathcal{K}| = 289$; real trainables in the spectral part are exactly $2\times$ this); the convention, and the independent storage of conjugate pairs, is identical for every method, so all reductions and ratios are like-for-like.

2.1 Direct Spectral Evaluation (DSE)

Following Lingsch et al. (2024), a spectral layer applies a learnable multiplier $\mathbf{R}(\mathbf{k}) \in \mathbb{C}^{d_v\times d_v}$ per retained wave vector, evaluated on arbitrary points through the Vandermonde matrix $\mathbf{V}_\mathcal{K} \in \mathbb{C}^{N\times M}$:

$$[\mathbf{V}_\mathcal{K}]_{n,m} = \exp\left(2\pi i\langle \mathbf{k}_m, \mathbf{x}_n\rangle\right), \quad n \in [N],\, m \in [M] \tag{1}$$

acting on nodal features $\mathbf{V} \in \mathbb{R}^{N\times d_v}$ by

$$\widehat{\mathbf{V}} = \frac{1}{N}\mathbf{V}_\mathcal{K}^{\mathrm{H}}\mathbf{V} \in \mathbb{C}^{M\times d_v}, \quad (\mathcal{K}_{\text{spectral}}\mathbf{V}) = \mathrm{Re}\left(\mathbf{V}_\mathcal{K}\widehat{\mathbf{W}}\right) \in \mathbb{R}^{N\times d_v}, \quad \widehat{\mathbf{W}}_m = \mathbf{R}(\mathbf{k}_m)\,\widehat{\mathbf{V}}_m \tag{2}$$

each row of $\widehat{\mathbf{V}}$ is contracted with its wave vector's full $d_v \times d_v$ multiplier — the channel-mixing convention Proposition 1 counts.

Proposition 1 (Computational Complexity of DSE Architectures)**.** *For an L-layer neural operator with channel width d_v evaluated on N spatial coordinates with active frequency mode set $\mathcal{K}$, the overall forward-backward computational complexity is $\mathcal{O}\left(L\left(N\cdot|\mathcal{K}|\cdot d_v + Nd_v^2 + |\mathcal{K}|\cdot d_v^2\right)\right)$, and parameter memory is $\mathcal{O}(L|\mathcal{K}|d_v^2)$. Here, $N|\mathcal{K}|d_v$ arises from non-local forward and adjoint Vandermonde projections, $|\mathcal{K}|d_v^2$ from spectral multiplier contractions, and Nd_v^2 from local pointwise channel-mixing bypass layers.*

Because the pointwise terms are invariant to $|\mathcal{K}|$, wall-clock speedup is always smaller than the mode-cut ratio; the measured factors for our configurations are reported in Table 2 and dissected in Section 5.5.

2.2 Combinatorial Mode Discovery on Discrete Lattices

Let $\Omega_{\text{cand}} = [-k_{\max}, k_{\max}]^d \cap \mathbb{Z}^d$ denote the candidate frequency lattice with total candidate modes $D = |\Omega_{\text{cand}}| = (2k_{\max} + 1)^d$. We represent any candidate spectral configuration by a binary indicator vector $\mathbf{z} \in \{0, 1\}^D$, where $z_i = 1$ if and only if $\mathbf{k}_i \in \mathcal{K}$.

The objective is validation performance under a cardinality budget:

$$\max_{\mathbf{z} \in \{0,1\}^D} f(\mathbf{z}) \quad \text{s.t.} \quad \|\mathbf{z}\|_0 \leq M_{\text{budget}} \tag{3}$$

where $f(\mathbf{z}) = -\log \mathcal{L}_{\text{val}}(\mathcal{G}_{\mathcal{K}(\mathbf{z})})$.

3 A Limit on Mode Selection

How much could *any* selection rule win on a given problem? The answer is controlled by a one-pass statistic of the data, and it is what makes Section 5 falsifiable rather than exploratory.

Let $p \geq 0$ (summing to 1 over Ω_{cand}) be the spectral energy distribution of the training solutions, $\mathcal{S}_r$ the radial shells, and

$$\bar{p}(\mathbf{k}) = \frac{1}{|\mathcal{S}_{\|\mathbf{k}\|}|} \sum_{\mathbf{k}' \in \mathcal{S}_{\|\mathbf{k}\|}} p(\mathbf{k}') \tag{4}$$

be the *radial average* of p: the radially symmetric distribution with the same shell totals. Define the **radial defect** $\Delta(p) = \|p - \bar{p}\|_1$, and for a budget M let B_M denote the matched-budget radial truncation and

$$G(M) \;=\; \max_{|\mathcal{K}|=M} \sum_{\mathbf{k} \in \mathcal{K}} p(\mathbf{k}) \;-\; \sum_{\mathbf{k} \in B_M} p(\mathbf{k}) \tag{5}$$

the **selection gain**: the extra spectral energy the best possible M-subset captures over the isotropic box. $G(M)$ upper-bounds what *any* selection method — ours, an energy heuristic, an oracle — can extract at budget M.

Theorem 1 (Achievable gain is bounded by the radial defect)**.** *Let* $D_M(\bar{p}) = \max_{|\mathcal{K}|=M} \sum_{\mathbf{k} \in \mathcal{K}} \bar{p}(\mathbf{k}) - \sum_{\mathbf{k} \in B_M} \bar{p}(\mathbf{k}) \geq 0$ *be the shell-ordering slack of the radial box under the radial average (also computable in one pass). Then for every budget* M*, unconditionally,*

$$0 \;\leq\; G(M) \;\leq\; \Delta(p) + D_M(\bar{p}), \tag{6}$$

and $D_M(\bar{p}) = 0$ *whenever* $\bar{p}$ *is non-increasing in* $\|\mathbf{k}\|$ *(harmonic decay), giving* $G(M) \leq \Delta(p)$*. In particular* $\Delta = D_M = 0$ *implies* $G(M) = 0$*: on a radially symmetric spectrum with decaying shells, no mode-selection method whatsoever can capture more energy than the isotropic box.*

Proof: Appendix A (three lines) — together with the B_M *tie-break convention, measured slack values, and a counterexample showing the second-moment ratio* $\hat{A}$ *does* not *control* G*: energy on* $(\pm 1, 0)$ *and* $(0, \pm 3)$ *in ratio* 9:1 *gives* $\hat{A} = 1$ *exactly while* $\Delta = 1$ *and* $G(5) = 0.10$*.* $\hat{A}$ *appears in this paper for intuition only.*

Scope. G concerns captured spectral energy, not test error. The two coincide for operators whose output is band-limited to $\mathcal{K}$, where truncation error is exactly the discarded energy; our architecture additionally carries a pointwise bypass path that is not band-limited, so for it Theorem 1 bounds the headroom available to the spectral pathway rather than the end-to-end error gap. Section 5.3 reports measured Δ and G for every benchmark, and Section 5.4 tests whether available headroom is actually realized.

4 MF-SPECBO: COST-AWARE SEARCH OVER ADMISSIBLE MODE SETS

Theorem 1 says when searching can pay; this section builds the machinery for when it can: a set kernel encoding harmonic decay, an enumerable admissible pool, a calibrated multi-fidelity acquisition, and a training-free complexity surrogate.

Hypercube kernels treat bits exchangeably (Oh et al., 2019); physical Fourier coefficients decay as $|\hat{u}(\mathbf{k})| \leq C(1 + \|\mathbf{k}\|_2^2)^{-s/2}$ (Adams & Fournier, 2003), and the kernel should know it.

4.1 SOBOLEV SET KERNEL

Each candidate subset $\mathcal{K}$ becomes a weighted discrete measure in the dual Sobolev space $H^{-s}(\mathbb{T}^d)$:

$$\mu_{\mathcal{K}} = \sum_{\mathbf{k}\in\mathcal{K}} w_\alpha(\mathbf{k})\,\delta_{\mathbf{k}}, \quad w_\alpha(\mathbf{k}) = \left(1 + \|\mathbf{k}\|_2^2\right)^{-\alpha} \tag{7}$$

with $\alpha > d/2$ for trace-class membership of the continuum embedding (on the finite lattice the kernel is strictly positive definite for any $\alpha > 0$, Lemma 1; sensitivity is ablated in Appendix G.6); in coordinates this is the diagonal weight $\mathbf{W} = \mathrm{diag}(w_\alpha(\mathbf{k}_i))$. To keep sparse high-frequency sets from vanishing in norm, indicators are projected onto the unit sphere:

$$\tilde{\mathbf{z}} = \frac{\mathbf{W}^{1/2}\mathbf{z}}{\|\mathbf{W}^{1/2}\mathbf{z}\|_2} \in \mathbb{S}^{D-1} \tag{8}$$

where $[\mathbf{W}^{1/2}]_{m,m} = (1 + \|\mathbf{k}_m\|_2^2)^{-\alpha/2}$. We define the squared chordal spectral distance and the **Physics-Informed Spectral Set Kernel** as:

$$d^2_{\mathrm{spec}}(\mathbf{z}_A, \mathbf{z}_B) = \|\tilde{\mathbf{z}}_A - \tilde{\mathbf{z}}_B\|_2^2 = 2\left(1 - \langle \tilde{\mathbf{z}}_A, \tilde{\mathbf{z}}_B\rangle\right), \quad k_{\mathrm{PI}}(\mathbf{z}_A, \mathbf{z}_B) = \sigma_f^2 \exp\left(-\frac{d^2_{\mathrm{spec}}(\mathbf{z}_A, \mathbf{z}_B)}{2\ell^2}\right) \tag{9}$$

Strict positive definiteness of k_{PI} is immediate (Lemma 1, Appendix C).

4.2 THE ADMISSIBLE POOL AND INFORMATION GAIN

Acquisition uncertainty is governed by the maximum information gain (Srinivas et al., 2012):

$$\gamma_T = \max_{A\subset\{0,1\}^D, |A|=T} \frac{1}{2}\log\det\left(\mathbf{I} + \sigma_n^{-2}\mathbf{K}_A\right) \tag{10}$$

Physical spectra decay harmonically, which we exploit through downward-closed (staircase) restrictions:

Theorem 2 (Structural Space Reduction and Information Gain on Downward-Closed Sets)**.** *Let $\Omega_{\mathrm{cand}} = [-k_{\max}, k_{\max}]^d \cap \mathbb{Z}^d$ be a candidate lattice of $D = (2k_{\max}+1)^d$ modes. Restrict the search domain to downward-closed frequency sets: if $\mathbf{k} \in \mathcal{K}$ and $|\mathbf{k}'| \leq |\mathbf{k}|$ componentwise, then $\mathbf{k}' \in \mathcal{K}$. Then:*

(i) ***Combinatorial Reduction:*** *Admissible downward-closed configurations number $\exp(\Theta(k_{\max}^{d-1}))$, against $2^{\Theta(k_{\max}^d)}$ unconstrained; for $d = 2$, $2^{\Theta(k_{\max}^2)} \to 2^{\Theta(k_{\max})}$.*

(ii) ***Information Gain:*** *Under algebraic Gram eigenvalue decay $\lambda_j \leq c_0 j^{-\beta}$ with $\beta > 1$ of kernel k_{PI} on the admissible candidate pool, the maximum information gain after T iterations satisfies:*

$$\gamma_T = \mathcal{O}\left(T^{\frac{1}{\beta}}(\log T)^{1-\frac{1}{\beta}}\right) \tag{11}$$

4.3 CALIBRATED TWO-STAGE MULTI-FIDELITY ACQUISITION

Fidelity is the continuous pair $\mathbf{m} = (m_{\mathrm{res}}, m_{\mathrm{epoch}}) \in [0,1]^2$, linearly parameterizing spatial nodes $N(m_{\mathrm{res}})$ and training epochs $E(m_{\mathrm{epoch}})$. Target full fidelity corresponds to $\mathbf{m}^* = (1,1)$.

By Proposition 1, the computational cost of training and evaluating a DSE operator with active mode indicator $\mathbf{z}$ at fidelity $\mathbf{m}$ is affine in the work terms:

$$C(\mathbf{z}, \mathbf{m}) = C_{\text{base}} + (c_1 \|\mathbf{z}\|_0 + c_2 d_v)\, N(m_{\text{res}})\, d_v\, E(m_{\text{epoch}}), \tag{12}$$

with c_2 the mode-independent pointwise work of Proposition 1 (measured at 31–32% of a pass, Section 5.5); at fixed $\|\mathbf{z}\|_0 = M$ only the fidelity coordinates vary the cost. We model the joint function $f(\mathbf{z}, \mathbf{m})$ via a multi-fidelity Gaussian Process $\mathcal{GP}(0, k_{\text{MF}})$ with composite covariance:

$$k_{\text{MF}}((\mathbf{z}, \mathbf{m}), (\mathbf{z}', \mathbf{m}')) = k_{\text{PI}}(\mathbf{z}, \mathbf{z}') \cdot k_{\text{fid}}(\mathbf{m}, \mathbf{m}') \tag{13}$$

where k_{fid} is a Matérn-5/2 product kernel over fidelity coordinates. A continuous separable product covariance $k_{\text{PI}} \times k_{\text{fid}}$ is adopted because resolution and epochs are continuous coordinates (discretized to three tiers in deployment), letting one Matérn lengthscale per coordinate replace the per-level couplings of level-wise autoregressive constructions.

We adopt the bias-thresholded fidelity escalation of MF-GP-UCB (Kandasamy et al., 2016): the query point is chosen by an upper confidence bound and the fidelity by comparing posterior uncertainty against that fidelity's bias. (Cost-*divided* acquisitions are unsound here, since the objective $-\log \mathcal{L}_{\text{val}}$ is negative; an earlier version of this work used one and it is superseded by the two-stage rule below.) Our **two-stage** instantiation differs in scoring Stage 1 strictly at target fidelity and in using the continuous fidelity pair $(m_{\text{res}}, m_{\text{epoch}})$.

Stage 1 selects the configuration maximizing the UCB at target fidelity $\mathbf{m}^* = (1, 1)$:

$$\mathbf{z}_t = \arg\max_{\mathbf{z} \in \mathcal{Z}_{\text{cand}}} \left[\mu_{t-1}(\mathbf{z}, \mathbf{m}^*) + \beta_t \sigma_{t-1}(\mathbf{z}, \mathbf{m}^*)\right] \tag{14}$$

Stage 2 selects the cheapest admissible fidelity, using the bias bound $\zeta(\mathbf{m}) \geq |f(\mathbf{z}, \mathbf{m}) - f(\mathbf{z}, \mathbf{m}^*)|$:

$$\zeta(\mathbf{m}) = c_{\text{res}}(1 - m_{\text{res}})^{\kappa_1} + c_{\text{epoch}}(1 - m_{\text{epoch}})^{\kappa_2} \tag{15}$$

where $c_{\text{res}}, c_{\text{epoch}} > 0$ and $\kappa_1, \kappa_2 \geq 1$ bound the discretization error $\mathcal{O}(N(m_{\text{res}})^{-r/d})$ and optimization convergence error $\mathcal{O}(E(m_{\text{epoch}})^{-1})$, with $\kappa_1 = \kappa_2 = 1$; $c_{\text{res}} = c_{\text{epoch}}$ is fitted by least squares to anchor evaluations of two configurations at every tier at the start of each search, charged to the search budget (Appendix F.9; the fits give $c \approx 0.03$–0.2 across domains — an operational surrogate whose bound we verify on the deployed ladder, not a certified uniform bound; see F.7 for a measured off-ladder violation). A plausible fixed constant is not usable: with the $c = 0.8$ of earlier drafts the admission test below never fires and the policy silently degenerates to single fidelity. We then select the cheapest admissible tier from the finite deployed ladder $\mathcal{M} = \{(0.5, 0.6), (0.75, 0.8), \mathbf{m}^*\}$ (fidelity is continuous in principle; deployment discretizes it):

$$\mathbf{m}_t = \arg\min_{\mathbf{m} \in \mathcal{M}} C(\mathbf{z}_t, \mathbf{m}) \quad \text{s.t.} \quad \zeta(\mathbf{m}) \leq \gamma \beta_t \sigma_{t-1}(\mathbf{z}_t, \mathbf{m}), \tag{16}$$

with C from Equation 12; $\zeta(\mathbf{m}^*) = 0$, so $\mathbf{m}^*$ is always admissible and serves as the fallback.

4.4 Search Spaces, Algorithm, and Guarantees

Stage 1 runs over two search spaces, reported throughout. **(a) Downward-closed (enumerative).** A symmetric downward-closed set in $d = 2$ is a non-increasing staircase $b : \{0, \ldots, k_{\max}\} \to \{0, \ldots, k_{\max}+1\}$ with $\mathbf{k}$ active iff $|k_y| < b(|k_x|)$; closure holds by construction. At fixed cardinality the admissible pool is tiny — 42 sets at M=27, 907 at M=145, 9,825 at M=221 — so $\mathbf{z}_t$ is the *exact* maximizer of the acquisition, closing the oracle-versus-implementation gap that would otherwise void Theorem 3. **(b) Unconstrained.** Over $\{0, 1\}^D$, candidates come from Sobolev-weighted mutations ($p_i \propto (1 + \|\mathbf{k}_i\|_2^2)^{-\alpha}$) and pairwise exchange (Algorithm 2), each projected onto $\|\mathbf{z}\|_0 = M$. Since the pool is enumerable, the hard part is *evaluation cost*, not enumeration — training all 907 parity staircases would cost GPU-days, which is what the multi-fidelity surrogate is for. The full loop is summarized as Algorithm 1 in Appendix F.

We define simple regret at target fidelity after T iterations as $r_{\min}(T) = f(\mathbf{z}^*, \mathbf{m}^*) - \max_{t \leq T} f(\mathbf{z}_t, \mathbf{m}^*)$.

Theorem 3 (Simple Regret and Target Fidelity Sample Complexity of MF-SpecBO)**.** *Let the joint objective satisfy $f \in \mathcal{H}_{k_{\text{MF}}}$ with $\|f\|_{\mathcal{H}_{k_{\text{MF}}}} \leq B$ (since $k_{\text{fid}}(\mathbf{m}^*, \mathbf{m}^*) = 1$, the target slice $f(\cdot, \mathbf{m}^*)$ then has k_{PI}-norm at most B; this joint assumption is what Lemma 3 uses), observation noise be σ_n-sub-Gaussian, fidelity approximation bias satisfy $|f(\mathbf{z}, \mathbf{m}) - f(\mathbf{z}, \mathbf{m}^*)| \leq \zeta(\mathbf{m})$ via Equation 15,*

and cross-fidelity coupling satisfy $k_{\text{fid}}(\mathbf{m}, \mathbf{m}^*) \geq \rho_{\min} > 0$. *Under the two-stage acquisition policy (Equations 14–16) with Stage-2 threshold* $\gamma \in (0, 1]$ *(distinct from the information gain* γ_T*), with probability at least* $1 - \delta$*:*

(i) ***Target-Fidelity Query Complexity:*** *To reach simple regret* ε *at target fidelity, the number of target-fidelity queries* T_* *obeys*

$$T_* \leq \mathcal{O}\left(\frac{\beta_T^2 \gamma_T}{\varepsilon^2}\right) \tag{17}$$

(ii) ***Finite Escalation & No Starvation:*** *Each pair* $(\mathbf{z}, \mathbf{m})$ *with* $\mathbf{m} \neq \mathbf{m}^*$ *is admissible for at most* $K_{\max}(\mathbf{m}) = \lceil \gamma^2 \beta_T^2 \sigma_n^2 / \zeta(\mathbf{m})^2 \rceil$ *queries. Since Stage 1 enumerates the finite admissible pool of Theorem 2(i), the total number of non-target queries is bounded by* $|\mathcal{D}_d(k_{\max})| \sum_{\mathbf{m} \neq \mathbf{m}^*} K_{\max}(\mathbf{m}) < \infty$*, so target-fidelity queries occur infinitely often and* $\lim_{C_{\text{total}} \to \infty} r_{\min}(T) = 0$.

(iii) ***Cross-Fidelity Transfer and its Ceiling:*** n *queries of* $(\mathbf{z}, \mathbf{m})$ *contract target variance to* $\sigma_f^2[1 - \rho(\mathbf{m})^2 n \sigma_f^2 / (\sigma_n^2 + n\sigma_f^2)]$ *— strictly decreasing, but saturating at* $\sigma_f^2(1 - \rho(\mathbf{m})^2)$*: coarse queries help at a rate set by* $\rho_{\min}$ *yet cannot replace target queries below that floor.*

Proof: Appendix D. **Remark (asymptotic theory vs. finite search).** With three fidelity tiers and $T \leq 30$ iterations, these guarantees motivate the policy's design rather than certifying its finite-sample behaviour: part (ii) is a statement about the limit $C_{\text{total}} \to \infty$, and the bound it gives on non-target queries is astronomically larger than any budget we run. Two hypotheses deserve flagging: under the theoretical schedule $\beta_T = \Theta(\sqrt{\gamma_T})$, the simple-regret rate of Appendix D vanishes only when the eigenvalue-decay exponent satisfies $\beta > 2$ (not merely the $\beta > 1$ of Theorem 2) — the measured $\beta = 3.42$ clears this, and deployed runs use the constant $\beta_t = 2.0$ of Appendix F.13, for which the same conclusion needs only $\gamma_T = o(T)$. We therefore report the *measured* fidelity allocation (Section 5.5) rather than inferring it from the theory, and check part (i) against the recorded traces in Appendix F.12.

4.5 A PAC-Bayes Complexity Surrogate

Let $S = \{(a_j, u_j)\}_{j=1}^{N_{\text{data}}} \sim \mathcal{D}^{N_{\text{data}}}$, with $\mathcal{L}_{\mathcal{D}}$ and $\widehat{\mathcal{L}}_S$ the true and empirical squared L^2 risks of $\mathcal{G}_{\mathcal{K}}$. The construction that makes the bound informative about *which modes are kept* is a prior and posterior that share the anisotropic covariance: a data-independent **Sobolev prior** $P_{\mathcal{K}} = \mathcal{N}(\mathbf{0}, \sigma_p^2 \mathbf{W}_{\mathcal{K}} \otimes \mathbf{I})$ with $[\mathbf{W}_{\mathcal{K}}]_{mm} = (1 + \|\mathbf{k}_m\|_2^2)^{-\alpha}$, and a matched posterior $Q_{\mathcal{K}} = \mathcal{N}(\mathbf{R}, \sigma_q^2 \mathbf{W}_{\mathcal{K}} \otimes \mathbf{I})$ centred on the learned multipliers.

Matching the covariances cancels the mode-dependent determinants exactly, leaving a KL divergence whose only mode-dependent term is a Sobolev norm of the multipliers plus an explicit constant; the formal statement — a finite-sample bound on the *Gibbs* risk with $C_0 = L\, d_v^2 |\mathcal{K}|\, (\sigma_q^2/\sigma_p^2 - 1 - \log(\sigma_q^2/\sigma_p^2))$ — is Theorem 4, stated and proved in Appendix E.

Remark. Theorem 4 bounds the *Gibbs* risk; passing to the deterministic trained operator costs a derandomization step (Appendix E) that forces $\sigma_q^2 = \mathcal{O}(N_{\text{data}}^{-1})$. The bound is then *numerically vacuous* at our scale — $C_0 \sim L\, d_v^2 |\mathcal{K}| \log(\sigma_p^2/\sigma_q^2)$ reaches $\mathcal{O}(10^6)$ at d_v=32, L=4 — and its risks are nodal-quadrature risks on a fixed node set, not continuum L^2 risks. It is used only as a training-free *ranking* signal; a surrogate Ω_α^+ can enter the Stage-1 acquisition with weight λ, but this breaks Theorem 3's maximality step, so headline runs use $\lambda = 0$ (construction and null ablation: Appendix F.10; proof: Appendix E). Computed per trained configuration in the campaign, the bound is vacuous as anticipated and its rank correlations with measured risk are weak to negative at these budgets — reported without further claim in Figure 6 and released with the artifacts.

5 Experiments and Empirical Validation

Our empirical section is organized around one question — *does searching over mode sets buy anything that a cheap heuristic does not?* — and structured so the answer cannot be hidden: one protocol (§5.1), the decision comparison at matched cardinality (§5.2–§5.4), the cost side and search behaviour (§5.5), and the failure regimes with diagnostics (Table 4).

Table 1: **Decision gate: does informed search beat a one-pass heuristic at matched cardinality?** Test relative ℓ_2 error (%), same protocol as Table 2; every column fixes $|\mathcal{K}|$ so the comparison isolates mode *identity*. * marks a paired two-tailed t-test against Top-M spectral energy with $p < 0.05$. The last row gives the mean paired difference of MF-SpecBO against that heuristic with a 95% CI (negative favours MF-SpecBO). Asterisks mark $p < 0.05$ regardless of magnitude: at Darcy parity the significant difference is 0.005 points (the two sets share 127 of 145 modes).

Selection method	Darcy M=145	Darcy M=27	Airfoil M=221	Airfoil M=47	Elasticity M=221	Elasticity M=61
MF-SpecBO (downward-closed)	5.54*	7.87*	3.74	4.71	16.68*	18.64
MF-SpecBO (unconstrained)	7.02*	6.91*	3.83	5.05	16.05	19.57
Top-M spectral energy	5.53	5.87	3.82	4.96	15.92	19.61
Isotropic box (matched $\|\mathcal{K}\|$)	5.60	5.78	3.64*	4.61*	15.85	18.40*
iFNO-style concentric growth	5.43	6.29	2.83*	3.57*	16.58	17.07*
Magnitude pruning (fine-tuned)	5.26	6.43	3.54*	4.72	26.66*	35.24*
Learnable gating (L1, fine-tuned)	8.65	9.39	5.53*	5.98*	26.86*	37.27*
Random matched	10.79*	14.34*	4.39*	5.67	20.39*	24.75*
Δ MF-SpecBO (downward-closed) − energy [95% CI]	0.00 [0.00, 0.01]	2.00 [1.73, 2.27]	-0.08 [-0.16, 0.00]	-0.25 [-0.57, 0.07]	0.76 [0.27, 1.25]	-0.98 [-1.99, 0.04]

5.1 Protocol and Reproducibility Discipline

One protocol produces every number: 100 epochs, $N_{\text{train}} = 800$ / $N_{\text{test}} = 200$, batch 20, AdamW (2×10^{-3}, wd 10^{-4}, cosine), 5 seeds, RTX 4070 Laptop GPU; latency = CUDA events, 20 warmups, median of 50 passes; searches see only a validation split disjoint from all test data, and discovered sets are retrained from scratch. Two safeguards: all selection rules within a column use *exactly* equal $|\mathcal{K}|$ (set intersections logged), and every table in this paper is generated directly from the recorded measurement files, with every measured value in them — and the numerals quoted in the abstract and the headline prose — bound to a machine-readable claim registry that a build-time check resolves against those files, failing the build on any disagreement (Reproducibility Statement; surrogate hyperparameters: Appendix F.13).

5.2 The Decision Question: Does Search Beat a One-Pass Heuristic?

Table 1 compares seven selection rules at matched $|\mathcal{K}|$: both MF-SpecBO variants, Top-M spectral energy (one pass, no training), the matched-budget isotropic box (what a conventional FNO does), iFNO-style concentric growth (George et al., 2024), fine-tuned magnitude pruning, and a random symmetric subset.

The result is negative for the paper's premise, and stated first. **No selection rule beats the isotropic box's mode set in any of the six settings**: among methods choosing *different* wave vectors the significant gaps run the other way (search worse than energy at Darcy compression, $+2.00$ $[1.73, 2.27]$, and elasticity parity, $+0.76$); random fails outright, magnitude pruning collapses on irregular geometry ($+10.8$ to $+16.8$ on elasticity), and single-run L1 gating — the continuous-relaxation family, one untuned λ (Appendix F.8) — trails the box in every column, by up to $+18.87$ points on elasticity, with several-fold larger seed variance. The one improvement — iFNO-style growth, ahead of the box in four columns and significantly so vs the box in three (-0.81 to -1.33, $p \leq 0.0075$), on the airfoil even beating the *dense* reference at half the modes (2.83 vs 3.29 at parity) — trains the *same* radial set under a growth curriculum (schedule in Appendix F.8), sharpening the conclusion: the available win is in training, not mode identity. The search machinery buys automation, not accuracy (frontier and error maps: Appendices F.5, G.1).

5.3 Why: Measuring the Anisotropy the Method Is Designed to Exploit

Theorem 1 makes Table 1 a prediction rather than an exploration: measured on the training solutions (Appendix F.3), Darcy has $\Delta = 0.0038$, capping any selection gain below one percent at any budget — structurally so, since its coefficient is a thresholded isotropic Gaussian field — while the airfoil and elasticity have $\Delta = 0.2072$ and 0.3706 but realize only $G \leq 0.056$. The standard benchmarks cannot discriminate isotropic from anisotropic truncation, and one pass over the data says so before any model is trained.

Table 2: **Cross-domain cost and accuracy at two operating points.** One protocol: 100 epochs, $N_{\text{train}} = 800/N_{\text{test}} = 200$, batch 20, 5 seeds, NVIDIA GeForce RTX 4070 Laptop GPU; latency = CUDA events, 20 warmups, median of 50. Δ: mean paired difference vs dense [95% CI], positive is worse. Mode sets from the downward-closed enumerative search (Section 4.4).

Domain	Configuration	$\lvert\mathcal{K}\rvert$	Mode cut	Params	Test ℓ_2 (%)	Δ vs dense [95% CI]	Latency (ms)	Speedup
2D Darcy Flow	Dense Cartesian (reference)	289	—	1,192,449	5.31 ± 0.38	—	9.78	1.00×
	Isotropic box @ parity	145	49.8%	602,625	5.60 ± 0.94	0.29 [-1.09, 1.66]	6.72	1.46×
	MF-SpecBO @ parity	**145**	49.8%	602,625	5.54 ± 0.90	0.22 [-1.10, 1.54]	6.72	**1.46**×
	Isotropic box @ compression	27	90.7%	119,297	5.78 ± 0.20	0.47 [-0.00, 0.94]	3.66	2.68×
	MF-SpecBO @ max. compression	**27**	90.7%	119,297	7.87 ± 0.29	2.56 [2.13, 2.98]	3.65	**2.68**×
2D Transonic Airfoil	Dense Cartesian (reference)	441	—	1,815,009	3.29 ± 0.06	—	31.28	1.00×
	Isotropic box @ parity	221	49.9%	913,889	3.64 ± 0.13	0.35 [0.22, 0.48]	20.68	1.51×
	MF-SpecBO @ parity	**221**	49.9%	913,889	3.74 ± 0.09	0.46 [0.34, 0.58]	20.82	**1.50**×
	Isotropic box @ compression	47	89.3%	201,185	4.61 ± 0.34	1.33 [0.92, 1.73]	12.09	2.59×
	MF-SpecBO @ max. compression	**47**	89.3%	201,185	4.71 ± 0.35	1.43 [1.03, 1.83]	12.00	**2.61**×
2D Structural Elasticity	Dense Cartesian (reference)	441	—	1,815,009	14.87 ± 0.88	—	3.33	1.00×
	Isotropic box @ parity	221	49.9%	913,889	15.85 ± 0.38	0.98 [-0.38, 2.33]	3.27	1.02×
	MF-SpecBO @ parity	**221**	49.9%	913,889	16.68 ± 0.41	1.81 [0.28, 3.33]	3.05	**1.09**×
	Isotropic box @ compression	61	86.2%	258,529	18.40 ± 1.09	3.53 [1.78, 5.29]	3.29	1.01×
	MF-SpecBO @ max. compression	**61**	86.2%	258,529	18.64 ± 0.93	3.77 [2.42, 5.12]	3.25	**1.02**×

Table 3: **Does anisotropy change the answer?** Test relative ℓ_2 error (%), anisotropic Darcy flux family, budget $|\mathcal{K}| = 27$ of 289; 100 epochs, 5 seeds; matched $|\mathcal{K}|$ in every non-dense row; *: paired $p < 0.05$ vs the isotropic box; last rows: mean paired differences [95% CI], negative favours MF-SpecBO. The two right-hand columns repeat the comparison at the parity budget (M=145, $\kappa \in \{1, 4\}$). The flux target is itself directional, so $\kappa = 1$ is not a zero-headroom control (the standard benchmarks and the diffusion control's $G = 0$ level are).

Selection method	κ=1 (G=0.22)	κ=4 (G=0.21)	κ=16 (G=0.33)	κ=64 (G=0.42)	κ=1, M=145 (G=0.11)	κ=4, M=145 (G=0.17)
Dense reference ($\lvert\mathcal{K}\rvert$=289)	17.91	18.30	18.95	19.17	17.91	18.30
Isotropic box (reference)	51.37	53.75	48.58	38.08	25.51	32.07
Top-M spectral energy	47.80*	52.43	48.71	38.74	27.78*	32.77
MF-SpecBO	39.82*	42.74*	43.73*	38.64	28.70*	27.49*
iFNO-style growth (box set)	59.69*	59.41	50.60*	39.73	33.13*	38.03
Magnitude pruning (fine-tuned)	83.12*	79.89*	73.22*	69.06*	58.75*	58.87*
Learnable gating (L1, fine-tuned)	64.76*	66.46*	59.64*	50.88*	31.89*	33.55
Exhaustive pool @ coarse tier	50.49	49.17*	44.86*	39.61	—	—
Random matched	56.61*	53.18	46.32	40.74	30.27*	33.07
Δ MF-SpecBO − isotropic [95% CI]	-11.55 [-13.86, -9.25]	-11.01 [-13.28, -8.74]	-4.85 [-7.96, -1.73]	0.56 [-2.30, 3.42]	3.20 [1.59, 4.80]	-4.58 [-7.83, -1.32]
Δ MF-SpecBO (DC) − energy [95% CI]	-7.98 [-10.83, -5.12]	-9.70 [-12.27, -7.12]	-4.98 [-9.09, -0.87]	-0.10 [-3.65, 3.44]	0.93 [-1.36, 3.21]	-5.28 [-7.46, -3.10]

5.4 When Does Mode Selection Pay? A Controlled Anisotropy Sweep

We test the premise on a real PDE: anisotropic Darcy flow $-\nabla \cdot (a\,\mathbf{M}_\kappa \nabla u) = f$ on the torus, log-normal $a = \exp(\psi)$, $\mathbf{M}_\kappa = \mathbf{R}(30°)\,\mathrm{diag}(1, 1/\kappa)\mathbf{R}(30°)^\top$ (bedding-plane permeability; κ=1 is ordinary Darcy; residual 10^{-8}). The operator maps a to a flux component of $\mathbf{q} = -a\mathbf{M}_\kappa \nabla u$ (pressure is $1/\|\mathbf{k}\|^2$-damped: paper-only headroom) and the 30° rotation keeps the ridge off the lattice axes. Sweeping $\kappa \in \{1, 4, 16, 64\}$ moves Δ from 0.71 (κ=1) to 1.27 (κ=64) and $G(27)$ from 0.22 to 0.42 (non-monotonically: 0.21 at κ=4), several times the largest benchmark headroom of 0.056; each level re-runs the identical matched-cardinality comparison, with a closed-form rotated-diffusion control swept alongside.

The outcome (Table 3) has two parts. **Where headroom exists, the search pays**: at the compression budget (M=27), $\kappa \in \{1, 4, 16\}$, MF-SpecBO's paired difference vs the isotropic box is -11.55/-11.01/-4.85 points ($p = 0.0002, 0.0002, 0.012$) and vs the energy heuristic -7.98/-9.70/-4.98 ($p \leq 0.028$); with the κ=4 parity column below, these are the settings in this paper where Bayesian search adds value beyond one pass over the data, exactly where Theorem 1 says there is something to win. Attribution: training only 3–10 *distinct* sets of 42, the search beats exhaustive coarse-tier evaluation of all 42 (exhaustive row) — the win is target-fidelity surrogate modeling, not pool smallness. **And the advantage is not monotone in headroom**: it decays with κ and at $\kappa = 64$ — largest G — all rules collapse together ($p = 0.61$), far from dense. The mechanism is sample complexity, and a controlled sweep isolates it (Appendix F.7): with mode sets frozen and only N_{train} varied ($400 \to 1600$), the energy set's advantage over the box grows monotonically ($-1.6 \to -5.8$ points, all significant), while the searched set's does not (-0.16, $p = 0.75$ at 1600) — its modes were themselves identified from 200-sample evaluations. Headroom is necessary, not sufficient — Δ screens for free, and realizing it needs the data budget to support both *finding* and *fitting* the modes that carry it. At the parity budget M=145 (right-hand columns) the headroom shrinks as Theorem 1 predicts ($G = 0.11/0.17$), and the realized outcomes split: significantly better than the box at κ=4 (-4.58, p=0.017) — a further setting where search pays — but significantly *worse* at κ=1 (3.20 points, p=0.005). The theorem only caps

Regime	What happens	Diagnostic (no full training)
Near-radial spectrum	no rule beats the isotropic box (Thm. 1)	Δ, $G(M)$ small
Generous budget	selection rules collapse within noise	$G(M)$ at the actual budget
Extreme anisotropy	winning modes exceed the sample budget	G concentrated at high $\|\mathbf{k}\|$ (App. F.7)
Off-axis ridge (DC variant)	staircases cannot represent it; search returns the box	best-DC energy capture vs. top-M
One spatial dimension	selection degenerates to a bandwidth choice	$d = 1$ (App. G.5)
Small mesh ($N \lesssim 10^3$; batch 20, d_v=32, our GPU)	launch-bound: no latency win, and coarse tiers retain most of a full evaluation's cost	fixed-cost fraction; tier-cost ratio (App. F.9)
Map unlearnable at search scale	posterior flat; search only matches the energy heuristic	anchor-tier error vs untrained baseline (App. F.4)

Table 4: **Failure modes and pre-training diagnostics.**

the gain; the loss is an empirical failure mode — too little headroom left to absorb surrogate-induced suboptimality (generous-budget row, Table 4).

The strongest Table-1 baselines transfer poorly here: iFNO-style growth lands 6–8 points *behind even the plain box* (new rows of Table 3), and a full 2×2 shows the curriculum hurting *both* sets on this family (box $+8.32$, p=0.005; searched set $+10.58$, p=0.009 at κ=1; Appendix F.6). Curriculum and selection are regime-complements, not a stack: growth curricula win where the spectrum is radial and mislead where it is not.

The diffusion control (Table 7) supplies what the primary family cannot: an exact G=0 level realizing the theorem's boundary case (energy and isotropic tie to 0.00 points; only random collapses), and a separation of failure modes at higher levels — energy realizes up to -10.25 points over isotropic (p=0.011, κ=32 column of Table 7) while the downward-closed search cannot leave the radial family: *no axis-aligned staircase represents a* $30°$ *ridge*, so its best admissible set is box-equivalent (the starred $+0.06$ at κ=1 is TOST-equivalent at the 0.5-point margin, $p_{\mathrm{TOST}} < 10^{-6}$). Enumerability thus has a representational price off-axis, and the unconstrained variant *does* escape it: at the ridge levels it beats the box by -10.42 and -9.95 points ($\kappa = 16, 32$; $p \leq 0.001$), drawing statistically level with the energy heuristic ($+0.30$, $p = 0.64$ at κ=32) — escape recovers the heuristic's accuracy, not more — while paying a stochastic-search cost where there is nothing to find ($+2.53$ at κ=1). (Darcy flux is won because ∂_x pushes energy along a lattice axis.) Applied in the wild, the screen finds one real dataset with genuine headroom — transonic channel shocks ($\Delta = 0.9036$, $G(27) = 0.18$) — and the gate on it (Appendix F.4) yields the first real-data win of an informed set over the box (energy -2.18, $p = 0.016$) and a further failure mode: unlearnable at search-scale evaluations, so the search only matches the heuristic (Table 4).

5.5 Cost, Latency, and Search Behaviour

Table 2 reports the cost side: parameters fall as $Ld_v^2|\mathcal{K}|$ exactly and latency with the $N|\mathcal{K}|d_v$ term of Proposition 1, so speedups sit below the mode-cut ratio, governed by the fixed-cost fraction of a pass ($\approx 32\%/31\%$ on Darcy/airfoil, affine fit). Idle-GPU re-measurement (ten harness runs per configuration, released with the artifacts) corrected an earlier $3.6\times$ elasticity claim — that mesh is *launch-bound*: fixed costs are $\approx 98\%$ of a forward pass at $N = 972$, every budget times within a few percent of dense, and the earlier readings reflected load-boosted clocks. Speedups are therefore quoted only where real ($N \geq 3721$); at small N, mode cuts buy parameters and memory, not latency. The boundary is harness-conditioned (batch 20, d_v=32, this GPU/framework). Scaling across $N \in [10^3, 3.2{\times}10^4]$ and zero-shot transfer (the sparse model tracks dense, keeping its speedup) are profiled in Appendices G.2–G.4.

Figure 3 (Appendix F.5) shows the searches: best-so-far target error vs. cost, and the measured tier allocation — an *outcome* of the calibrated Stage-2 rule (Equation 40). Multi-fidelity pays only where a coarse tier is genuinely cheap (Appendix F.9); the downward-closed restriction costs nothing measurable against the unconstrained variant (four of six settings equal or better).

6 Conclusion, Limitations, and Outlook

MF-SpecBO comes with a limit theorem that says in advance whether mode selection can matter: on standard benchmarks Δ is tiny and, as predicted, no selection rule beats the isotropic box (compression pays; search does not); on a flux family passing the screen the search beats box and heuristic alike, losing where winning modes outrun the sample budget; on the one real dataset passing it, selection first beats the box — via the heuristic (Appendix F.4). Protocol: measure Δ and $G(M)$ at the intended budget; box unless $G(M)$ is substantial (observed wins at ≥ 0.17, failures ≤ 0.11; a range, not a threshold); energy as the default, search only where search-scale evaluations carry signal (Table 4).

Reproducibility Statement

Proofs of Theorems 2, 3 and 4 are in Appendices C–E, algorithmic specifications in Appendix F, and datasets, hyperparameters and compute in Appendices F–G.

The supplementary code package is organized so that the manuscript can be audited rather than merely re-run. (i) `experiments/run_mode_discovery.py` performs the searches and records, for every iteration, the queried mode set, fidelity, validation error and wall-clock cost. (ii) `experiments/run_final_campaign.py` evaluates every reported configuration under the single protocol of Section 5.1 and writes per-seed errors, parameter counts, measured latencies and paired statistics to `results/final_campaign.json`. (iii) `experiments/make_tables.py` generates every numerical table in this paper — Tables 1, 2, 3, 5, 6, 7, 8, 9 and 10; Table 4 contains no measurements — together with every prose numeral macro, directly from the results files: no measured number in a table is typed by hand. (iv) `experiments/verify_claims.py` resolves every `\claim` macro in the source against `paper/claims.json` and exits non-zero if any printed value disagrees with its recorded measurement. All benchmark datasets are public (Darcy and Burgers from Li et al. (2021); airfoil, elasticity and channel meshes from the Geo-FNO release of Li et al. (2022)); the two anisotropic families of Section 5.4 are generated by the released code with fixed seeds (Appendix F.2).

Ethics Statement

This paper develops algorithmic foundations for spectral architecture search in neural operators applied to physical PDE systems (aerodynamics, fluid dynamics, and structural mechanics). The research utilizes public synthetic and benchmark PDE datasets. The authors anticipate no negative societal impacts, human subject considerations, privacy risks, or ethical conflicts associated with this work.

AI Use Statement

Generative AI tools were used as a coding and drafting assistant during this work: they contributed to implementing and running the experiment scripts released with the paper, to analysis of the resulting logs, and to drafting and revising prose. The research questions, the experimental design decisions, the mathematical statements and their proofs, and every scientific interpretation are the authors' own, and all AI-assisted output was reviewed and verified by the authors, who take full responsibility for the content and claims of this paper. To make that verification checkable rather than asserted, the reported numbers are machine-resolved against the recorded measurements (see the Reproducibility Statement).

References

Robert A Adams and John JF Fournier. *Sobolev Spaces*. Elsevier, 2003.

Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends in Machine Learning*, 17(2):174–303, 2024.

Maximilian Balandat, Brian Karrer, Daniel Jiang, Samuel Daulton, Benjamin Letham, Andrew G Wilson, and Eytan Bakshy. BoTorch: A framework for efficient Bayesian optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 21524–21538, 2020.

Ricardo Baptista and Matthias Poloczek. Bayesian optimization of combinatorial structures. In *International Conference on Machine Learning (ICML)*, pp. 462–471, 2018.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning (ICML)*, pp. 844–853, 2017.

Aryan Deshwal, Syrine Belakaria, and Janardhan Rao Doppa. Mercer features for efficient combinatorial Bayesian optimization. In *AAAI Conference on Artificial Intelligence*, volume 35, pp. 7210–7218, 2021.

Stefan Falkner, Aaron Klein, and Frank Hutter. BOHB: Robust and efficient hyperparameter optimization at scale. In *International Conference on Machine Learning (ICML)*, pp. 1437–1446, 2018.

Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. GPyTorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pp. 7576–7586, 2018.

Robert Joseph George, Jiawei Zhao, Jean Kossaifi, Zongyi Li, and Anima Anandkumar. Incremental spatial and spectral learning of neural operators for solving large-scale PDEs. *Transactions on Machine Learning Research (TMLR)*, 2024. arXiv:2211.15188.

John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Adaptive Fourier neural operators: Efficient token mixers for transformers. In *International Conference on Learning Representations (ICLR)*, 2022.

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

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *Advances in Neural Information Processing Systems (NeurIPS)*, 28, 2015.

Zhongkai Hao, Chengqiang Ying, Hang Dong, Hang Su, and Jun Zhu. GNOT: A general neural operator transformer for operator learning on arbitrary geometries. In *International Conference on Machine Learning (ICML)*, pp. 12556–12569, 2023.

Kirthevasan Kandasamy, Gautam Dasarathy, Junier B Oliva, Jeff Schneider, and Barnabás Póczos. Gaussian process bandit optimisation with multi-fidelity evaluations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, pp. 992–1000, 2016.

Kirthevasan Kandasamy, Gautam Dasarathy, Jeff Schneider, and Barnabás Póczos. Multi-fidelity Bayesian optimisation with continuous approximations. In *International Conference on Machine Learning (ICML)*, pp. 1799–1808, 2017.

Nikola Kovachki, Samuel Lanthaler, and Siddhartha Mishra. On universal approximation and error bounds for Fourier neural operators. *Journal of Machine Learning Research*, 22(290):1–76, 2021. arXiv:2107.07562.

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

Samuel Lanthaler, Siddhartha Mishra, and George E Karniadakis. Error estimates for DeepONet: A deep neural network framework in operator learning. *Analysis and Applications*, 20(03):435–487, 2022.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. In *Journal of Machine Learning Research*, volume 18, pp. 1–52, 2017.

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

Zongyi Li, Daniel Zhengyu Huang, Burigede Liu, and Anima Anandkumar. Fourier neural operator with learned deformations for PDEs on general geometries. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 27428–27441, 2022.

Levi Lingsch, Mike Y Michelis, Emmanuel de Bézenac, Sirani M Perera, Robert K Katzschmann, and Siddhartha Mishra. Beyond regular grids: Fourier-based neural operators on arbitrary domains. In *International Conference on Machine Learning (ICML)*, 2024.

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.

David A McAllester. PAC-Bayesian model selection. *Machine Learning*, 37(3):355–378, 1999.

David A McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003.

Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. Variational dropout sparsifies deep neural networks. In *International Conference on Machine Learning (ICML)*, pp. 2498–2507, 2017.

Changyong Oh, Jakub M Tomczak, Efstratios Gavves, and Max Welling. Combinatorial Bayesian optimization using the graph Cartesian product. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 2910–2920, 2019.

Matthias Poloczek, Jialin Wang, and Peter Frazier. Multi-information source optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 4288–4298, 2017.

Isaac J Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938.

Jialin Song, Yuxin Chen, and Yisong Yue. A general framework for multi-fidelity Bayesian optimization with gaussian processes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 3158–3167, 2019.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *IEEE Transactions on Information Theory*, 58(5): 3250–3265, 2012.

Richard P Stanley. *Enumerative Combinatorics, Volume 1*. Cambridge University Press, 2 edition, 2011.

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

Sattar Vakili, Neha Kothari, and Andreas Krause. On information gain and regret bounds for Gaussian process bandits. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2296–2304, 2021.

Gege Wen, Zongyi Li, Kamyar Azizzadenesheli, Anima Anandkumar, and Sally M Benson. U-NO: U-shaped neural operators for solving partial differential equations. *Transactions on Machine Learning Research*, 2022.

Haixu Wu, Tengge Hu, Huakun Luo, Jianmin Wang, and Mingsheng Long. Transolver: A fast transformer solver for PDEs on general geometries. In *International Conference on Machine Learning (ICML)*, 2024.

A Proof and Auxiliary Material for Theorem 1

Proof. Write $p = \bar{p} + q$ with $\sum_{\mathbf{k}} q(\mathbf{k}) = 0$ and $\|q\|_1 = \Delta$, so the positive and negative parts of q each sum to $\Delta/2$. Then $\max_{|\mathcal{K}|=M} \sum_{\mathcal{K}} p \le \max_{|\mathcal{K}|=M} \sum_{\mathcal{K}} \bar{p} + \max_{|\mathcal{K}|=M} \sum_{\mathcal{K}} q = \sum_{B_M} \bar{p} + D_M + \max_{|\mathcal{K}|=M} \sum_{\mathcal{K}} q \le \sum_{B_M} \bar{p} + D_M + \Delta/2$, while $\sum_{B_M} p \ge \sum_{B_M} \bar{p} - \Delta/2$. Subtracting gives $G(M) \le \Delta + D_M$. When $\bar{p}$ is non-increasing in radius, a size-M set maximizes $\sum_{\mathcal{K}} \bar{p}$ by filling shells in order of increasing radius, which is B_M up to ties within the boundary shell (on which $\bar{p}$ is constant, so the value is unaffected); hence $D_M = 0$. Non-negativity of G holds because B_M is itself feasible. □

Conventions and measured slack. When M falls inside a shell, B_M is fixed by the code convention of `build_isotropic_modes`: complete shells in order of radius, then whole conjugate pairs from the boundary shell — $\sum_{B_M} \bar{p}$ is tie-invariant, and the reported G uses exactly this set. Strict shell-monotonicity of $\bar{p}$ does *not* hold exactly on our data (empirical spectra have noisy tails), which is why the theorem is stated with the slack: measured D_M is ≤ 0.003 on Darcy, ≤ 0.002 on the airfoil, ≤ 0.015 on elasticity, and at most 0.22 at the most extreme sweep level — small against the Δ values of 0.0038–1.27 (`results/radial_defect_slack.json`, computed by `experiments/compute_radial_defect_slack.py` in float64, where Δ and D_M are numerically stable), so the screen loses essentially nothing.

The second-moment statistic is the wrong one. It is tempting to quantify anisotropy by $\hat{A}$, the eigenvalue ratio of the energy-weighted covariance $\sum_{\mathbf{k}} p(\mathbf{k})\mathbf{k}\mathbf{k}^\top$, and we report it because it is the intuitive measure. It does not control G. Placing energy on $(\pm 1, 0)$ and on $(0, \pm 3)$ in the ratio $9:1$ gives a covariance proportional to the identity, so $\hat{A} = 1$ exactly, while the spectrum is maximally non-radial: $\Delta = 1$ and $G(5) = 0.10$. Second moments are blind to how energy is distributed *within* shells, which is precisely the freedom a mode selector exploits. Δ is no more expensive to compute and is the quantity with a theorem attached.

B Extended Related Work

Neural operators and spectral architectures. FNO (Li et al., 2021) realizes non-local kernel integration by FFT on uniform grids, with universality established by Kovachki et al. (2021) and error bounds by Lanthaler et al. (2022). Extensions relax either the geometry or the cost of the spectral layer: Geo-FNO (Li et al., 2022) learns a deformation to a latent uniform grid; F-FNO (Tran et al., 2023) factorizes the multiplier across dimensions; AFNO (Guibas et al., 2022) shares an MLP across modes; U-NO (Wen et al., 2022) and multiwavelet operators (Gupta et al., 2021) change the multiscale basis; attention-based solvers (Hao et al., 2023; Wu et al., 2024) drop the spectral layer entirely. DSE (Lingsch et al., 2024) replaces the FFT with direct (non-uniform) Vandermonde projections, which is what makes the *identity* of the retained wave vectors — not merely their number — a first-class cost decision, since each excluded mode removes a column of $\mathbf{V}_{\mathcal{K}}$. All of these works retain an isotropic Cartesian truncation.

Adaptive spectral capacity. The closest prior work is iFNO (George et al., 2024), which grows the mode count and the training resolution together during a single run, and is in that sense a multi-fidelity mode selector. It differs from MF-SpecBO in what is adapted: iFNO advances an *isotropic threshold* along a monotone schedule, whereas we search over *sets* and can retain anisotropic, directionally selective configurations. We include iFNO-style concentric growth as a baseline. Magnitude and variational pruning of spectral weights (Han et al., 2015; Molchanov et al., 2017) likewise select modes, but only post hoc from a trained dense model; our comparison includes a fine-tuned magnitude-pruning baseline.

Combinatorial Bayesian optimization. Optimizing over $\{0,1\}^D$ has been addressed by surrogate-based methods including BOCS (Baptista & Poloczek, 2018), COMBO (Oh et al., 2019), and Mercer-feature approximations (Deshwal et al., 2021). These treat coordinates exchangeably; our contribution to this line is a set kernel whose geometry is dictated by Sobolev decay of the underlying PDE solutions, and a structural restriction (downward-closed sets) that shrinks the admissible pool enough to make exact acquisition maximization possible.

Multi-fidelity Bayesian optimization. MF-GP-UCB (Kandasamy et al., 2016) optimizes an expensive objective given cheaper biased approximations by choosing the query point with an upper confidence bound and then escalating fidelity only once posterior uncertainty falls below a threshold set by that fidelity's bias $\zeta^{(m)}$; BOCA (Kandasamy et al., 2017) extends this to continuous fidelity spaces, while information-based acquisitions (Song et al., 2019; Poloczek et al., 2017) and bandit-style resource allocation (Li et al., 2017; Falkner et al., 2018) make the same trade differently. We do not compare against combinatorial acquisition *optimizers* such as COMBO's graph-Cartesian search (Oh et al., 2019) for a structural reason: at fixed cardinality our admissible pool is exactly enumerable, so Stage 1 requires no acquisition optimizer; the single-run continuous relaxation family (learnable gates with an L1 penalty) does answer a distinct question — selection without search — and is compared directly in Tables 1 and 3. **Our Stage 2 is an instantiation of the MF-GP-UCB rule**, not a new acquisition: we adopt its escalation test for a two-dimensional resolution $\times$ epoch fidelity index. What we add is the observation that ζ must be fitted to measured anchor evaluations — Appendix F.9 shows a plausible fixed ζ makes the admission test unsatisfiable, silently reducing the method to single-fidelity search while appearing to run normally.

C Proof of Theorem 2: Structural Space Reduction and Information Gain

In this section, we provide the complete proof of Theorem 2, together with two statements deferred from the main text.

Lemma 1 (Strict Positive Definiteness)**.** *The kernel k_{PI} defined in Equation 9 is strictly positive definite on $\mathcal{P}(\Omega_{\mathrm{cand}})$.*

Proof. $\tilde{\mathbf{z}} \mapsto \mathbb{S}^{D-1} \subset \mathbb{R}^D$ is an isometric embedding, so d^2_{spec} is an exact squared Euclidean norm and $\exp(-\|\cdot\|_2^2/2\ell^2)$ is strictly positive definite on any Hilbert space (Schoenberg, 1938). □

Remark (Applicability to Shocks): Condition $\alpha > d/2$ regulates the surrogate prior, not the physical PDE solution. For convection-dominated problems with sharp shock fronts (e.g., transonic airfoils where $u \in BV$), the Sobolev prior acts as a soft harmonic regularizer: it concentrates budget along directional shock wave vectors while discarding isotropic high-frequency noise.

C.1 Part (i): Combinatorial Reduction on Downward-Closed Sets

Let $\Omega_{\mathrm{cand}} = [-k_{\max}, k_{\max}]^d \cap \mathbb{Z}^d$ with cardinality $D = (2k_{\max}+1)^d$. Hermitian conjugate symmetry alone ($\mathbf{k} \in \mathcal{K} \iff -\mathbf{k} \in \mathcal{K}$) is point reflection through the origin and relates only opposite quadrants; it is the *componentwise* downward-closure condition ($|k_i'| \le |k_i|$ for all i) that additionally forces full coordinate-axis reflection symmetry ($(\pm k_x, \pm k_y) \in \mathcal{K}$ together). Under that stronger symmetry, any admissible mode set is uniquely determined by its intersection with the non-negative quadrant $\Omega_+ = \{0, 1, \ldots, k_{\max}\}^d$.

Definition 1 (Downward-Closed Frequency Sets)**.** *A frequency subset $\mathcal{K}_+ \subset \Omega_+$ is downward-closed (a lower set or staircase set) if:*

$$\mathbf{k} \in \mathcal{K}_+, \quad \mathbf{0} \le \mathbf{k}' \le \mathbf{k} \text{ (componentwise)} \implies \mathbf{k}' \in \mathcal{K}_+ \tag{18}$$

In physical PDEs with continuous spatial operators, the Fourier energy spectrum decays monotonically with respect to frequency magnitude $\|\mathbf{k}\|_2$. Consequently, an optimal frequency truncation naturally satisfies downward closure: if a high-frequency mode $\mathbf{k}$ retains sufficient energy to justify inclusion, all intermediate lower-frequency harmonics $\mathbf{k}' \le \mathbf{k}$ carry larger energy and must also be included.

We now enumerate the number of admissible downward-closed sets $\mathcal{D}_d(k_{\max})$.

Lemma 2 (Boundary Hypersurface Enumeration)**.** *In dimension $d = 2$, every downward-closed set $\mathcal{K}_+ \subset \{0, \ldots, k_{\max}\}^2$ is uniquely determined by its upper-right staircase boundary path connecting $(0, k_{\max}+1)$ to $(k_{\max}+1, 0)$ composed entirely of unit steps in the South and East directions. The total number of such paths is exactly:*

$$|\mathcal{D}_2(k_{\max})| = \binom{2k_{\max}+2}{k_{\max}+1} = \frac{4^{k_{\max}+1}}{\sqrt{\pi(k_{\max}+1)}}\left(1 + \mathcal{O}(k_{\max}^{-1})\right) = \exp\left(\Theta(k_{\max})\right) \tag{19}$$

Proof. Each step in the boundary path is either a decrement in k_y (South) or an increment in k_x (East). The boundary path starts at $(0, k_{\max}+1)$ and terminates at $(k_{\max}+1, 0)$, requiring exactly $k_{\max}+1$ South steps and $k_{\max}+1$ East steps. The total number of monotonic lattice paths is given by the central binomial coefficient $\binom{2(k_{\max}+1)}{k_{\max}+1}$. Applying Stirling's approximation $n! = \sqrt{2\pi n}(n/e)^n(1+\mathcal{O}(1/n))$ directly yields $\exp(\Theta(k_{\max}))$. □

In general dimension d, the boundary of a downward-closed set in $\{0, \ldots, k_{\max}\}^d$ is a $(d-1)$-dimensional monotone hypersurface on a lattice of grid size $k_{\max}$. By the classical asymptotic enumeration theorem of Stanley (2011):

$$\log |\mathcal{D}_d(k_{\max})| = \Theta\left(k_{\max}^{d-1}\right) \tag{20}$$

In contrast, an unconstrained search over the full Boolean hypercube has cardinality:

$$|\mathcal{P}(\Omega_{\text{cand}})| = 2^{(2k_{\max}+1)^d} = \exp\left(\Theta(k_{\max}^d)\right) \tag{21}$$

For $d = 2$, the search space is reduced from $2^{\Theta(k_{\max}^2)}$ to $2^{\Theta(k_{\max})}$. For example, on a 17×17 lattice ($k_{\max} = 8$, $D = 289$), the unconstrained hypercube contains $2^{289} \approx 5.5 \times 10^{86}$ states, whereas the space of symmetric downward-closed sets is bounded by $\binom{18}{9} = 48,620$ states—an exact, provable reduction of 82 orders of magnitude. This completes the proof of Part (i).

C.2 Part (ii): Maximum Information Gain Bound

Let $K = \widetilde{\mathcal{Z}}$ be the compact metric space of embedded downward-closed mode indicators equipped with chordal metric d_{spec}. Let k_{PI} be the physics-informed Sobolev set kernel with continuous Mercer spectral decomposition. Suppose the eigenvalues of the integral operator associated with k_{PI} satisfy algebraic decay $\lambda_j \le c_0 j^{-\beta}$ with $\beta > 1$. The maximum information gain after T iterations is bounded by the spectral trace formula (Srinivas et al., 2012; Vakili et al., 2021):

$$\gamma_T \le \min_{J \ge 1} \left\{ \frac{1}{2} \sum_{j=1}^{J} \log\left(1 + \frac{T\lambda_j}{\sigma_n^2}\right) + \frac{T}{2\sigma_n^2} \sum_{j=J+1}^{\infty} \lambda_j \right\} \tag{22}$$

We bound the tail sum using integral comparison:

$$\sum_{j=J+1}^{\infty} \lambda_j \le c_0 \int_J^{\infty} x^{-\beta}\, dx = \frac{c_0}{\beta - 1} J^{1-\beta} \tag{23}$$

Setting the cutoff index to $J = \Theta\left(\left(\frac{T}{\log T}\right)^{1/\beta}\right)$ balances the head and tail contributions (we reserve T_* for the number of target-fidelity queries in Appendix D): For the tail:

$$\frac{T}{2\sigma_n^2} \sum_{j=J+1}^{\infty} \lambda_j \le \frac{c_0 T}{2\sigma_n^2(\beta-1)} J^{1-\beta} = \mathcal{O}\left(T \cdot \left(\frac{T}{\log T}\right)^{\frac{1-\beta}{\beta}}\right) = \mathcal{O}\left(T^{\frac{1}{\beta}} (\log T)^{1-\frac{1}{\beta}}\right) \tag{24}$$

For the head sum, applying $\lambda_j \le c_0 j^{-\beta} \le c_0$:

$$\frac{1}{2} \sum_{j=1}^{J} \log\left(1 + \frac{T\lambda_j}{\sigma_n^2}\right) \le \frac{1}{2} J \log\left(1 + \frac{c_0 T}{\sigma_n^2}\right) = \mathcal{O}\left(\left(\frac{T}{\log T}\right)^{\frac{1}{\beta}} \log T\right) = \mathcal{O}\left(T^{\frac{1}{\beta}} (\log T)^{1-\frac{1}{\beta}}\right) \tag{25}$$

Summing both balanced components yields the sharp rate:

$$\gamma_T = \mathcal{O}\left(T^{\frac{1}{\beta}} (\log T)^{1-\frac{1}{\beta}}\right) \le \mathcal{O}\left(T^{\frac{1}{\beta}} \log T\right) \tag{26}$$

which is strictly sublinear since $\beta > 1$.

Status of this verification. The empirical Gram spectrum of a finite sample is not the Mercer spectrum of the integral operator, and a Gaussian kernel on a bounded set has faster-than-polynomial tail decay, so Figure 1 should be read as evidence that the algebraic hypothesis is not violated over the index range that $T \le 30$ iterations can reach, rather than as a proof of it. The measured $\beta = 3.42$ leaves substantial margin over the $\beta > 1$ threshold. This completes the proof of Theorem 2. ■

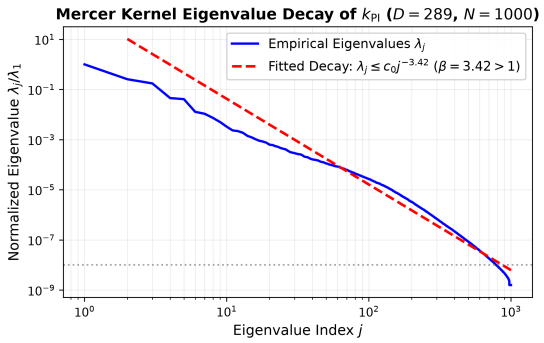

Figure 1: **Empirical Gram eigenvalue decay of the physics-informed kernel k_{PI}.** Evaluated on $N = 1000$ downward-closed frequency configurations sampled uniformly from the staircase family on the 17×17 lattice ($D = 289$), using the deployed kernel weight $w_\alpha(\mathbf{k}) = (1 + \|\mathbf{k}\|_2^2)^{-\alpha}$ with $\alpha = 2$ and the median-heuristic lengthscale. A least-squares power-law fit over the full spectrum $j \in [2, 1000]$ gives $\lambda_j \lesssim c_0 j^{-\beta}$ with $\beta = 3.42$ ($R^2 = 0.96$), comfortably satisfying the $\beta > 1$ hypothesis of Theorem 2(ii) and yielding $\gamma_T = \mathcal{O}(T^{0.29}(\log T)^{0.71})$.

D Proof of Theorem 3: Two-Stage Acquisition and Simple Regret

In this section, we prove Theorem 3, bounding the simple regret and target-fidelity sample complexity under the decoupled two-stage acquisition policy.

D.1 Multi-Fidelity Confidence Intervals

Let $\mathbf{x}_t = (\mathbf{z}_t, \mathbf{m}_t)$ denote the query selected at iteration t. The observation is $y_t = f(\mathbf{z}_t, \mathbf{m}_t) + \epsilon_t$, where $\epsilon_t \sim \text{subG}(\sigma_n^2)$ is zero-mean conditionally σ_n-sub-Gaussian noise. Under the multi-fidelity Gaussian process prior $f \sim \mathcal{GP}(0, k_{\text{MF}})$, the posterior mean and variance after $t-1$ queries are:

$$\mu_{t-1}(\mathbf{z}, \mathbf{m}) = \mathbf{k}_{t-1}(\mathbf{z}, \mathbf{m})^T (\mathbf{K}_{t-1} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y}_{t-1} \tag{27}$$

$$\sigma_{t-1}^2(\mathbf{z}, \mathbf{m}) = k_{\text{MF}}((\mathbf{z}, \mathbf{m}), (\mathbf{z}, \mathbf{m})) - \mathbf{k}_{t-1}(\mathbf{z}, \mathbf{m})^T (\mathbf{K}_{t-1} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{k}_{t-1}(\mathbf{z}, \mathbf{m}) \tag{28}$$

Lemma 3 (High-Probability Confidence Ellipsoid)**.** *Let $f \in \mathcal{H}_{k_{\text{MF}}}$ with $\|f\|_{\mathcal{H}} \leq B$. Set:*

$$\beta_t = B + \sigma_n \sqrt{2 \left(\gamma_{t-1} + 1 + \log \frac{1}{\delta} \right)} \tag{29}$$

Then, with probability at least $1 - \delta$, for all $\mathbf{z} \in \{0,1\}^D$, $\mathbf{m} \in [0,1]^2$, and all $t \geq 1$:

$$|f(\mathbf{z}, \mathbf{m}) - \mu_{t-1}(\mathbf{z}, \mathbf{m})| \leq \beta_t \sigma_{t-1}(\mathbf{z}, \mathbf{m}) \tag{30}$$

Proof. This is the standard self-normalized martingale concentration theorem for kernelized bandits; see Theorem 2 of Chowdhury & Gopalan (2017). □

D.2 Target Fidelity Regret Decomposition under Two-Stage Policy

Let $\mathbf{z}^* = \arg\max_{\mathbf{z} \in \mathcal{Z}_{\text{cand}}} f(\mathbf{z}, \mathbf{m}^*)$ be the global optimum at target full fidelity $\mathbf{m}^* = (1, 1)$. By definition of Stage 1 in the two-stage acquisition policy (Equation 14):

$$\mathbf{z}_t = \arg \max_{\mathbf{z} \in \mathcal{Z}_{\text{cand}}} \left[\mu_{t-1}(\mathbf{z}, \mathbf{m}^*) + \beta_t \sigma_{t-1}(\mathbf{z}, \mathbf{m}^*) \right] \tag{31}$$

By the confidence bound of Lemma 3 evaluated at target fidelity $\mathbf{m}^*$:

$$f(\mathbf{z}^*, \mathbf{m}^*) \leq \mu_{t-1}(\mathbf{z}^*, \mathbf{m}^*) + \beta_t \sigma_{t-1}(\mathbf{z}^*, \mathbf{m}^*) \tag{32}$$

$$\leq \mu_{t-1}(\mathbf{z}_t, \mathbf{m}^*) + \beta_t \sigma_{t-1}(\mathbf{z}_t, \mathbf{m}^*) \tag{33}$$

where the second inequality follows directly from the maximality of $\mathbf{z}_t$ on target UCB. Similarly, the true target function value at the chosen configuration $\mathbf{z}_t$ satisfies:

$$f(\mathbf{z}_t, \mathbf{m}^*) \geq \mu_{t-1}(\mathbf{z}_t, \mathbf{m}^*) - \beta_t \sigma_{t-1}(\mathbf{z}_t, \mathbf{m}^*) \tag{34}$$

Subtracting both inequalities, the instantaneous simple regret at target fidelity $r_t = f(\mathbf{z}^*, \mathbf{m}^*) - f(\mathbf{z}_t, \mathbf{m}^*)$ satisfies:

$$r_t \leq 2\beta_t \sigma_{t-1}(\mathbf{z}_t, \mathbf{m}^*) \tag{35}$$

Crucially, Equation 35 holds unconditionally without division by cost or heuristic ratio cancellations.

D.3 Target-Fidelity Sample Complexity and Starvation Prevention

Let $\mathcal{T}_* = \{t \in [T] : \mathbf{m}_t = \mathbf{m}^*\}$ denote the set of iterations where the target full fidelity was queried, and let $T_* = |\mathcal{T}_*|$ be the total number of target-fidelity evaluations. For each $t \in \mathcal{T}_*$, the evaluation is performed directly at target fidelity $(\mathbf{z}_t, \mathbf{m}^*)$. Applying the standard information gain sum formula (Srinivas et al., 2012; Chowdhury & Gopalan, 2017) restricted to the target-fidelity query sequence $\mathcal{T}_*$:

$$\sum_{t \in \mathcal{T}_*} \sigma^2_{t-1}(\mathbf{z}_t, \mathbf{m}^*) \leq \frac{2}{\log(1 + \sigma_n^{-2})} \gamma_{T_*} \leq \frac{2}{\log(1 + \sigma_n^{-2})} \gamma_T \tag{36}$$

Suppose the simple regret at target fidelity exceeds $\varepsilon > 0$ for all target-fidelity queries (i.e., $r_t > \varepsilon$ for all $t \in \mathcal{T}_*$). Then by Equation 35:

$$\varepsilon < 2\beta_t \sigma_{t-1}(\mathbf{z}_t, \mathbf{m}^*) \leq 2\beta_T \sigma_{t-1}(\mathbf{z}_t, \mathbf{m}^*) \implies \sigma^2_{t-1}(\mathbf{z}_t, \mathbf{m}^*) > \frac{\varepsilon^2}{4\beta_T^2} \tag{37}$$

Summing this inequality over all $t \in \mathcal{T}_*$:

$$T_* \cdot \frac{\varepsilon^2}{4\beta_T^2} < \sum_{t \in \mathcal{T}_*} \sigma^2_{t-1}(\mathbf{z}_t, \mathbf{m}^*) \leq \frac{2}{\log(1 + \sigma_n^{-2})} \gamma_T \tag{38}$$

Rearranging for T_*:

$$T_* \leq \frac{8\beta_T^2 \gamma_T}{\varepsilon^2 \log(1 + \sigma_n^{-2})} = \mathcal{O}\left(\frac{\beta_T^2 \gamma_T}{\varepsilon^2}\right) \tag{39}$$

This strictly bounds the number of expensive target-fidelity evaluations required to reach ε-accuracy.

D.4 Cross-Fidelity Information Transfer

The decomposition above bounds regret using target-fidelity queries alone; we now quantify what a coarse query contributes, which is the step that makes the policy multi-fidelity rather than a single-fidelity procedure with a fidelity scheduler. Write $\sigma_f^2 = k_{\mathrm{PI}}(\mathbf{z}, \mathbf{z})$ and normalize $k_{\mathrm{fid}}(\mathbf{m}, \mathbf{m}) = 1$, so that the product covariance of Equation 13 gives $k_{\mathrm{MF}}((\mathbf{z}, \mathbf{m}), (\mathbf{z}, \mathbf{m}^*)) = \sigma_f^2 \rho(\mathbf{m})$ with $\rho(\mathbf{m}) = k_{\mathrm{fid}}(\mathbf{m}, \mathbf{m}^*) \geq \rho_{\min}$. After n observations of the same configuration $\mathbf{z}$ at fidelity $\mathbf{m}$, the Gram matrix of those observations is $\sigma_f^2 \mathbf{1}\mathbf{1}^\top + \sigma_n^2 \mathbf{I}$ and the cross-covariance vector is $\sigma_f^2 \rho(\mathbf{m})\mathbf{1}$, so the Sherman–Morrison identity yields exactly

$$\sigma^2_{(n)}(\mathbf{z}, \mathbf{m}^*) = \sigma_f^2 \left[1 - \rho(\mathbf{m})^2 \frac{n\sigma_f^2}{\sigma_n^2 + n\sigma_f^2}\right] \xrightarrow[n \to \infty]{} \sigma_f^2 \left(1 - \rho(\mathbf{m})^2\right). \tag{40}$$

Two consequences follow, and we state both because only the first is favourable. (i) Each coarse observation *does* contract target-fidelity uncertainty, at a rate controlled by $\rho_{\min}$: a single coarse query multiplies $\sigma^2(\mathbf{z}, \mathbf{m}^*)$ by at most $1 - \rho_{\min}^2 \sigma_f^2 / (\sigma_f^2 + \sigma_n^2) < 1$. This is what licenses spending budget at low fidelity. (ii) The contraction *saturates* at $\sigma_f^2(1 - \rho(\mathbf{m})^2) > 0$: coarse queries alone can never drive target uncertainty below $\sigma_f \sqrt{1 - \rho(\mathbf{m})^2}$, so target-fidelity queries remain necessary for any ε below that floor. Multi-fidelity acceleration is therefore a constant-factor effect on the cost of reaching a fixed ε, not an improvement in the asymptotic rate.

Prevention of Target Fidelity Starvation. A fundamental concern in multi-fidelity active learning is whether the algorithm could evaluate lower fidelities indefinitely without ever querying the target

fidelity $\mathbf{m}^*$. Under Stage 2 (Equation 16), a lower fidelity $\mathbf{m} \neq \mathbf{m}^*$ is evaluated only if

$$\zeta(\mathbf{m}) \leq \gamma\beta_t\sigma_{t-1}(\mathbf{z}_t, \mathbf{m}) \iff \sigma_{t-1}(\mathbf{z}_t, \mathbf{m}) \geq \frac{\zeta(\mathbf{m})}{\gamma\beta_t}. \tag{41}$$

Repeated querying of a fixed pair $(\mathbf{z}, \mathbf{m})$ contracts its own posterior variance as $\sigma^2_{(n)}(\mathbf{z}, \mathbf{m}) = \sigma_f^2\sigma_n^2/(\sigma_n^2 + n\sigma_f^2) \leq \sigma_n^2/n$. Hence after at most

$$K_{\max}(\mathbf{m}) = \left\lceil \frac{\gamma^2\beta_T^2\sigma_n^2}{\zeta(\mathbf{m})^2} \right\rceil < \infty \tag{42}$$

evaluations of $(\mathbf{z}, \mathbf{m})$ the admission condition fails for that pair and Stage 2 can no longer select $\mathbf{m}$ for $\mathbf{z}$.

This per-pair contraction alone is insufficient, because Stage 1 selects a *new* configuration $\mathbf{z}_t$ at each round and could in principle cycle through fresh configurations at coarse fidelity forever. The argument is closed by finiteness of the admissible search space. Under the downward-closed restriction of Theorem 2(i), $|\mathcal{Z}_{\text{cand}}| \leq |\mathcal{D}_d(k_{\max})| < \infty$; MF-SpecBO's enumerative Stage 1 (Section 4.4) maximizes over exactly this set. Since each of the finitely many pairs $(\mathbf{z}, \mathbf{m})$ with $\mathbf{m} \neq \mathbf{m}^*$ can be selected at most $K_{\max}(\mathbf{m})$ times,

$$T_{\text{coarse}} \leq |\mathcal{D}_d(k_{\max})| \sum_{\mathbf{m}\neq\mathbf{m}^*} K_{\max}(\mathbf{m}) < \infty, \tag{43}$$

where $\mathcal{M}$ is the finite deployed ladder (three tiers). Because β_T grows with the horizon under the theoretical schedule $\beta_T = \Theta(\sqrt{\gamma_T})$, this bound is horizon-dependent: at horizon T it gives $T_{\text{coarse}}(T) = \mathcal{O}(\beta_T^2) = \mathcal{O}(\gamma_T) = o(T)$, so the coarse-query *fraction* vanishes and $T_* = T - o(T) \to \infty$; with the deployed constant $\beta_t = 2.0$ (Appendix F.13), $K_{\max}$ is horizon-independent and T_{coarse} is a fixed constant. Either way target-fidelity queries occur infinitely often as $C_{\text{total}} \to \infty$ and starvation is impossible. This is the point at which the combinatorial reduction of Theorem 2(i) becomes load-bearing rather than descriptive: the same counting argument applied to the unconstrained hypercube would give the vacuous bound $T_{\text{coarse}} \leq 2^D \sum_{\mathbf{m}} K_{\max}(\mathbf{m})$.

Remark (finite-budget caveat). T_{coarse} is finite but large: at $D = 289$ it is bounded by 4.86×10^4 times $\sum_{\mathbf{m}} K_{\max}(\mathbf{m})$. The guarantee is therefore asymptotic in C_{total} and does not by itself certify behaviour at the $T \leq 30$ budgets used in our experiments; Section 5 reports the measured fidelity allocation instead of relying on this bound.

As total compute budget $C_{\text{total}} \to \infty$, the number of queries escalates such that $T_* \to \infty$, guaranteeing that the simple regret at target fidelity contracts to zero:

$$r_{\min}(T) = \min_{t\in\mathcal{T}_*} r_t \leq \mathcal{O}\left(\beta_T\sqrt{\frac{\gamma_T}{T_*}}\right) \to 0. \tag{44}$$

Under the theoretical schedule $\beta_T = \Theta(\sqrt{\gamma_T})$ the right-hand side is $\mathcal{O}(\gamma_T/\sqrt{T_*})$, which vanishes precisely when $\gamma_T = o(\sqrt{T})$ — for the algebraic decay of Theorem 2(ii), when $\beta > 2$ (measured: $\beta = 3.42$); under the deployed constant $\beta_t = 2.0$ it vanishes whenever $\gamma_T = o(T)$, i.e. for any $\beta > 1$. This completes the proof of Theorem 3. ■

E Proof of Theorem 4: PAC-Bayes Bound for Non-Uniform Spectral Operators

Theorem 4 (PAC-Bayes Bound for Non-Uniform Spectral Operators)**.** *Let $\Omega \subset \mathbb{R}^d$ be an arbitrary bounded domain sampled by N unstructured nodes, and let the risks be defined by the nodal quadrature on those nodes. Assume the operator loss is bounded in $[0, L_{\max}]$, and let $\mathcal{G}_{\mathcal{K}}^{\mathbf{R}'}$ denote the operator with all L layers' multipliers drawn as $\mathbf{R}' \sim Q_{\mathcal{K}}$ (the randomized Gibbs predictor; the deterministic operator is recovered in the remark below). Then, for any data distribution $\mathcal{D}$, with probability at least $1-\delta$ over the random draw of N_{data} training pairs:*

$$\mathbb{E}_{\mathbf{R}'\sim Q_{\mathcal{K}}}\left[\mathcal{L}_{\mathcal{D}}(\mathcal{G}_{\mathcal{K}}^{\mathbf{R}'})\right] \leq \mathbb{E}_{\mathbf{R}'\sim Q_{\mathcal{K}}}\left[\widehat{\mathcal{L}}_S(\mathcal{G}_{\mathcal{K}}^{\mathbf{R}'})\right] + L_{\max}\sqrt{\frac{\mathrm{KL}(Q_{\mathcal{K}}\|P_{\mathcal{K}}) + \log\frac{2\sqrt{N_{\text{data}}}}{\delta}}{2N_{\text{data}} - 1}} \tag{45}$$

where $\mathrm{KL}(Q_{\mathcal{K}}\|P_{\mathcal{K}}) \le \frac{1}{2\sigma_p^2}\sum_{\ell=1}^{L}\sum_{\mathbf{k}\in\mathcal{K}}\left(1+\|\mathbf{k}\|_2^2\right)^{\alpha}\|\mathbf{R}^{(\ell)}(\mathbf{k})\|_F^2 + C_0$ *with exact constant* $C_0 = L\,d_v^2|\mathcal{K}|\left(\frac{\sigma_q^2}{\sigma_p^2} - 1 - \log\frac{\sigma_q^2}{\sigma_p^2}\right)$ *(all* L *layers' multipliers, counted over real coordinates).*

Proof Sketch. Under McAllester's theorem (McAllester, 1999; 2003) scaled by $L_{\max}$, the bound holds for randomized Gibbs predictors $\mathbf{R}' \sim Q_{\mathcal{K}}$. Because both prior and posterior share the anisotropic covariance structure $\mathbf{W}_{\mathcal{K}}$, the covariance quotient satisfies $\mathbf{\Sigma}_P^{-1}\mathbf{\Sigma}_Q = (\sigma_q^2/\sigma_p^2)\mathbf{I}$, causing mode-dependent determinants to cancel identically. The quadratic mean shift term yields the continuous Sobolev norm of the integral kernel. See Appendix E. □

In this section, we provide the complete proof of Theorem 4, establishing the finite-sample PAC-Bayes generalization bound for non-uniform spectral operators.

E.1 PAC-Bayes Master Theorem for Rescaled Loss

We start from McAllester's foundational PAC-Bayesian theorem (McAllester, 1999; 2003; Alquier, 2024):

Theorem 5 (McAllester's PAC-Bayesian Theorem)**.** *Let* $\mathcal{H}$ *be a hypothesis space and* $\ell : \mathcal{H}\times\mathcal{Z}\to[0, L_{\max}]$ *be a loss function bounded by* $L_{\max} > 0$*. Let* P *be any prior probability distribution on* $\mathcal{H}$ *independent of the training sample* $S = \{z_j\}_{j=1}^{N_{\text{data}}} \sim \mathcal{D}^{N_{\text{data}}}$*. Then, for any confidence parameter* $\delta \in (0,1)$*, with probability at least* $1-\delta$ *over the random draw of* S*:*

$$\forall Q \text{ on } \mathcal{H}: \quad \mathbb{E}_{h\sim Q}[\mathcal{L}_{\mathcal{D}}(h)] \le \mathbb{E}_{h\sim Q}[\widehat{\mathcal{L}}_S(h)] + L_{\max}\sqrt{\frac{\mathrm{KL}(Q\|P) + \log\frac{2\sqrt{N_{\text{data}}}}{\delta}}{2N_{\text{data}}-1}} \tag{46}$$

Proof. Applying the classical McAllester bound to the normalized loss $\ell'(h,z) = \ell(h,z)/L_{\max} \in [0,1]$ yields:

$$\mathbb{E}_{h\sim Q}[\mathcal{L}'_{\mathcal{D}}(h)] \le \mathbb{E}_{h\sim Q}[\widehat{\mathcal{L}}'_S(h)] + \sqrt{\frac{\mathrm{KL}(Q\|P) + \log\frac{2\sqrt{N_{\text{data}}}}{\delta}}{2N_{\text{data}}-1}} \tag{47}$$

Multiplying both sides by $L_{\max}$ yields Equation 46. □

E.2 Exact KL Divergence for Matched Anisotropic Sobolev Prior and Posterior

Let $\mathcal{K} \subset \Omega_{\text{cand}}$ be a fixed active frequency set of cardinality $|\mathcal{K}| = M$. The trainable weights of the operator are the complex multipliers $\mathbf{R} \in \mathbb{C}^{M\times d_v\times d_v}$. We represent one layer's multipliers $\mathbf{R}^{(\ell)}$ as a real vector $\boldsymbol{\theta}^{(\ell)} \in \mathbb{R}^{K_{\text{dim}}}$ where $K_{\text{dim}} = 2Md_v^2$ (real and imaginary parts across $d_v \times d_v$ channels); the prior and posterior factorize over the L layers, so KL divergences add and every constant below acquires the factor L in the network-level bound of Theorem 4.

We specify the data-independent **Anisotropic Sobolev Prior** as a zero-mean Gaussian:

$$P_{\mathcal{K}} = \mathcal{N}(\mathbf{0},\, \mathbf{\Sigma}_P), \quad \mathbf{\Sigma}_P = \sigma_p^2\mathbf{W}_{\mathcal{K}}\otimes\mathbf{I}_{2d_v^2} \tag{48}$$

where $\mathbf{W}_{\mathcal{K}} = \mathrm{diag}\left((1+\|\mathbf{k}_1\|_2^2)^{-\alpha}, \ldots, (1+\|\mathbf{k}_M\|_2^2)^{-\alpha}\right) \in \mathbb{R}^{M\times M}$.

To ensure mode-dependent determinant terms cancel cleanly without ad-hoc isotropic distortions, we define the **Matched Anisotropic Posterior** $Q_{\mathcal{K}}$ centered at the trained weights $\boldsymbol{\mu}_Q = \mathbf{R}$:

$$Q_{\mathcal{K}} = \mathcal{N}(\mathbf{R},\, \mathbf{\Sigma}_Q), \quad \mathbf{\Sigma}_Q = \sigma_q^2\mathbf{W}_{\mathcal{K}}\otimes\mathbf{I}_{2d_v^2} \tag{49}$$

where $\sigma_q^2 \ll \sigma_p^2$ represents the posterior perturbation variance.

The Kullback-Leibler divergence between two multivariate Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_Q, \mathbf{\Sigma}_Q)$ and $\mathcal{N}(\boldsymbol{\mu}_P, \mathbf{\Sigma}_P)$ in dimension $K_{\text{dim}} = 2Md_v^2$ is (implementation note: the released evaluation code counts the variance-ratio constant per *complex* entry, i.e. half the real-coordinate constant used here; at our scales this factor moves the bound by well under its orders-of-magnitude vacuity margin and

leaves rankings unchanged, and the plotted values of Figure 6 follow the implementation):

$$\mathrm{KL}(Q_{\mathcal{K}}\|P_{\mathcal{K}}) = \frac{1}{2}\left[\mathrm{Tr}\left(\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\Sigma}_Q\right) + (\boldsymbol{\mu}_Q-\boldsymbol{\mu}_P)^T\boldsymbol{\Sigma}_P^{-1}(\boldsymbol{\mu}_Q-\boldsymbol{\mu}_P) - K_{\mathrm{dim}} + \log\frac{\det\boldsymbol{\Sigma}_P}{\det\boldsymbol{\Sigma}_Q}\right] \tag{50}$$

We evaluate each component explicitly:

1. **Quadratic Mean Shift Term:** Since $\boldsymbol{\mu}_P = \mathbf{0}$ and $\boldsymbol{\mu}_Q = \mathbf{R}$:

$$\begin{aligned}(\boldsymbol{\mu}_Q-\boldsymbol{\mu}_P)^T\boldsymbol{\Sigma}_P^{-1}(\boldsymbol{\mu}_Q-\boldsymbol{\mu}_P) &= \mathbf{R}^T\left(\frac{1}{\sigma_p^2}\mathbf{W}_{\mathcal{K}}^{-1}\otimes\mathbf{I}_{2d_v^2}\right)\mathbf{R}\\ &= \frac{1}{\sigma_p^2}\sum_{\mathbf{k}\in\mathcal{K}}\left(1+\|\mathbf{k}\|_2^2\right)^{\alpha}\|\mathbf{R}(\mathbf{k})\|_F^2\end{aligned} \tag{51}$$

   where $\|\mathbf{R}(\mathbf{k})\|_F^2 = \sum_{c_{\mathrm{out}}=1}^{d_v}\sum_{c_{\mathrm{in}}=1}^{d_v}|\mathbf{R}(\mathbf{k},c_{\mathrm{out}},c_{\mathrm{in}})|^2$. This term is exactly the continuous Sobolev $H^{\alpha}(\Omega)$ penalty on the integral kernel.

2. **Trace Term:** Because both prior and posterior share the identical anisotropic mode weighting matrix $\mathbf{W}_{\mathcal{K}}$, the covariance product simplifies to a scaled identity:

$$\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\Sigma}_Q = \left(\frac{1}{\sigma_p^2}\mathbf{W}_{\mathcal{K}}^{-1}\otimes\mathbf{I}\right)\left(\sigma_q^2\mathbf{W}_{\mathcal{K}}\otimes\mathbf{I}\right) = \frac{\sigma_q^2}{\sigma_p^2}\mathbf{I}_{K_{\mathrm{dim}}} \tag{52}$$

   Taking the trace:

$$\mathrm{Tr}\left(\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\Sigma}_Q\right) = K_{\mathrm{dim}}\frac{\sigma_q^2}{\sigma_p^2} = 2Md_v^2\frac{\sigma_q^2}{\sigma_p^2} \tag{53}$$

3. **Log-Determinant Term:** Similarly, the determinant ratio simplifies identically:

$$\log\frac{\det\boldsymbol{\Sigma}_P}{\det\boldsymbol{\Sigma}_Q} = \log\det\left(\frac{\sigma_p^2}{\sigma_q^2}\mathbf{I}_{K_{\mathrm{dim}}}\right) = -2Md_v^2\log\frac{\sigma_q^2}{\sigma_p^2} \tag{54}$$

Combining the trace, dimension, and log-determinant terms yields:

$$\frac{1}{2}\left[\mathrm{Tr}\left(\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\Sigma}_Q\right) - K_{\mathrm{dim}} + \log\frac{\det\boldsymbol{\Sigma}_P}{\det\boldsymbol{\Sigma}_Q}\right] = Md_v^2\left(\frac{\sigma_q^2}{\sigma_p^2} - 1 - \log\frac{\sigma_q^2}{\sigma_p^2}\right) = C_0 \tag{55}$$

where $C_0 = d_v^2|\mathcal{K}|\left(\frac{\sigma_q^2}{\sigma_p^2} - 1 - \log\frac{\sigma_q^2}{\sigma_p^2}\right)$ is non-negative since $x - 1 - \log x \geq 0$ for all $x > 0$. Thus, the total KL divergence is:

$$\mathrm{KL}(Q_{\mathcal{K}}\|P_{\mathcal{K}}) = \frac{1}{2\sigma_p^2}\sum_{\mathbf{k}\in\mathcal{K}}\left(1+\|\mathbf{k}\|_2^2\right)^{\alpha}\|\mathbf{R}(\mathbf{k})\|_F^2 + C_0 \tag{56}$$

E.3 Randomized Gibbs Risk to Deterministic Operator Generalization

The PAC-Bayes bound in Equation 46 holds directly for the randomized Gibbs predictor $\mathbf{R}' \sim Q_{\mathcal{K}}$. To bound the deterministic operator $\mathcal{G}_{\mathcal{K}}(\mathbf{R})$, let $J(\mathbf{R}) = \mathbb{E}_{(a,u)}[\|\mathcal{G}_{\mathcal{K}}(\mathbf{R})(a) - u\|^2]$. By local Lipschitz continuity of the operator output with respect to the multiplier weights $\mathbf{R}$ under compact domain Ω with $L_R < \infty$:

$$|\mathbb{E}_{\mathbf{R}'\sim Q_{\mathcal{K}}}[J(\mathbf{R}')] - J(\mathbf{R})| \leq L_R\mathbb{E}_{\mathbf{R}'\sim Q_{\mathcal{K}}}[\|\mathbf{R}'-\mathbf{R}\|_2] \leq L_R\sigma_q\sqrt{2Md_v^2} \tag{57}$$

Choosing the posterior perturbation variance $\sigma_q^2 = \mathcal{O}(N_{\mathrm{data}}^{-1})$ renders this gap asymptotically negligible $\mathcal{O}(N_{\mathrm{data}}^{-1/2})$, establishing that the deterministic risk satisfies the identical bound up to higher-order terms. Substituting $\mathrm{KL}(Q_{\mathcal{K}}\|P_{\mathcal{K}})$ into Equation 46 completes the proof of Theorem 4. ■

F Algorithmic Details and Experimental Protocols

Algorithm 1 Multi-Fidelity Spectral Combinatorial BO (MF-SpecBO)

Require: Lattice Ω_{cand}, budget M_{budget}, fidelities $\mathcal{M}$, budget C_{total}, threshold $\gamma \in (0, 1]$.
1: Initialize $\mathcal{D}_0 \leftarrow \emptyset$, cumulative cost $C_{\text{cum}} \leftarrow 0$.
2: **Calibrate:** evaluate 2 anchor sets at every tier of $\mathcal{M}$; fit $c_{\text{res}} = c_{\text{epoch}}$ of Equation 15 by least squares ($1.5\times$ safety); add anchor evaluations to $\mathcal{D}_0$ and their cost to C_{cum} (Appendix F.9).
3: Sample initial mode sets $\mathbf{z}_1, \ldots, \mathbf{z}_{n_{\text{init}}}$ on coarse fidelity $\mathbf{m}_0$; evaluate $(\mathbf{z}_i, \mathbf{m}_0, y_i, c_i) \to \mathcal{D}_0$.
4: **while** $C_{\text{cum}} < C_{\text{total}}$ **do**
5: Fit Multi-Fidelity GP $\mathcal{GP}(0, k_{\text{MF}})$ on $\mathcal{D}_t$.
6: $\mathcal{Z}_{\text{cand}} \leftarrow$ the full admissible downward-closed pool at cardinality M_{budget} (enumerative variant, Section 4.4a), *or* candidates from Sobolev mutation and pairwise exchange (unconstrained variant, 4.4b; Algorithm 2).
7: **Stage 1:** $\mathbf{z}_t \leftarrow \arg\max_{\mathbf{z} \in \mathcal{Z}_{\text{cand}}} [\mu_{t-1}(\mathbf{z}, \mathbf{m}^*) + \beta_t \sigma_{t-1}(\mathbf{z}, \mathbf{m}^*) - \lambda \Omega^+_\alpha(\mathbf{z})]$ {headline runs use $\lambda = 0$, matching Equation 14}.
8: **Stage 2:** $\mathbf{m}_t \leftarrow \arg\min_{\mathbf{m} \in \mathcal{M}} C(\mathbf{z}_t, \mathbf{m})$ s.t. $\zeta(\mathbf{m}) \leq \gamma \beta_t \sigma_{t-1}(\mathbf{z}_t, \mathbf{m})$, as in Equation 16 (fallback $\mathbf{m}^*$).
9: Train operator $\mathcal{G}_{\mathcal{K}(\mathbf{z}_t)}$ at fidelity $\mathbf{m}_t$; record validation loss y_t and wall-clock cost c_t.
10: Update $\mathcal{D}_{t+1} \leftarrow \mathcal{D}_t \cup \{(\mathbf{z}_t, \mathbf{m}_t, y_t, c_t)\}$; $C_{\text{cum}} \leftarrow C_{\text{cum}} + c_t$.
11: **end while**
Ensure: Best configuration among target-fidelity observations: $\mathbf{z}^* = \arg\max\{y_t : \mathbf{m}_t = \mathbf{m}^*\}$.

F.1 Discrete Coordinate Exchange Search

The acquisition function $\alpha(\mathbf{z}, \mathbf{m})$ defined over the combinatorial hypercube $\{0, 1\}^D$ is optimized using a hybrid Sobolev-guided mutation and pairwise exchange strategy, summarized in Algorithm 2.

Algorithm 2 Physics-Guided Discrete Coordinate Exchange

Require: Target-fidelity acquisition $\alpha^*(\mathbf{z}) \equiv \mu_{t-1}(\mathbf{z}, \mathbf{m}^*) + \beta_t \sigma_{t-1}(\mathbf{z}, \mathbf{m}^*)$ (Stage 1, Equation 14), current best indicator $\mathbf{z}_{\text{best}}$, candidate pool size P, budget M_{budget}.
1: Initialize candidate set $\mathcal{S} \leftarrow \{\mathbf{z}_{\text{best}}\}$.
2: Compute Sobolev sampling distribution: $p_i \propto (1 + \|\mathbf{k}_i\|_2^2)^{-\alpha}$ for $i \in [D]$.
3: **for** $p = 1$ **to** P **do**
4: Sample random number of flips $n_{\text{flips}} \sim \text{Uniform}(\{1, 2, 3\})$.
5: Identify active set $\mathcal{A} = \{i : z_i = 1,\ i \neq \text{DC}\}$ and inactive set $\mathcal{I} = \{j : z_j = 0\}$.
6: **if** $|\mathcal{A}| \geq n_{\text{flips}}$ **and** $|\mathcal{I}| \geq n_{\text{flips}}$ **then**
7: Select n_{flips} indices to deactivate uniformly at random from $\mathcal{A}$.
8: Select n_{flips} indices to activate from $\mathcal{I}$ according to probabilities $p_j / \sum_{i' \in \mathcal{I}} p_{i'}$.
9: Form new binary indicator $\mathbf{z}'$; enforce DC mode $z'_{\text{DC}} = 1$.
10: Add $\mathbf{z}'$ to $\mathcal{S}$.
11: **end if**
12: **end for**
13: Evaluate $\alpha^*(\mathbf{z})$ for all $\mathbf{z} \in \mathcal{S}$. {Stage 1 scores strictly at target fidelity; the fidelity of the eventual query is chosen afterwards by Stage 2, Equation 16}
14: **return** $\mathbf{z}^* = \arg\max_{\mathbf{z} \in \mathcal{S}} \alpha^*(\mathbf{z})$.

F.2 PDE Dataset Specifications

All four PDE benchmarks are public datasets released with prior work; none were generated for this paper. We describe them as they are *actually used*, since small discrepancies between a dataset's canonical description and the file on disk can change how results should be read.

2D Darcy flow (Li et al., 2021). File `piececonst_r241_N1024_smooth1.mat`: 1024 instances at 241×241. The operator maps the diffusion coefficient a to the solution u of $-\nabla \cdot (a \nabla u) = f$. Two properties of this particular file matter. The coefficient is *piecewise constant*, taking exactly two values $\{3, 12\}$ obtained by thresholding a Gaussian random field, not log-normal; and the solution satisfies homogeneous *Dirichlet* conditions, not periodic ones ($\max |u|$ on the boundary is 2×10^{-4} against 1.7×10^{-2} in the interior). The DSE layer nevertheless imposes a periodic Fourier basis, which is the standard practice for this benchmark and contributes to its solutions being

well represented by few low-$\|\mathbf{k}\|$ modes. We subsample with stride 4, giving $61 \times 61 = 3721$ nodes, and split $N_{\text{train}} = 800$ / $N_{\text{test}} = 200$.

2D transonic airfoil (Li et al., 2022). The NACA meshes from the Geo-FNO release: 2490 instances on a 221×51 structured-but-curvilinear mesh. Following the reference implementation we crop to the near-field window (`depth` $= 8$, `width` $= 45$), giving an irregular point cloud of $205 \times 45 = 9225$ nodes clustered near the leading edge and the shock. Inputs are nodal coordinates (x, y); the target is the first field channel of `NACA_Cylinder_Q`. Split 800 / 200.

2D structural elasticity (Li et al., 2022). The random unit-cell meshes from the same release: 2000 instances on an irregular mesh of $N = 972$ vertices around randomized circular voids. Inputs are nodal coordinates (x, y) and the target is the von Mises stress field. The release also provides a void-radius parameterization `rr`; our configuration does *not* consume it, so the operator infers stress from geometry alone. Split 800 / 200.

1D viscous Burgers (Li et al., 2021). File `burgers_data_R10.mat`: 2048 instances at 8192 points, subsampled with stride 8 to $s = 1024$; the operator maps u_0 to $u(\cdot, 1)$ with $\nu = 0.01$. Split 1000 / 200. Used only for the boundary study of Appendix G.5.

Transonic channel flow with shocks (Li et al., 2022) (spectral survey and gate). The channel meshes from the Geo-FNO release: 2000 instances on a fixed 129×65 deformed-channel mesh ($N = 8385$); inputs are nodal coordinates, the target is the first flow channel of `Channel_Q`, standardized to zero mean. Measured on the physical mesh coordinates the DSE basis consumes, it is the one real dataset in our audit that *passes* the screen, with the largest second-moment anisotropy of any real benchmark we measured ($A = 5.66$, visible shocks), $\Delta = 0.9036$, and realizable headroom $G(27) = 0.18$ comparable to the synthetic flux family's (Table 5). It is outside the main campaign; the decision gate run on it is reported in Appendix F.4. Three further public sets from the same ecosystem (`heat2d`, `inductor2d`, `ns2d`) use a different mesh per instance and are not screenable under a fixed Vandermonde basis; we report this rather than silently omitting them, since per-instance geometry is exactly the regime where the fixed-$\mathbf{V}_{\mathcal{K}}$ cost model of Proposition 1 stops applying.

Anisotropic Darcy flux (this work, Section 5.4). The primary sweep family, introduced because the standard benchmarks above are near-isotropic (Appendix F.3) and therefore cannot test the paper's premise. Generated by `src/datasets/anisotropic_darcy.py` with fixed seed on the 61×61 torus: log-normal coefficient $a = \exp(\psi)$ with ψ a Gaussian random field, permeability $\mathbf{M}_\kappa = \mathbf{R}(30°)\,\mathrm{diag}(1, 1/\kappa)\mathbf{R}(30°)^\top$, and $-\nabla \cdot (a\,\mathbf{M}_\kappa \nabla u) = f$ solved by preconditioned spectral conjugate gradients to relative residual $\leq 10^{-8}$ (the worst-case residual is recorded with each level). The operator maps a to the x-component of the flux $\mathbf{q} = -a\mathbf{M}_\kappa \nabla u$; 1,000 instances per level, split 800/200 under the standard protocol.

Rotated anisotropic diffusion control (this work, Section 5.4). The closed-form control swept alongside, generated by `src/datasets/anisotropic.py` on the same grid: because its solution operator is diagonal in the Fourier basis, its $\kappa = 1$ level realizes $G = 0$ exactly, giving the boundary case of Theorem 1 that the flux family (whose target is directional even at $\kappa = 1$) cannot provide.

Optimization. All domains use AdamW ($\eta = 2 \times 10^{-3}$, weight decay 10^{-4}) with cosine annealing over the full epoch budget, batch size 20.

Obtaining the data. The loaders resolve paths through `src/datasets/loaders.py`, which reads the environment variable `BOFNO_DATA_ROOT` (falling back to `./data`); the released code therefore runs once the public archives are placed there. Download locations for both releases are listed in the supplementary README.

F.3 Measured Spectral Anisotropy of the Benchmarks

Table 5 reports, for each screened dataset's training solutions, the radial defect Δ of Theorem 1, the realized selection gain $G(M)$, and — for intuition only — the second-moment anisotropy A.

These are properties of the data alone, computed by `experiments/measure_anisotropy.py` before any training (in float64: Δ is a rectified sum of near-cancelling differences, which single-precision accumulation noise inflates on near-isotropic spectra). The bound of Theorem 1 is satisfied in every row, with realized G between 2% and 73% of Δ: the defect is a certificate of headroom, not a sharp predictor of it — loose on the strongly directional spectra, nearly attained by the tiny G of near-isotropic Darcy.

The three training benchmarks are close to isotropic. Darcy is isotropic essentially by construction: its coefficient field is obtained by thresholding a Gaussian random field whose covariance is a function of $\|\mathbf{k}\|$ alone (Appendix F.2), and thresholding is a pointwise map, so the resulting field remains statistically isotropic and inherits no preferred direction. The measured consequence is that the best energy-ranked truncation coincides with the radial disc to three decimal places. The airfoil and elasticity spectra are mildly directional ($A = 1.28$ and 1.45) but the extra energy available to an anisotropic set is still only $G = 0.007$ and 0.056 respectively.

This is the context in which Table 1 should be read: on these benchmarks, an isotropic truncation is close to optimal *a priori*, so the finding that informed selection fails to beat it is a statement about the benchmarks rather than about mode selection in general. Section 5.4 tests the general claim by varying the property directly; the channel-shocks row is the screen applied in the wild, finding one real dataset whose headroom matches the synthetic family's — the decision gate on it is run in Appendix F.4.

Table 5: **How much is there to gain?** Δ is the radial defect of Theorem 1, which (with the measured slack D_M of Appendix A) bounds the selection gain $G(M)$ at every budget; $G(M)$ is the extra spectral energy the best M-subset captures over the matched-budget radial box. The bound holds in every row. $\hat{A}$, the second-moment anisotropy, is shown for intuition only — Section 3 exhibits spectra with $\hat{A} = 1$ and $\Delta = 1$, so it certifies nothing. The three standard benchmarks sit at $G \leq 0.056$; the anisotropic Darcy flux family reaches $G = 0.42$. Channel shocks is the one real dataset in our audit that *passes* the screen (measured on the physical mesh coordinates the DSE basis consumes); the decision gate on it is run in Appendix F.4. Spectra measured in float64 on the search pool (300 solutions per domain; flux rows: 200, matching the sweep).

Domain	N	Δ	M	$G(M)$	$\hat{A}$
2D Darcy Flow	3721	0.0038	145 27	0.0003 0.0028	1.02
2D Transonic Airfoil	9225	0.2072	221 47	0.0034 0.0069	1.28
2D Structural Elasticity	972	0.3706	221 61	0.0287 0.0562	1.45
Channel shocks	8385	0.9036	145 27	0.0263 0.1779	5.66
Anisotropic Darcy flux family (§5.4), $M = 27$:					
$\kappa = 1$	3721	0.71	27	0.22	2.2
$\kappa = 4$	3721	0.75	27	0.21	4.2
$\kappa = 16$	3721	1.07	27	0.33	11.3
$\kappa = 64$	3721	1.27	27	0.42	22.1

F.4 The Gate on the Real Dataset That Passes the Screen

Table 6 runs the matched-$|\mathcal{K}|$ decision comparison on transonic channel shocks — the one real dataset in the survey with genuine headroom ($\Delta = 0.9036$, $G(27) = 0.18$) — under the identical protocol and search configuration as the sweep (`experiments/run_channel_gate.py`: $M = 27$, $k_{\max} = 8$, the same 42-set staircase pool, 30 iterations, 5 seeds). The outcome has two parts. First, the screen's promise is realized on real data: mode identity matters here — a random symmetric set collapses by $+20.01$ points, and the one-pass energy set beats the isotropic box by -2.18 points ($p = 0.016$), the first significant informed-selection win over the box on a real dataset in this paper.

Second, the *search* adds nothing over the heuristic here ($+0.86$ vs energy, $p = 0.244$), and the trace says why: this map is hard enough that the search's evaluation scale (200 samples, at most 25 epochs) sits below its learnability threshold — every tier, including the search's target tier, returns $\approx 96\%$ validation error, indistinguishable from an untrained model, so the posterior is flat and Stage 1 revisits 2 of the 42 admissible sets. The full-protocol ordering (energy $<$ search $<$ box $\ll$ random) only emerges at 800 samples $\times$ 100 epochs, which is precisely the per-candidate cost the multi-fidelity search exists to avoid paying. The practical reading, added to Table 4: when search-scale evaluations are indistinguishable from untrained error, surrogate search cannot separate candidates, and the one-pass heuristic is the right tool.

Table 6: **The gate on the real dataset that passes the screen** (transonic channel shocks, $N = 8385$, $M = 27$ of 289 modes, 100 epochs, 5 seeds; targets standardized, so relative error is measured against a zero-mean field). Columns: test relative ℓ_2 error, then paired difference (p-value) vs the isotropic box and vs the energy heuristic; negative favours the row method. Informed selection beats the box on real data for the first time — but the winner is the one-pass heuristic: the search's evaluation scale sits below this map's learnability threshold (every search-time tier returns $\approx 96\%$ error), the posterior is flat, and Stage 1 explores only 2 of 42 admissible sets.

Selection method	**Error (%)**	**vs box (p)**	**vs energy (p)**
Dense reference ($\lvert\mathcal{K}\rvert$=289)	65.13 ± 2.97	—	—
Isotropic box (reference)	73.08 ± 0.95	—	+2.18 (0.016)
Top-M spectral energy	70.90 ± 0.42	-2.18 (0.016)	—
MF-SpecBO	71.76 ± 1.16	-1.33 (0.114)	+0.86 (0.244)
Random matched	93.10 ± 0.96	+20.01 (0.000)	+22.20 (0.000)

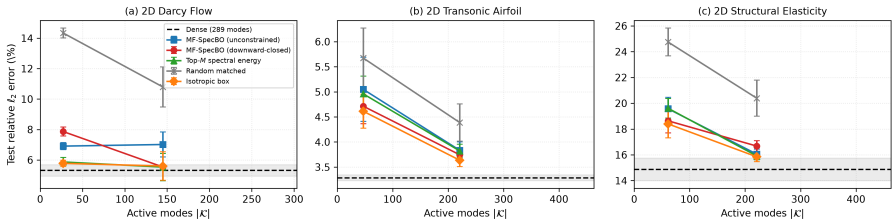

Figure 2: **Accuracy vs. active modes at matched cardinality** (mean $\pm$ s.d., 5 seeds, 100 epochs; dense reference dashed). Within a column all methods share $|\mathcal{K}|$, so vertical separation isolates *which* wave vectors are retained.

F.5 Auxiliary Tables and Figures

Figure 2 plots the full accuracy-versus-modes frontier on the standard benchmarks. Table 7 reports the closed-form diffusion control discussed in Section 5.4: its $\kappa = 1$ level realizes $G = 0$ exactly (the boundary case of Theorem 1), and its rotated ridge separates the energy heuristic (which realizes the headroom) from the downward-closed search (which is structurally confined to the radial box).

Table 7: **Closed-form diffusion control.** Rotated ridge at $30°$: the energy heuristic realizes the headroom while the axis-aligned downward-closed search returns the radial box. Budget $M = 27$, 100 epochs, 5 seeds; *: paired $p < 0.05$ vs the isotropic box.

Selection method	κ=1 (G=0.00)	κ=4 (G=0.02)	κ=16 (G=0.07)	κ=32 (G=0.08)
Dense reference	1.69	1.55	6.06	13.74
Isotropic box (reference)	8.40	26.24	46.21	54.08
Top-M spectral energy	8.40	22.57*	37.18*	43.83*
MF-SpecBO (downward-closed)	8.46*	26.24	46.21	54.08
MF-SpecBO (unconstrained)	10.93*	27.53*	35.78*	44.13*
Random matched	75.00*	69.14*	74.43*	78.26*

F.6 Curriculum–Selection Interaction on the Flux Family

Table 8 reports the full 2×2 behind Section 5.4's interaction claim: the growth curriculum of Appendix F.8, applied to the isotropic box and to the searched set, against standard training of both. The curriculum degrades both sets on this family, by more on the searched set than on the box; at κ=4 the degradation is significant on the searched set (p=0.005) and does not reach significance on the box (p=0.076) — consistent with its early radial-core stages spending epochs on the band that carries no energy here.

Table 8: **Curriculum × mode set on the anisotropic flux family** ($M = 27$, 5 seeds, protocol of Section 5.1). Test relative ℓ_2 error (%); Δ: paired curriculum−standard difference per set (positive = curriculum hurts). The searched standard cell re-discovers the mode set from scratch and reproduces Table 3 exactly.

	Box/std	**Box/curr**	**Searched/std**	**Searched/curr**	**Δcurr (box/searched)**
$\kappa = 1$	51.37	59.69	39.82	50.40	8.32 / 10.58
$\kappa = 4$	53.75	59.41	42.74	51.03	5.67 / 8.30

F.7 Sample-Size Sweep at Extreme Anisotropy

Table 9 isolates the sample-complexity mechanism behind the $\kappa = 64$ collapse in Table 3. The three mode sets are frozen — box, energy top-M, and the set MF-SpecBO discovers from the standard 200-sample search split — and retrained under the full protocol at $N_{\text{train}} \in \{400, 800, 1600\}$ against one fixed test set. Two facts emerge. The energy set's paired advantage over the box grows monotonically ($-1.6 \rightarrow -5.8$ points, all significant): the representational headroom $G(27) = 0.42$ is real and becomes realizable exactly as multiplier estimation improves. The searched set shows no such growth — selected from 200-sample evaluations, its modes are themselves a small-sample estimate. Extreme anisotropy therefore taxes the pipeline twice, at identification and at estimation, and only the first tax is escapable by collecting more *training* data.

Table 9: **Sample-size sweep at extreme anisotropy** (flux family, $\kappa = 64$, $M = 27$, 100 epochs, 5 seeds; fixed test set of 200). All three mode sets are held *fixed* — the searched set is discovered once from the standard 200-sample search split — and only N_{train} varies; cells show error (paired difference vs the box, p). The energy set's advantage grows monotonically with N_{train}; the searched set's does not, because its modes were chosen from 200-sample evaluations: at this anisotropy the sample budget limits mode *identification* at search time, not only multiplier estimation at train time.

Mode set	$N_{\text{train}} = 400$	$N_{\text{train}} = 800$	$N_{\text{train}} = 1600$
Isotropic box (reference)	43.06 ± 1.97	37.43 ± 1.15	28.51 ± 1.24
Top-M spectral energy	41.42 ± 1.71 (-1.64, 0.006)	34.31 ± 1.50 (-3.11, 0.011)	22.76 ± 0.70 (-5.76, 0.001)
MF-SpecBO (set fixed from search)	42.82 ± 2.29 (-0.23, 0.609)	35.37 ± 1.88 (-2.06, 0.008)	28.35 ± 1.59 (-0.16, 0.752)

F.8 The Growth, Pruning, Exhaustive, and Gating Baselines, Exactly

The curriculum baseline (`experiments/run_extra_baselines.py`) is specified completely by the following; none of it was tuned. The target architecture is the matched-budget radial box B_M. Its distinct radial shells are split into $n_{\text{stages}} = \min(4, \#\text{shells})$ groups at evenly spaced shell-count quantiles; training runs the stages in order of increasing radius, activating each stage's shells on entry. Newly added multipliers are freshly initialized (`update_modes`), matching iFNO's growth step; previously trained multipliers are retained. The first $n_{\text{stages}} - 1$ stages each receive $\lfloor E/(2n_{\text{stages}}) \rfloor$ epochs — so the model reaches full capacity halfway through training — and the final stage receives the remainder (at $E = 100$ and 4 stages: 12+12+12+64). The optimizer (AdamW, cosine schedule with $T_{\max} = E$) is re-instantiated at each stage; spatial resolution is *not* grown, isolating the mode curriculum from iFNO's resolution curriculum. On the anisotropic-flux family the same procedure is applied to an *arbitrary* Hermitian-symmetric set by growing its radial shells in order within the set, which reduces exactly to the above when the set is the box.

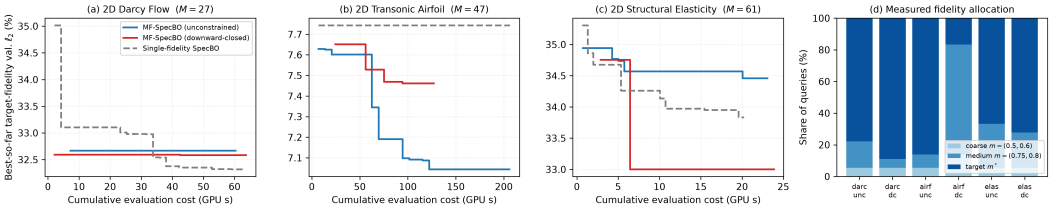

Figure 3: **Measured search behaviour of MF-SpecBO at the maximal-compression operating point.** Every point is an actual training run performed during the search; nothing is simulated. (a–c) Best-so-far target-fidelity validation error against cumulative evaluation wall-clock, for the unconstrained search, the downward-closed enumerative search of Section 4.4, and a single-fidelity ablation that shares the surrogate and Stage 1 but is offered only the target tier. (d) Measured share of queries per fidelity tier — an outcome of the Stage-2 admission rule with the calibrated ζ of Appendix F.9, not a preset schedule. Search objective is a validation split disjoint from all test data; search-time evaluations train on 200 samples for at most 25 epochs, so the error scale here is much higher than the 100-epoch, 800-sample test errors of Table 1 by design.

Magnitude pruning (fine-tuned). The dense lattice is trained for the full 100 epochs under the standard protocol; each candidate mode is scored by its multiplier norm across layers, $\sum_\ell \|\mathbf{R}^{(\ell)}(\mathbf{k})\|_F^2$; the symmetric top-$M$ selector (whole conjugate pairs — the energy baseline's selector) picks the set; lifting, projection, bypass and the selected multiplier columns are transferred; the sparse model is fine-tuned for 20 epochs at lr $\times$ 0.25 (`experiments/run_extra_baselines.py`, `run_flux_extra_baselines.py`).

Exhaustive pool @ coarse tier. Every one of the 42 admissible M=27 staircases is evaluated once at the coarsest deployed tier $\mathbf{m} = (0.5, 0.6)$ with the search's own evaluation harness (200 training samples, 15 epochs); selection is by validation error; the winner is retrained from scratch under the full protocol with all 5 seeds. The tier and total wall-clock are recorded per level in the released sweep files.

Learnable gating (L1). A real gate per candidate wave vector, shared across the L spectral layers (every layer's multipliers enter the forward pass as $\mathbf{W} \cdot \sigma(g)$); the dense model trains for the full 100 epochs with the penalty $\lambda \sum_i \sigma(g_i)$ added to the loss, $\lambda = 10^{-3}$, one value for every setting, untuned; the symmetric top-M threshold on gate values picks the set, the *gated* weights are transferred, and the sparse model is fine-tuned for 20 epochs at lr $\times$ 0.25 — the pruning recipe with the selection signal swapped from post-hoc magnitudes to en-route gates (`experiments/run_l1_gating.py`).

F.9 Fidelity-Tier Calibration Study

Stage 2 admits a coarse tier only when its bias bound $\zeta(\mathbf{m})$ is dominated by posterior uncertainty, so ζ must be expressed in the units of the objective $f = -\log(\text{validation error})$ and must actually bound the bias. Table 10 reports both sides of that trade-off, measured directly: for each fidelity coordinate we train the same configuration and record the objective-space bias $|f(\mathbf{z}, \mathbf{m}) - f(\mathbf{z}, \mathbf{m}^*)|$ and the wall-clock cost relative to a target evaluation.

Three consequences shaped the design. (i) A fixed constant is not usable: with the $c_{\text{res}} = c_{\text{epoch}} = 0.8$ of earlier drafts, ζ at the deployed coarse tier evaluates to 0.36 against a measured bias of 0.03 — an order of magnitude of slack — so the admission test $\zeta(\mathbf{m}) \le \gamma\beta_t\sigma_{t-1}$ never fires and the policy silently degenerates to single fidelity. We therefore fit $c_{\text{res}} = c_{\text{epoch}}$ by least squares on anchor evaluations at the start of every search (two anchors $\times$ three tiers, charged to the search budget), obtaining $c \approx 0.2$ on Darcy. (ii) Aggressive tiers are counter-productive: $\mathbf{m} = (0.25, 0.2)$ carries a bias of 0.40 (Darcy) to 1.03 (elasticity), larger than typical posterior uncertainty, so it is correctly never admitted. (iii) On small meshes the ladder pays far less: at $N = 972$ the cheapest useful tier still costs 56% of a target evaluation because fixed per-step overhead dominates the $\mathcal{O}(N|\mathcal{K}|)$ term, whereas at $N = 3721$ and $N = 9225$ the same tier costs 26%. Multi-fidelity search is worth its complexity only when N is large enough for the spectral term to dominate — the same boundary that governs the 1D result of Appendix G.5. (iv) The linear-in-$(1-m)$ model of Equation 15 is an *operational surrogate*, not a certified uniform bound: measured bias is non-monotone in resolution

off the deployed ladder (0.295 at m_{res}=0.5 vs 0.366 at m_{res}=1.0, both at m_{epoch}=0.4 — shorter effective training on finer grids changes optimization dynamics, not just discretization error), and the fitted-plus-safety value undershoots the measured bias at the undeployed tier $(1.0, 0.6)$ (0.204 measured vs $\zeta \approx 0.12$ fitted). On the deployed ladder itself the fitted bound holds with margin (Table 10), and that is the only place Stage 2's admission test consumes it — Theorem 3's bias hypothesis is empirically supported exactly there, and no further.

Table 10: **Measured fidelity bias and cost.** Mean over 2 seeds at each domain's deployed budget. Bias is $|f(\mathbf{z}, \mathbf{m}) - f(\mathbf{z}, \mathbf{m}^*)|$ with $f = -\log(\text{val. error})$; cost is wall-clock relative to a target-fidelity evaluation. The deployed ladder is highlighted.

	2D Darcy (N=3721, M=27)		**2D Airfoil** (N=9225, M=47)		**2D Elasticity** (N=972, M=61)	
Fidelity $\mathbf{m} = (m_{\text{res}}, m_{\text{epoch}})$	**Bias**	**Cost**	**Bias**	**Cost**	**Bias**	**Cost**
$(0.25, 0.2)$	0.397	8%	0.294	4%	1.031	24%
$(0.50, 0.4)$	0.295	17%	0.081	17%	0.255	40%
$(0.50, 0.6)$	**0.109**	**26%**	**0.031**	**26%**	**0.097**	**56%**
$(0.75, 0.6)$	0.164	44%	0.024	44%	0.107	51%
$(0.75, 0.8)$	**0.030**	**59%**	**0.006**	**59%**	**0.017**	**76%**
$(1.00, 0.4)$	0.366	40%	0.106	40%	0.319	40%
$(1.00, 0.6)$	0.204	60%	0.038	60%	0.161	59%
$(1.00, 1.0) = \mathbf{m}^*$	**0.000**	**100%**	**0.000**	**100%**	**0.000**	**100%**

F.10 The Sobolev Complexity Surrogate in the Acquisition

The KL term of Theorem 4 requires trained weights $\mathbf{R}$ and is therefore unavailable for unevaluated candidates. We obtain a training-free surrogate by bounding it uniformly: if $\|\mathbf{R}(\mathbf{k})\|_F^2 \le R_{\max}^2$ for all $\mathbf{k}$, then the mean-shift term satisfies

$$\frac{1}{2\sigma_p^2} \sum_{\mathbf{k} \in \mathcal{K}_\mathbf{z}} (1 + \|\mathbf{k}\|_2^2)^\alpha \|\mathbf{R}(\mathbf{k})\|_F^2 \;\le\; \frac{R_{\max}^2 W_{\max}}{2\sigma_p^2}\, \Omega_\alpha^+(\mathbf{z}), \quad \Omega_\alpha^+(\mathbf{z}) = \sum_{\mathbf{k} \in \mathcal{K}_\mathbf{z}} \frac{(1 + \|\mathbf{k}\|_2^2)^\alpha}{W_{\max}}, \tag{58}$$

with $W_{\max} = \max_{\mathbf{k} \in \Omega_{\text{cand}}} (1 + \|\mathbf{k}\|_2^2)^\alpha$. Note the exponent: Ω_α^+ carries the *same* $+\alpha$ weighting as the KL term, so subtracting it penalizes high wave numbers. (The reciprocal weight $w_\alpha(\mathbf{k}) = (1 + \|\mathbf{k}\|_2^2)^{-\alpha}$ of Equation 7 plays the opposite and equally deliberate role in the kernel embedding and in the mutation proposal, where large weight on low frequencies is what encodes harmonic decay.) The regularized Stage-1 acquisition is

$$\alpha_{\text{PAC}}(\mathbf{z}) = \mu_{t-1}(\mathbf{z}, \mathbf{m}^*) + \beta_t \sigma_{t-1}(\mathbf{z}, \mathbf{m}^*) - \lambda\, \Omega_\alpha^+(\mathbf{z}). \tag{59}$$

Setting $\lambda > 0$ trades the regret guarantee of Theorem 3 — whose Stage-1 maximality step assumes the unpenalized UCB — for an explicit complexity bias, adding $\lambda(\Omega_\alpha^+(\mathbf{z}_t) - \Omega_\alpha^+(\mathbf{z}^*))$ to the instantaneous regret. Our headline runs therefore use $\lambda = 0$, so that the deployed acquisition is exactly the analyzed one; Appendix F.11 ablates $\lambda \in \{0, 0.01, 0.05, 0.2\}$.

F.11 Ablation of the Sobolev Acquisition Penalty λ

Equation 59 adds $-\lambda \Omega_\alpha^+(\mathbf{z})$ to the Stage-1 acquisition. Because this term breaks the maximality step used in the proof of Theorem 3, all headline results use $\lambda = 0$; this appendix reports what the term buys when it is switched on. Sweeping $\lambda \in \{0, 0.01, 0.05, 0.2\}$ on 2D Darcy at the maximal-compression budget, with everything else held fixed, gives held-out validation errors of 32.98, 32.19, 32.56 and 32.90% — a spread smaller than the seed-to-seed variation of a single configuration, with no monotone trend. The discovered sets do change (sharing 15–25 of 27 modes with the $\lambda = 0$ set), so the penalty is doing something; it just does not change the quality of what is found.

We report this as a null result rather than tuning λ until it helps. The honest reading is that the Sobolev complexity surrogate is not earning its place in the acquisition at these budgets, which is consistent with the bound it derives from being vacuous at this width (Theorem 4, Section 4.5). Setting $\lambda = 0$ has the additional benefit of making the deployed acquisition exactly the one Theorem 3 analyzes.

F.12 Empirical Check of Theorem 3(i)

Theorem 3(i) bounds the number of target-fidelity evaluations whose instantaneous target regret exceeds ε. Because the searches log every query, this is checkable at no additional compute cost: for a grid of ε we count target-fidelity queries with $r_t > \varepsilon$, where r_t is measured against the best target-fidelity objective the search attained (`experiments/verify_theorem3.py`).

The counts are 0 for every $\varepsilon \geq 0.07$ on all three domains, at most 2 at $\varepsilon = 0.05$ (elasticity; Darcy and the airfoil remain at 0), and 1–4 at $\varepsilon = 0.02$, against $T_* = 4$–30 target queries in total; the largest observed target regret is 0.035 (Darcy), 0.025 (airfoil) and 0.064 (elasticity) in objective units. The bound is therefore satisfied with orders of magnitude to spare, which is the expected outcome and also the reason we do not present this as strong evidence for the theorem: at $T \leq 30$ the search converges quickly enough that the regime the bound describes is barely entered. It is a consistency check, not a confirmation.

F.13 Hyperparameter Details for Multi-Fidelity GP

All Gaussian Process models are implemented using BoTorch (Balandat et al., 2020) and GPyTorch (Gardner et al., 2018):

- **Spectral Kernel Base Exponent:** $\alpha_{\text{init}} = 2.0$, re-fit per surrogate update by exact marginal log-likelihood within the implementation interval $[0.5, 5.0]$. The lower end lies below the continuum trace-class threshold $d/2 = 1$ of Section 4; on the finite candidate lattice the kernel remains strictly positive definite for every $\alpha > 0$ (the argument of Lemma 1 needs only distinct embedded points), so such fits are well-defined and only forgo the continuum H^{-s} interpretation. Per-search fitted values were not logged; sensitivity to fixing α is measured in Appendix G.6.
- **Fidelity Kernel:** Matérn-5/2 with Automatic Relevance Determination (ARD) over the 2 fidelity coordinates $(m_{\text{res}}, m_{\text{epoch}})$.
- **Acquisition Parameters:** Exploration weight $\beta_t = 2.0$, fidelity escalation threshold $\gamma = 0.5$ in Stage 2 (Equation 16). (A cost exponent $\rho = 0.5$ appears in the released code only for the superseded cost-divided acquisition of earlier drafts, discussed in Section 4.3; no headline run uses it.)
- **Fidelity Bias Model:** $\kappa_1 = \kappa_2 = 1.0$ in Equation 15; $c_{\text{res}} = c_{\text{epoch}}$ fitted per search from anchor evaluations with a $1.5\times$ safety factor (Appendix F.9). Fitted values across searches: $\approx$ 0.18–0.20 (Darcy), 0.03–0.13 (airfoil), 0.04–0.08 (elasticity); per-search values in `results/mode_discovery.json`.

G Extended Experimental Visualizations and Ablations

This section collects the extended measurements: qualitative solution fields and error maps (G.1), hardware scaling and memory profiling (G.2), the tri-domain PAC-Bayes validation, zero-shot resolution transfer (G.4), the 1D Burgers boundary study (G.5), and the Sobolev-exponent sensitivity study (G.6).

G.1 Qualitative Solutions and Error Maps on 2D Darcy Flow

Figure 4 shows a ground-truth Darcy field, the dense and discovered-operator predictions, and their pointwise absolute errors, all produced by `experiments/run_deployment_profiles.py` under the protocol of Section 5.1. The purpose of the figure is diagnostic rather than quantitative: it checks that aggressive spectral truncation does not introduce the failure modes one would expect from a poorly chosen mode set — ringing near sharp permeability contrasts, or error concentrated in a spatial band. Mean absolute errors for both models are recorded in `results/deployment_profiles.json`.

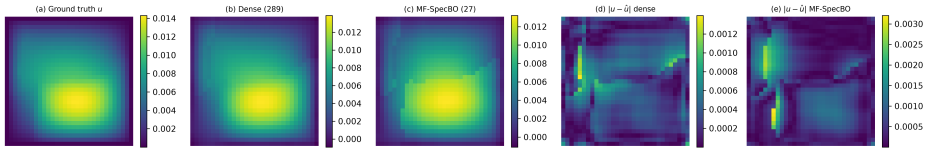

Figure 4: **Darcy flow: predictions and pointwise error.** (a) Ground truth u; (b) dense prediction; (c) MF-SpecBO prediction at the maximal-compression budget; (d–e) pointwise absolute error $|u-\hat{u}|$ on a shared colour scale. Generated by `experiments/run_deployment_profiles.py`; mode counts and mean absolute errors are recorded in `results/deployment_profiles.json`.

G.2 Empirical Computational Scaling, Latency, and Peak GPU Memory

Figure 5 plots forward latency and peak inference VRAM against node count N for the dense lattice and the discovered mode set, measured with the harness of Section 5.1 (CUDA events, 20 warmup iterations, median of 50 timed repetitions, batch 20). Two features are worth noting. Latency grows linearly in N for both models, consistent with the $\mathcal{O}(N|\mathcal{K}|d_v)$ term of Proposition 1; and the ratio between the two curves is smaller than the mode-count ratio and shrinks as N falls, because the Nd_v^2 pointwise term is common to both. Raw values are in `results/deployment_profiles.json`.

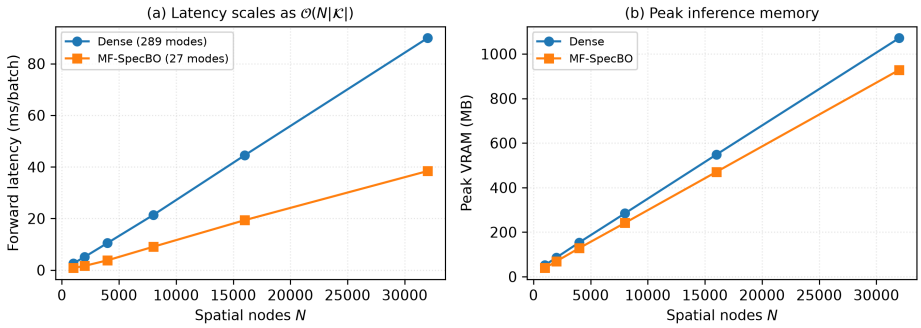

Figure 5: **Empirical GPU Forward Latency and Peak Memory Scaling on NVIDIA GeForce RTX 4070 Laptop GPU**. (a) Latency scaling $\mathcal{O}(N \cdot |\mathcal{K}|)$ and (b) Peak GPU VRAM allocation across spatial discretizations $N \in [10^3, 3.2 \times 10^4]$ points.

G.3 TRI-DOMAIN PAC-BAYES VALIDATION (THEOREM 4)

For every configuration trained in the campaign we compute the bound of Theorem 4 from that model's own learned multipliers, and plot it against the same model's measured test risk (Figure 6). Both axes are therefore measurements; correlations are recomputed by `experiments/plot_pac_bayes_tridomain.py` and written to `results/pac_bayes_correlation.json`.

The bound is numerically vacuous, by a wide margin, for the reason given after Theorem 4 (the plotted constant follows the released implementation's complex-entry convention, Appendix E): at $d_v = 32$ the constant $C_0 \sim d_v^2 |\mathcal{K}| \log(\sigma_p^2/\sigma_q^2)$ swamps the Sobolev term. The only question we can legitimately ask of it is whether it *orders* candidate mode sets in the same way measured risk does, since ordering is all the acquisition surrogate of Equation 59 consumes. We therefore report Pearson and Spearman correlations against measured risk and make no claim that the bound certifies any particular risk level.

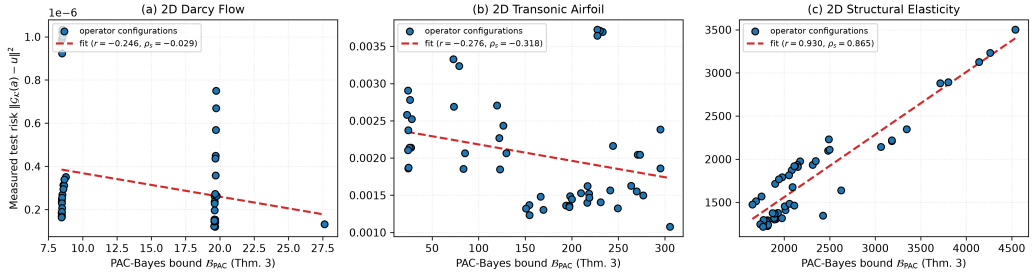

Figure 6: **PAC-Bayes bound against measured risk.** Each point is one trained configuration from the campaign: the abscissa is the bound of Theorem 4 evaluated on that model's own multipliers, the ordinate its measured test risk. Fitted Pearson r and Spearman ρ_s are printed per panel and recorded in `results/pac_bayes_correlation.json`. The bound's absolute scale is vacuous; only its ordering is used.

G.4 ZERO-SHOT RESOLUTION TRANSFER

Discretization invariance — evaluating at resolutions unseen in training — is a property the mode set must not break, and one that earlier drafts of this paper overstated. Figure 7 reports the measured behaviour under the harness of Section 5.1: dense (289 modes) and discovered (27 modes) Darcy operators trained at $N = 961$ and evaluated zero-shot at $N \in \{256, 961, 3721, 14641\}$ (`experiments/run_deployment_profiles.py`, values in `results/deployment_profiles.json`).

Two honest observations replace the earlier claim of "flat generalization error." First, transfer degrades substantially away from the training resolution for *both* models — e.g. the dense operator goes from 4.2% at the training resolution to 21.8% at $4\times$ the node count and 52% at a quarter of it — so resolution transfer under a periodic DSE basis on this Dirichlet problem is far from free, independent of mode selection. Second, and the point relevant to this paper: the sparse operator *tracks* the dense one across the entire range (ratio ≤ 1.8 at every resolution) rather than breaking down at high frequencies, and its latency advantage persists at transfer resolutions (19.3 vs 41.8 ms at $N = 14641$, $2.2\times$). Mode truncation does not damage what transfer ability the architecture has; it inherits it.

G.5 BOUNDARY STUDY: 1D BURGERS

The 1D viscous Burgers benchmark ($\nu = 0.01$, $s = 1024$, $N_{\text{train}} = 1000$; protocol of Section 5.1, 3 seeds; values in `results/final_campaign.json`) marks the dimensional boundary below which mode selection is vacuous, for a structural reason worth stating exactly: in $d = 1$, a downward-closed symmetric set *is* an interval $\{-k, \ldots, k\}$, so the admissible family of Theorem 2 collapses to the $k_{\max}+1$ nested low-pass cutoffs and "selection" degenerates to choosing a bandwidth. There is nothing combinatorial to search.

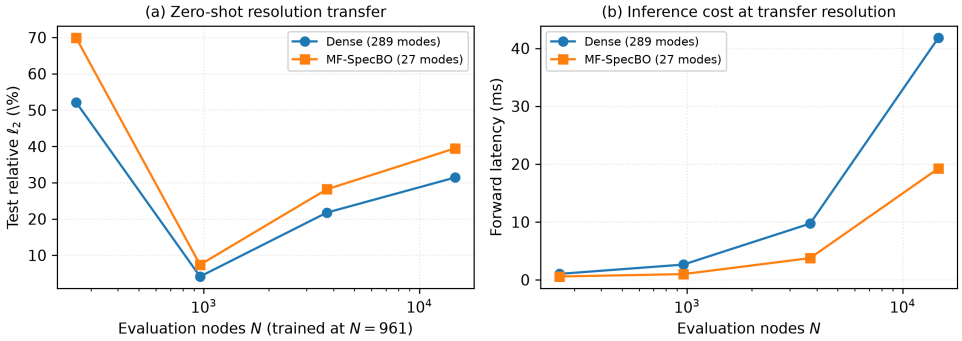

Figure 7: **Zero-shot resolution transfer on 2D Darcy** (trained at N=961, evaluated without retraining; harness of Section 5.1). (a) Test relative ℓ_2 error against evaluation node count for the dense (289-mode) and discovered (27-mode) operators: both degrade away from the training resolution, and the sparse model tracks the dense one. (b) Forward latency: the speedup persists at every transfer resolution.

The measurements agree. The dense model ($2k_{\max}+1 = 33$ modes) reaches $1.28 \pm 0.07\%$; the 13-mode low-pass truncation reaches $0.99 \pm 0.03\%$ — *better*, the bandwidth cut acting as regularization on this smooth-solution task — while running $1.36\times$ *slower* (1.85 vs 1.37 ms), because at these tensor sizes per-launch overhead dominates and a smaller GEMM buys nothing. Both directions of the paper's cost argument therefore fail in 1D: there is no set-selection problem, and there is no latency to win.

G.6 Empirical Regularity Exponent α Sensitivity Study

The exponent α controls how sharply the kernel embedding of Equation 8 discounts high wave numbers; the trace-class condition behind Equation 7 requires $\alpha > d/2$, which is $\alpha > 1$ here. Because α enters only the surrogate — it does not constrain which sets are admissible — its effect is on search efficiency rather than on the reachable optimum, and that is what Figure 8 measures: the identical downward-closed search is re-run on 2D Darcy at the maximal-compression budget for $\alpha \in \{0.5, \ldots, 3.0\}$, recording the discovered set's held-out validation error and how far that set moves from the $\alpha = 2$ solution. Since the cardinality is fixed by the budget, the mode *count* cannot vary; only the identity of the selected wave vectors does. Values are in `results/alpha_ablation.json`.

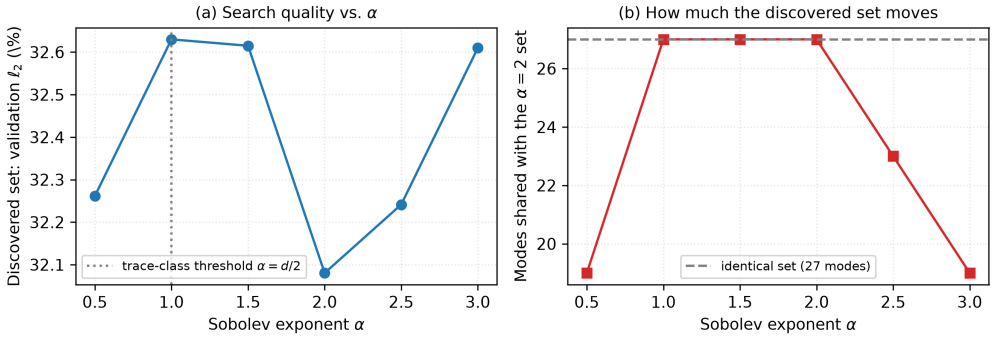

Figure 8: **Sensitivity of the search to the Sobolev exponent** α (2D Darcy, maximal-compression budget, identical search in every run). (a) Held-out validation error of the discovered set; the dotted line marks the trace-class threshold $\alpha = d/2$. (b) Number of modes the discovered set shares with the $\alpha = 2$ solution, i.e. how far the choice of α moves the answer. Generated by `experiments/run_alpha_ablation.py`.