TEST-TIME COMPUTE ALLOCATION IN RANDOM-FEATURE PDE SOLVERS: WIDTH VERSUS SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Under a fixed work budget, should a random-feature partial differential equation (PDE) solver widen one trial space or search across several? We compare these allocations using the same feature distribution and fitting procedure. An exact error decomposition separates candidate-set quality from final selection loss using a post-solve reference-error oracle. Across 60 independently sampled linear advection–diffusion–reaction problems that activate search, six budgets, and four seeds, the task-geometric search/wide error ratio is 9.32 (95% CI 7.92–10.91); search wins on none of the 60 tasks under this paired statistic. Oracle selection from the same candidate sets still yields a ratio of 9.26, while final selection adds only 0.7% to the geometric mean error. Thus, almost the entire gap is already present in the available candidates. Ablations nevertheless show that physics-based selection is useful within those populations. Finite-set ranking guarantees and a spectral residual bound for a restricted elliptic operator class clarify when verification controls ranking and solution error. These findings identify candidate generation and compute allocation as the main bottlenecks for the tested search policy: accurate final verification cannot compensate for the absence of a competitive candidate.

1 INTRODUCTION

Random-feature methods solve a PDE by sampling basis functions and fitting their linear coefficients (Rahimi & Recht, 2007; Chen et al., 2022; Liao, 2026). The basis determines which solutions can be represented, while the coefficient solve finds a fit within that space. Two banks with the same number of features can therefore produce different errors. This variability suggests two ways to spend additional computation: widen one bank, or solve several banks and select a promising one.

The second strategy needs a way to compare candidates without knowing the solution. PDE residuals provide such a signal. A search policy can fit candidates on one set of points, rank them using independent physics evaluations, and spend the remaining budget on the survivors. This is a form of test-time scaling: all computation takes place after a new PDE instance is supplied, without a pretrained solution model. As in other generation–verification systems, the balance between producing candidates and evaluating them matters (Snell et al., 2025).

There is also a cost that more accurate verification cannot remove. Fitting and checking several banks leaves less computation for each bank. A selector may identify the best candidate correctly and still return a worse solution than a single wider solve. Distinguishing these two sources of error is essential: a ranking failure calls for a better verifier, whereas a candidate quality failure calls for a different allocation or representation.

We examine this distinction in random-feature solvers for two-dimensional advection–diffusion–reaction equations. HYBRIDWIDE fits the largest affordable bank from a mixture of tanh, exponential, Gaussian, and sine features. GATE-RFM+ uses the same feature distribution and linear solver, but splits the budget across banks, ranks them by held-out physics, and widens selected survivors. An input-based gate restricts search to a specified regime; outside it, the policy returns the same wide solution. This design lets us separate the value of the representation from the value of searching over it.

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.

The primary experiment favors the single wide model on every task under the paired task statistic. More revealingly, even the best candidate in each realized search population, selected with access to reference error, remains far behind the wide baseline on aggregate. The result is not explained by final ranking noise. At the same time, random selection is much worse than physics selection within those populations. Both observations can hold: verification helps choose among available candidates, but the allocation may fail to produce a competitive candidate in the first place.

The paper makes three contributions. First, it gives a controlled comparison of width and population search with shared features, fitting, and compute accounting. Second, it uses an exact population-gain/selection-loss decomposition to locate the performance gap, supported by full comparator and component-ablation results. Third, it connects independent physics verification to finite-set ranking guarantees and, for a specified elliptic operator, to solution error. Together these provide a way to assess whether a search policy needs better candidates, better selection, or both.

2 Related work

Random-feature PDE solvers. The random-feature method (RFM) combines randomized bases, collocation, and linear least squares; multiscale and local constructions extend its approximation capacity (Chen et al., 2022; Liao, 2026). Weak formulations change the fitting objective (Kuvakin et al., 2025), local filtering improves the feature system (van Beek et al., 2025), and structure-adaptive methods exploit coordinate information (Linghu et al., 2026). These approaches emphasize the representation or its numerical solution. We study a complementary question: once a representation is fixed, is computation better spent on one large space or a search over several smaller spaces?

Inference-time computation for physics. SCaSML corrects a pretrained surrogate through a defect equation (Fan et al., 2026); RecurrSim varies recurrent depth to trade accuracy against cost (Abdul Majid et al., 2026). CodePDE uses language-model inference to generate solver code (Li et al., 2026), while neural operator splitting searches compositions of pretrained operators (Serrano et al., 2026). Our computational units are freshly fitted trial spaces. The comparison concerns how to divide a fixed solve budget across them, including the cost of verification.

Residuals, verification, and allocation. PINNs use residuals as training losses (Raissi et al., 2019), although optimization imbalance can weaken the link between training progress and solution accuracy (Wang et al., 2022). Residual-adaptive sampling changes fitting points (Wu et al., 2023); certification instead bounds a fixed model's residual (Eiras et al., 2024). Here held-out residuals rank separately fitted models. Successive halving and Hyperband supply the allocation templates (Jamieson & Talwalkar, 2016; Li et al., 2018). Our frequency-weighted score is motivated by dual-residual and optimal-test ideas (Demkowicz & Gopalakrishnan, 2011); its finite dictionary makes it an efficient ranking statistic. Section 3.4 states the additional conditions needed to connect that statistic to solution error.

3 Comparing breadth and width

3.1 A shared trial space and numerical solver

We consider the boundary-value problem

$$\mathcal{L}u = -\nabla\cdot(\epsilon(x)\nabla u) + \beta\cdot\nabla u + cu = f \quad \text{in } \Omega = (0,1)^2, \qquad u|_{\partial\Omega} = 0. \tag{1}$$

Here $\epsilon(x) > 0$ is diffusion, $\beta \in \mathbb{R}^2$ is constant advection, $c \geq 0$ is reaction, and f is the supplied forcing. A bank θ specifies M frozen features and a fitted solution

$$\mathcal{V}_\theta = \operatorname{span}\{\phi_{\theta j}\}_{j=1}^{M}, \qquad u_\theta(x) = \sum_{j=1}^{M} \alpha_{\theta j}\phi_{\theta j}(x). \tag{2}$$

Here $\alpha_{\theta j}$ is the fitted coefficient of feature j. All banks cover the same domain. Their feature draws change the function space, rather than partitioning the domain into subregions.

The shared hybrid bank contains tanh ridges (3/8), outflow-oriented exponentials (1/4), local Gaussians (1/4), and sine products (1/8). Nonsine features are multiplied by $q(x) = x_1(1-x_1)x_2(1-x_2)$; sine products already vanish on the boundary. Exponential rates depend on the supplied advection and diffusion, and feature anchors use forcing summaries. No feature parameter is obtained from the reference solution. Banks from different seeds form separate lineages; widening a lineage takes a longer prefix of its fixed feature draw. The exact feature laws are in Appendix A.

For fitting points $X_f = \{x_i\}_{i=1}^{N_f}$ drawn from a known density $p(x)$, we form $A_{ij} = (\mathcal{L}\phi_{\theta j})(x_i)$, $b_i = f(x_i)$, and $W_{ii} = p(x_i)^{-1/2}$. The fitting objective is

$$\min_{\alpha\in\mathbb{R}^M} \|W(A\alpha - b)\|_2^2 + \lambda\|D\alpha\|_2^2, \qquad D_{jj} = \max\left\{\sqrt{N_f^{-1}\sum_i (WA)_{ij}^2}, 10^{-12}\right\}. \tag{3}$$

The weights preserve the uniform-domain residual objective in expectation. Column scaling and a truncated singular value decomposition (SVD) solve this objective in the retained singular subspace, with regularization strength $\lambda = 10^{-10}$. The same sampling density, scaling, truncation rule, and seed construction are used for the wide and search policies. Fitting points are shared across banks of equal width for a given task and seed.

3.2 Allocating one budget

HybridWide chooses the largest affordable width and fits one bank. Gate-RFM+ first evaluates a gate based on the PDE coefficients and forcing concentration, location, boundary mass, and roughness. The gate decides between one wide solve and the search procedure in Table 1. Appendix A gives the observable statistics and fixed thresholds. No gate decision uses a fitted solution's reference error.

Table 1: Budgeted population search. The candidate list Θ stores every completed fit; the active set A contains only banks eligible for promotion. $C(M)$ is the full cost in Eq. (4); S_A collects active scores, $S_{\min}$ is their minimum, and MAD is median absolute deviation.

Input: PDE, work budget B, random seed, ordered width ladder.
1. Gate. If closed, return HybridWide with the same budget and seed.
2. Initialize. Set spent work to zero and $A = \Theta = \varnothing$. At the smallest affordable width, fit up to four independently drawn banks. Charge each fit, evaluate its selection score, and add it to A and Θ.
3. Prune. Sort A by selection score. Keep up to its best $\max(1, \lceil |A|/2 \rceil)$ banks whose scores are at most $S_{\min} + 1.4826\ \mathrm{MAD}(S_A)$.
4. Widen. In score order, refit survivors at the next ladder width if the remaining budget covers $C(M)$. Add each fit to Θ and replace A by these new fits. Repeat steps 3–4 until no promotion is affordable.
5. Accept. Draw a fresh acceptance sample. Return the member of Θ with the smallest acceptance score.

Every fit is charged against a common deterministic work budget B. For width M, fitting size N_f, selection size N_s, acceptance size N_a, and verifier dictionary size K_v, the standard action costs

$$C(M) = 3(N_f + N_s + N_a)M + (N_s + N_a)K_v + \frac{4N_fM^2 + 8M^3}{64}. \tag{4}$$

The implementation takes the integer part. The terms account for feature evaluations, moment projection, and dense linear algebra. Refitting a survivor incurs a full action cost; the total must not exceed B. Single-bank methods use $N_s = N_a = 0$. Extra QR, weak-fitting, and residual-refinement operations are charged separately. This work proxy provides reproducible budget matching; measured online time also includes proposal construction, gate evaluation, and library overhead.

The budget makes breadth and width compete even before selection error enters. With $N_f = 8M$, the dominant fit term is $(5/8)M^3$. In an idealized one-fit-per-bank allocation, splitting that part of a budget equally across K banks reduces affordable width by approximately $K^{-1/3}$. Verification and repeated fits impose further costs. Thus sampling more banks does not mean sampling more solutions from an unchanged error distribution.

Write $\mathcal{E}(v) = \|v - u\|_{L^2}/\|u\|_{L^2}$ for relative error, with $\|u\|_{L^2} > 0$. Fix the fitting points and let $F_M(e)$ be the conditional error distribution of an independently drawn width-M bank. At fixed width, the oracle among K independent banks obeys

$$\Pr\left\{\min_{k\le K} \mathcal{E}(\theta_k) > e\right\} = \big(1 - F_M(e)\big)^K. \tag{5}$$

Increasing K improves this fixed-width oracle. Under a fixed budget, however, M also changes, so the distribution on the right changes with it. Adaptive descendants additionally share their parent features. We therefore use the realized candidate populations to diagnose the implemented search, rather than treating Eq. (5) as its scaling law.

3.3 Separating candidate quality from selection loss

For a fixed task, budget, and seed, let b be the wide baseline, Θ the candidates generated by search, o the member of Θ with smallest reference error, and $\widehat{v}$ the returned candidate. Reference error is evaluated only after the online solve; $\log$ denotes the natural logarithm.

Proposition 1 (Population gain and selection loss)**.** *For positive candidate and baseline errors, define* $G_{\mathrm{pop}} = \log[\mathcal{E}(b)/\mathcal{E}(o)]$ *and* $T_{\mathrm{sel}} = \log[\mathcal{E}(\widehat{v})/\mathcal{E}(o)]$*. Then*

$$\log \frac{\mathcal{E}(\widehat{v})}{\mathcal{E}(b)} = T_{\mathrm{sel}} - G_{\mathrm{pop}}, \qquad T_{\mathrm{sel}} \geq 0. \tag{6}$$

Search beats the baseline if and only if $G_{\mathrm{pop}} > T_{\mathrm{sel}}$.

The identity follows by canceling $\mathcal{E}(o)$. It separates three cases. If $G_{\mathrm{pop}} \leq 0$, the population contains no solution better than the baseline. If $G_{\mathrm{pop}} > 0$ but $T_{\mathrm{sel}} \geq G_{\mathrm{pop}}$, selection loses the available advantage. If $G_{\mathrm{pop}} > T_{\mathrm{sel}}$, search converts candidate diversity into a net improvement. The first two cases suggest different interventions: change candidate generation or resource allocation in the first, and improve selection in the second.

For aggregate comparisons, we average each term using the same paired task weights. Exponentiating gives a product of the oracle/baseline error ratio and the selected/oracle penalty. This remains exact for the geometric aggregate and avoids interpreting a ratio of separately computed medians as a paired effect. The oracle diagnoses the completed population: changing intermediate ranking decisions could generate a different population.

3.4 Physics scores and independent verification

Search ranks banks on a common uniform sample X_s, initially independent of fitting. Once the candidate set is fixed, a new i.i.d. uniform sample X_a makes the final choice. Neither sample contains solution labels. For $r_\theta = \mathcal{L}u_\theta - f$, sine tests $\psi_k(x) = \sin(\pi k_1 x_1)\sin(\pi k_2 x_2)$ for positive integer pairs $k = (k_1, k_2)$, and frequencies $\omega_k = \pi(k_1, k_2)$, define

$$\widehat{m}_{\theta k}(X) = \frac{4}{|X|\sigma_k} \sum_{x \in X} r_\theta(x)\psi_k(x), \qquad S_X(\theta) = \frac{\|\widehat{m}_\theta(X)\|_2}{\sqrt{|X|^{-1}\sum_{x\in X}|f(x)|^2} + 10^{-30}}. \tag{7}$$

The weight is $\sigma_k = \sqrt{(\epsilon_{\mathrm{ref}}|\omega_k|^2 + c)^2 + (\beta^\top \omega_k)^2 + 1}$, using the supplied nominal diffusion ϵ_{ref} for variable-coefficient tasks. Low-frequency tests are ordered by squared frequency. Dividing by the forcing scale makes scores comparable within a task and does not change their ranking. The score estimates a finite vector of weak residual moments, weighted by an operator-symbol proxy.

Independent acceptance prevents reuse of the search sample from invalidating the following finite-set statement. For a uniform random point X, write $Z_{\theta k}(X) = 4r_\theta(X)\psi_k(X)/\sigma_k$, $m_{\theta k} = \mathbb{E}Z_{\theta k}(X)$, and $V(\theta) = \|m_\theta\|_2$.

Theorem 1 (Ranking after adaptive search)**.** *Conditional on the completed search, let* Θ *contain* J *candidates. Suppose* X_a *contains* $n_a = N_a$ *i.i.d. uniform points independent of that search, and* $|Z_{\theta k}(X)| \leq R$ *almost surely for all candidates and tests. For* $0 < \eta < 1$*, with probability at least* $1 - \eta$*, simultaneously*

$$|\widehat{V}_a(\theta) - V(\theta)| \leq \delta, \qquad V(\widehat{\theta}_a) \leq \min_{\theta\in\Theta} V(\theta) + 2\delta,$$
$$\delta = R\sqrt{\frac{2K_v \log(2JK_v/\eta)}{n_a}}, \tag{8}$$

where $\widehat{\theta}_a$ *minimizes the acceptance score and* $\widehat{V}_a = \|\widehat{m}(X_a)\|_2$.

The proof uses conditional Hoeffding bounds and a union bound (Appendix B). It controls selection in the population weak score. Relating that score to solution error requires information about the operator and residual components outside the test dictionary.

For a unique population-score minimizer, let $\Delta > 0$ be its score gap to the next candidate. On the event in Theorem 1, final selection recovers that minimizer whenever $2\delta < \Delta$. A sufficient acceptance size is therefore

$$n_a > \frac{8R^2 K_v \log(2JK_v/\eta)}{\Delta^2}. \tag{9}$$

For fixed R and Δ, the dependence on the number of candidates is logarithmic, while resolving a smaller score gap requires quadratically more samples. The test count also enters the bound and the work charge. This gives a reason to allocate verification effort according to ranking difficulty. The experiment uses a fixed acceptance size; Eq. (9) is a sufficient condition, not an estimate of its actual sample requirement.

Theorem 2 (Spectral residuals and solution error)**.** *For $\mathcal{L} = -\nu\Delta + c$ on the unit square with homogeneous Dirichlet data, constants $\nu > 0$, $c \geq 0$, and orthonormal sine eigenpairs (e_k, λ_k) of $\mathcal{L}$, indexed by positive integer pairs k, let $r_v = \mathcal{L}v - f \in L^2$ and $v - u \in H_0^1$. For a finite index set I, define $V_I(v)^2 = \sum_{k\in I} |\langle r_v, e_k\rangle|^2/\lambda_k^2$ and $T_I(v) = \|r_v\|_{L^2}/\inf_{k\notin I}\lambda_k$. Then*

$$V_I(v) \leq \|v - u\|_{L^2} \leq \sqrt{V_I(v)^2 + T_I(v)^2}. \tag{10}$$

If all empirical V_I scores on a fixed candidate set are within δ of their population values, their minimizer $\widehat{v}$ satisfies

$$\|\widehat{v} - u\|_{L^2} \leq \min_{v\in\Theta} \|v - u\|_{L^2} + 2\delta + T_I(\widehat{v}). \tag{11}$$

This bound exposes the two remaining terms: sampling error and unresolved residual content. A larger acceptance sample addresses the first; it does not address the second. The advection score in Eq. (7) uses a symbol approximation, so its empirical performance must be assessed separately from this constant-coefficient spectral result.

The unresolved term is necessary, not just a feature of the proof. Choose a sine mode e_ℓ outside I and a candidate $v = u + ae_\ell$. Its residual is $a\lambda_\ell e_\ell$, so every tested population moment vanishes although $\|v - u\|_{L^2} = |a|$. Any finite dictionary admits such a direction. Accurate estimation of the tested moments alone cannot give a uniform error guarantee over all candidates. This is why the empirical oracle decomposition and the sampling theorem answer different questions.

A complementary stability statement applies to the advection operator. For $v - u \in H_0^1$, diffusion $\epsilon \geq \epsilon_0 > 0$, and $c - \frac{1}{2}\nabla\cdot\beta \geq 0$, an energy estimate gives

$$\|v - u\|_{L^2} \leq \frac{C_P}{\epsilon_0}\|\mathcal{L}v - f\|_{H^{-1}}, \tag{12}$$

where C_P is the domain's Poincaré constant and the dual norm uses the H_0^1 gradient norm. Appendix B gives the derivation. This estimate concerns the full dual residual. The finite score in Eq. (7) can approximate useful directions in that residual, but requires additional control of the untested directions. Its guarantee also becomes less informative as diffusion decreases, consistent with the need to resolve narrow layers in the trial space.

4 Controlled evaluation of the allocation tradeoff

Tasks and budgets. The primary sample contains 60 independently drawn PDEs conditional on gate activation: 31 with variable diffusion and 29 with superposed layers. Parameters were drawn from a mixture of variable-diffusion, multiscale, and superposed-layer generators, followed by rejection using the observable gate. All 38 rejected draws were multiscale tasks. The primary comparison therefore describes the accepted two-family population. Additional tests contain 32 unconditioned tasks, 32 transformations of eight base PDEs, and 12 problems with independently specified forcing, for 136 task instances. Appendix A specifies their construction.

These families vary where the approximation difficulty lies. Variable diffusion changes the operator across the domain; superposed layers combine different decay scales; multiscale tasks mix localized

layers with oscillatory content. Smooth tasks in the broad suite remove advection and use sine mixtures. The independently forced problems avoid prescribing an exact solution, while the symmetry tests change coordinates without changing the underlying boundary-value problem. These roles motivate the separate suites rather than pooling them into a single performance statistic.

Each primary task uses six work budgets, $B \in \{1.5, 3, 6, 12, 24, 48\} \times 10^6$, and four paired random seeds. The feature ladder runs from 32 to 512, with $N_f = \max(256, 8M)$. Search uses $N_s = 256$, $N_a = 512$, $K_v = 32$, and four initial banks. The same settings apply to the main comparison at every budget. Verification costs are charged whether or not they change the chosen solution.

Comparators and ablations. The default matrix contains 14 algorithms. Besides HYBRIDWIDE and GATE-RFM+, it includes equal-budget best-of-2/4/8, successive halving, a two-bracket Hyperband variant, residual row adaptation, QR feature pruning, and weak-moment fitting. Broad and forcing-conditioned tanh banks, local Gaussian features, and sine features test the role of representation. These are controlled implementations of the stated strategies; their definitions and cost adjustments appear in Appendix A.

Component tests use 24 primary tasks, three budgets, and three seeds. They vary the feature mixture, fitting rows, conditioning, initial breadth, pruning, scoring rule, acceptance independence, and gate. A separate 12-task grid varies acceptance size and dictionary size. Gate variants are also evaluated on broad and transformed tasks. The resulting linear-PDE study contains 51,584 evaluations, including finite-difference reference checks; Appendix A gives the count by experiment.

Errors, timing, and statistical units. Online timing stops before reference solutions, oracle errors, or fine-grid evaluations are computed. Relative L^2 errors use a 128^2 midpoint grid, with selected 256^2 checks. Independently forced tasks use a centered finite-difference reference and a two-grid discrepancy estimate.

The primary statistic averages log error ratios over budgets and seeds within each task, then equally over tasks. Confidence intervals use 10,000 task-bootstrap draws; the primary p-value is a two-sided task sign-flip test. Secondary contrasts share a Holm correction. Transformed instances are grouped by their eight base PDEs. A task win means its average log ratio is below zero; harm means its ratio exceeds 1.05. Repeated budgets and seeds are not counted as independent tasks.

The frozen superiority criterion required search/wide ratio ≤ 0.85, a 95% upper confidence endpoint below one, $p < 0.05$, a task-harm upper bound ≤ 0.05, and median relative error $\leq 1\%$. We report that original decision alongside the measured effects. The candidate-quality explanation is an analysis of the observed result.

5 RESULTS: ONE STRONG WIDE MODEL BEATS THE TESTED SEARCH

5.1 ACCURACY AND COMPUTATION

The primary GATE-RFM+/HYBRIDWIDE task-geometric error ratio is **9.318** (95% confidence interval (CI) 7.924–10.911), with **0/60 task wins** and 60/60 task harms. Of 1,440 paired runs, 1,269 are more than 5% worse. The two-sided sign-flip test gives $p = 2.00 \times 10^{-5}$ in the direction of deterioration, so the superiority criterion is not met.

Table 2 reports accuracy and online time for all methods. HYBRIDWIDE has median relative error 6.67×10^{-5}, compared with 3.92×10^{-4} for search, and reaches 1% error in 92.3% versus 80.3% of runs. Thus both can produce accurate solutions, but the single wide solve does so more reliably on this sample. Its median time is also lower: 0.109 versus 0.312 seconds. This comparison of marginal medians is descriptive; only 539 of the 1,440 error-blind nearest-time pairs are within 10% of one another. Figure 1 shows the full budget and time curves without treating every budget pair as a wall-time match.

GATE-RFM+ improves on best-of-4, best-of-8, Hyperband, and weak-moment fitting, but is worse than best-of-2, successive halving, and QR-pruned RF. The row-adaptive comparison is inconclusive after multiplicity correction ($p = 0.340$). Appendix C gives the complete paired intervals and win counts. In particular, the large gain over conditioned tanh wide combines changes in representation and allocation; the hybrid wide control is the comparison that isolates allocation.

Table 2: Primary comparison: 1,440 runs per method over 60 PDEs, six budgets, and four seeds. Errors are relative L^2; online time excludes reference evaluation; q95 denotes the 95th percentile.

Method	Median error	q95 error	Median time (s)	Error $\leq 1\%$ (%)
Broad tanh wide	3.36	6.28	0.088	0.0
Conditioned tanh wide	2.43	4.35	0.091	0.0
Local Gaussian RF	3.05	5.63	0.122	0.0
Sine spectral RF	3.63	5.88	0.172	0.0
HybridWide	6.67e-05	0.0148	0.109	92.3
QR-pruned RF	0.000139	0.0148	0.121	92.2
Best-of-2	0.000193	0.0404	0.254	82.9
Best-of-4	0.000954	0.0604	0.247	73.1
Best-of-8	0.00324	0.116	0.310	61.4
Successive halving	0.000403	0.0465	0.261	81.9
Hyperband variant	0.000999	0.0926	0.362	71.1
Row-adaptive RF	0.00055	0.0664	0.266	79.8
Weak-moment RF	0.00141	0.134	0.129	74.0
Gate-RFM+	0.000392	0.0704	0.312	80.3

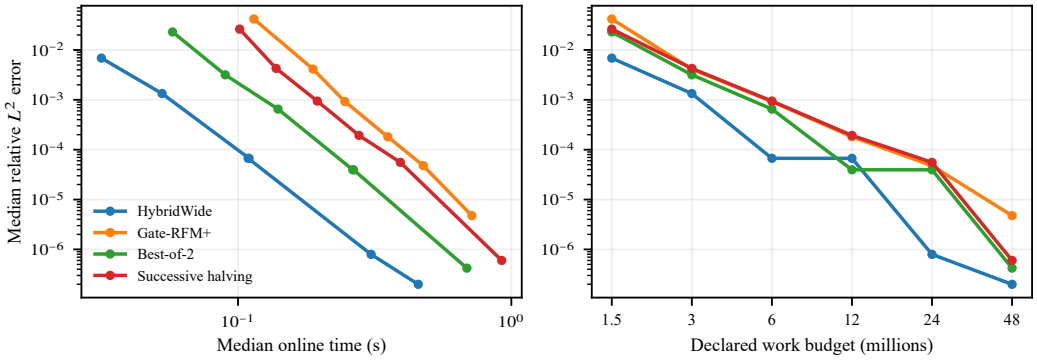

Figure 1: Median relative error versus measured online time (left) and work budget (right), aggregated over 60 tasks and four seeds at each budget. Reference evaluation is excluded from online time.

5.2 The candidate set accounts for most of the gap

The same-population oracle has a geometric error ratio of 9.255 to HybridWide. Final selection adds a factor of only 1.0068; the median selected/oracle ratio is 1.000. Multiplying the two factors recovers 9.318 up to rounding, as required by Eq. (6). On aggregate, even perfect final selection from these populations would leave nearly the entire gap to the wide baseline.

The median population contains nine fitted candidates, including intermediate widths. Their diversity is obtained by sharing computation across lineages, refitting survivors, and reserving work for verification. This is consistent with the breadth–width cost in Section 3.2: the best available searched candidate need not match a bank that spends the budget on one solve. The oracle comparison measures the resulting candidate deficit; it does not isolate width from the other costs or rule out a better adaptive search path.

Table 3: Allocation and candidate quality at each primary budget (240 paired runs per row). Widths and candidate counts are medians. Error ratios are geometric means over paired tasks and seeds; oracle selection uses reference error only after the solve.

Budget (10^6)	Wide width	Search width	Candidates	Search/wide	Oracle/wide
1.5	96	48	6	5.63	5.50
3	128	96	8	3.20	3.17
6	192	128	9	11.99	11.94
12	192	160	10	2.55	2.54
24	320	192	11	59.96	59.96
48	384	256	12	19.78	19.78

The budget-level results locate this deficit more precisely (Table 3). The selected search model is narrower at every budget, while the number of fitted candidates rises from six to twelve. At each budget, the oracle error ratio remains above one. At the two largest budgets, final selection matches the oracle in every recorded run, yet the geometric error ratios are 60.0 and 19.8. For these populations, increasing final ranking accuracy has no remaining benefit.

The gap is not monotone in budget. Between 6 and 12 million work units, the wide model stays at width 192 while the median searched width increases from 128 to 160; the paired error ratio falls from 12.0 to 2.55. At 24 million units, the wide model reaches 320 features while search reaches a median of 192, and the gap increases. Discrete width choices and the cost of refitting therefore matter when interpreting a scaling curve. The measured ratios should not be summarized by a smooth power law.

Nor do the policies spend exactly the same amount of their permitted work. Median budget utilization ranges from 44.2% to 95.6% for the wide policy and 55.4% to 84.9% for search. Both stop when their next prescribed action is unaffordable. Thus the experiment compares policies under a common ceiling, including their unused budget. A denser width ladder or a rule that reallocates the final remainder would change this allocation problem and is a concrete alternative to spending more on final verification.

This diagnosis also separates intermediate and final verification. Improving the final acceptance score cannot recover a solution absent from the population. Improving the intermediate ranking could change which lineages are widened, and hence the population itself. The present result therefore directs attention toward candidate generation and resource allocation, without claiming that every possible verifier-guided search must fail.

5.3 What the components contribute

Table 4: Component ablations on 24 primary PDEs. Ratios above one favor the full method. The score-reuse control saves acceptance cost, so it changes affordable widths as well as statistical independence.

Change from full method	Changed/full	95% CI	p_{Holm}
Tanh features only	9.5e+03	6.95e+03–1.27e+04	0.00274
Uniform fit rows	2.04	1.63–2.64	0.00274
No conditioning	0.639	0.462–0.9	0.34
One initial bank	1.54	1.19–2.02	0.173
No MAD pruning	1	0.938–1.08	1
Reuse selection sample	0.526	0.496–0.556	0.00274
Random final choice	255	136–500	0.00274
Strong residual score	0.996	0.883–1.13	1
Training residual score	0.944	0.815–1.09	1
Hybrid residual score	0.966	0.881–1.04	1
Two initial banks	1.57	1.19–2.11	0.216
Eight initial banks	1.04	0.771–1.38	1
Coordinate-specific gate	1	1–1	1
Always search	1	1–1	1

The representation and fitting distribution have large effects (Table 4, changed/full-method error ratios). Replacing the hybrid basis with tanh features raises error by about 9.50×10^3, and uniform fitting rows raise it by a factor of 2.04. Both changes are significant after Holm correction. These comparisons support the value of the shared numerical representation, which benefits the wide baseline as well as the search policy.

Physics selection is also useful: choosing a random member of the same completed population raises error by a factor of 255. This does not conflict with the small observed selection penalty. The score already finds a near-oracle member on aggregate, so replacing it with a random choice can be damaging even when further gains from final ranking are small.

Other changes provide less evidence of a benefit. Removing MAD pruning gives ratio 1.00. Strong, training, and hybrid residual scores have ratios near one, with no significant improvement. One, two, or eight initial banks do not improve on the default four-bank policy after correction. Because all

primary tasks satisfy the gate, changing to an always-search policy leaves these runs unchanged; gate utility must be assessed on other tasks.

Reusing the selection sample for final choice gives ratio 0.526. That change also saves acceptance work and makes wider fits affordable. Its gain cannot be assigned to statistical dependence alone. Likewise, the acceptance-size and dictionary-size grid charges the actual additional work. Its ratios to the wide control remain between 36.2 and 103 on the separate sensitivity subset (Appendix C). More verification is not free, and none of these settings restores search superiority there.

5.4 TRANSFER AND THE ROLE OF THE GATE

On 32 unconditioned tasks, the paired search/wide ratio is 2.427 (95% CI 1.688–3.571). On transformed instances it is 11.229 (9.638–13.230), but the eight independent base-task clusters give a Holm-adjusted $p = 0.305$. All 12 independently forced tasks are off gate and return the same HYBRIDWIDE solution. Detailed absolute outcomes for every comparator are in Appendix C.

Removing the gate on broad tasks makes error 30.1 times that of the gated policy (95% CI 6.88–153; adjusted $p = 0.00320$). The gate therefore avoids some harmful searches. Yet the gated policy still loses to always using the wide solve on aggregate. Avoiding the worst branch is a weaker property than selecting a branch that improves on the incumbent.

Coordinate normalization also illustrates this distinction. It activates search on all transformed instances. The coordinate-specific gate, which falls back more often, has ratio 0.0964 to the normalized gate, with adjusted $p = 0.305$. A more symmetric decision rule can be less useful if it routes more tasks to an inferior action. The off-gate equality is exact by construction; the accuracy of the activated branch remains an empirical property.

For relative harm tolerance $\epsilon > 0$ and gate indicator g, let $H_\epsilon = \{\mathcal{E}(\text{GATE-RFM+}) > (1 + \epsilon)\mathcal{E}(\text{HYBRIDWIDE})\}$, $\pi = \Pr(g = 1)$, and $q_\epsilon = \Pr(H_\epsilon \mid g = 1)$ under a specified task distribution. Exact fallback gives

$$\Pr(H_\epsilon) = \pi q_\epsilon. \tag{13}$$

Reducing gate activation can therefore reduce overall harm even when the activated search never improves on the wide solve. The gate-conditioned primary sample estimates performance within the activated regime; it cannot estimate the deployment activation rate π. Broad-suite performance and conditional search performance answer different questions. Appendix B gives a task-level confidence bound when π is known.

Reference checks support the main error comparisons. Across 1,904 selected fine-quadrature checks, the median relative change in reported error is 8.44×10^{-5} and the 95th percentile is 4.41%. For independently forced tasks, two-grid reference discrepancies range from 0.123% to 0.322%, setting the resolution at which those errors can be interpreted.

6 LIMITATIONS AND CONCLUSION

The experiments identify a concrete regime in which one strong wide solve outperforms test-time population search. Under the shared hybrid feature law and work budget, most of the gap is already present in the best searched candidate. Final selection contributes little additional error. This is the main distinction the analysis makes: the usefulness of a verifier within a population does not establish the usefulness of generating that population.

The main comparison covers linear PDEs under the specified sampling and budget rules. The 136 instances include correlated transformations and do not represent 136 equation families. Controlled implementations allow direct ablation, but do not constitute a comprehensive comparison with numerical PDE software. The work proxy is reproducible, while wall-time matching is incomplete. The spectral error theorem applies to a narrower operator class than the advection experiments. A predictive crossover between breadth and width, and hybrid-feature nonlinear transfer, remain open.

Allocation studies should first establish a strong wide baseline, then measure whether the searched population can beat it at the same budget and how much of that advantage selection retains. This separates candidate-generation losses from selection losses and tests search against a concrete computational alternative.

AI USE STATEMENT

OpenAI Codex was used for coding assistance and language editing. The authors reviewed all AI-assisted content and take full responsibility for the manuscript, implementation, and reported results.

REPRODUCIBILITY STATEMENT

Section 3 defines the solver, allocation, and error decomposition. Appendix A gives feature laws, sampling, parameters, and statistical procedures. Appendix B contains proofs; Appendix C gives paired comparisons, sensitivity tests, and transfer results. The artifact retains frozen configurations, per-run outcomes, candidate records, and generation scripts. Appendix D specifies 6,192 additional evaluations of nonlinear and geometric extensions and their fitting methods; Appendix E presents their results and analysis.

REFERENCES

Harris Abdul Majid, Pietro Sittoni, and Francesco Tudisco. Test-time accuracy-cost control in neural simulators via recurrent-depth. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=U2j9ZNgHqw.

Jingrun Chen, Xurong Chi, Weinan E, and Zhouwang Yang. Bridging traditional and machine learning-based algorithms for solving PDEs: The random feature method. *Journal of Machine Learning*, 1 (3):268–298, 2022. doi: 10.4208/jml.220726. URL https://global-sci.org/intro/article_detail/jml/21027.html.

Leszek Demkowicz and Jay Gopalakrishnan. A class of discontinuous petrov–galerkin methods. part II: Optimal test functions. *Numerical Methods for Partial Differential Equations*, 27(1):70–105, 2011. doi: 10.1002/num.20640.

Francisco Eiras, Adel Bibi, Rudy R. Bunel, Krishnamurthy Dvijotham, Philip Torr, and M. Pawan Kumar. Efficient error certification for physics-informed neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12318–12347, 2024. URL https://proceedings.mlr.press/v235/eiras24a.html.

Zexi Fan, Yan Sun, Shihao Yang, and Yiping Lu. Physics-informed inference time scaling for solving high-dimensional partial differential equations via defect correction. In *International Conference on Learning Representations*, 2026.

Kevin Jamieson and Ameet Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 240–248, 2016. URL https://proceedings.mlr.press/v51/jamieson16.html.

Mikhail Kuvakin, Zijian Mei, and Jingrun Chen. Weak random feature method for solving partial differential equations. *arXiv preprint arXiv:2505.00508*, 2025. URL https://arxiv.org/abs/2505.00508.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018. URL https://www.jmlr.org/papers/v18/16-558.html.

Shanda Li, Tanya Marwah, Junhong Shen, Weiwei Sun, Andrej Risteski, Yiming Yang, and Ameet Talwalkar. CodePDE: An inference framework for LLM-driven PDE solver generation. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=eG3Qy5Oux6.

Chunyang Liao. Solving partial differential equations with random feature models. *Communications in Nonlinear Science and Numerical Simulation*, 152:109343, 2026. doi: 10.1016/j.cnsns.2025.109343. URL https://doi.org/10.1016/j.cnsns.2025.109343.

Jiale Linghu, Hao Dong, and Yangshuai Wang. A structure-adaptive random feature method for high-dimensional elliptic PDEs. *arXiv preprint arXiv:2607.19786*, 2026. URL https://arxiv.org/abs/2607.19786.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045.

Louis Serrano, Jiequn Han, Edouard Oyallon, Shirley Ho, and Rudy Morel. Test-time generalization for physics through neural operator splitting. *arXiv preprint arXiv:2602.00884*, 2026. URL https://arxiv.org/abs/2602.00884.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

Jan Willem van Beek, Victorita Dolean, and Ben Moseley. Local feature filtering for scalable and well-conditioned domain-decomposed random feature methods. *arXiv preprint arXiv:2506.17626*, 2025. URL https://arxiv.org/abs/2506.17626.

Sifan Wang, Xinling Yu, and Paris Perdikaris. When and why PINNs fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, 449:110768, 2022. doi: 10.1016/j.jcp.2021.110768.

Chenxi Wu, Min Zhu, Qinyang Tan, Yadhu Kartha, and Lu Lu. A comprehensive study of non-adaptive and residual-based adaptive sampling for physics-informed neural networks. *Computer Methods in Applied Mechanics and Engineering*, 403:115671, 2023. doi: 10.1016/j.cma.2022.115671. URL https://arxiv.org/abs/2207.10289.

A Implementation and experimental setup

Feature distributions and fitting. The feature type is assigned by index within each eight-feature block, giving three tanh ridges, two exponentials, two Gaussians, and one sine product. Let z_j be an anchor, d_j a unit direction, and $s = \operatorname{sign}(\beta)$, with zero components assigned positive sign. The four feature forms are

$$\begin{aligned} \phi_j^{\tanh}(x) &= q(x)\tanh(w_j^\top x + b_j), & \phi_j^{\exp}(x) &= q(x)\exp[-a_j^\top d_{\mathrm{out}}(x)], \\ \phi_j^{\mathrm{gauss}}(x) &= q(x)\exp[-\|a_j \odot (x - z_j)\|_2^2], & \phi_j^{\mathrm{sine}}(x) &= \sin(\pi k_{j1} x_1)\sin(\pi k_{j2} x_2). \end{aligned} \tag{14}$$

Here $\odot$ denotes componentwise multiplication, and $d_{\mathrm{out},\ell}(x) = 1 - x_\ell$ for $\beta_\ell \geq 0$ and x_ℓ otherwise. The operator evaluations use analytic first and second derivatives, including the diffusion-gradient term when ϵ varies.

Anchors are sampled from a 48^2 midpoint probe with probability proportional to $0.08 + \sqrt{|f(x_i)|/(\max_r |f(x_r)| + 10^{-30})}$. For tanh features, d_j is a normalized two-dimensional standard Gaussian, $w_j = h\gamma_j d_j$, and $b_j = -w_j^\top z_j + \xi_j$, with $\xi_j \sim N(0, 0.15^2)$. The multiplier γ_j is uniform on $\{0.35, 0.7, 1, 1.8, 3\}$ and $h = \mathrm{clip}(2 + 1.5 n_{\mathrm{sign}} + 3\kappa_f, 2, 80)$. Here κ_f is the forcing-concentration statistic defined below, and ϵ_{ref} is the nominal diffusion coefficient. The quantity n_{sign} counts adjacent forcing sign changes on the probe, divided by the grid side and rounded. Exponential rates have $a_{j\ell} = \max(|\beta_\ell|/\max(\epsilon_{\mathrm{ref}}, 10^{-9}), 2)\exp(U_{j\ell})$, with $U_{j\ell}$ uniform on $[-1.7, 0.6]$. Gaussian inverse scales are independently uniform on $\{3, 8, 20, 50\}$. Sine indices are positive integer pairs ordered by $k_1^2 + k_2^2$, with lexicographic tie-breaking.

For fitting, define outflow coordinates $z_\ell = x_\ell$ if $\beta_\ell \geq 0$ and $z_\ell = 1 - x_\ell$ otherwise. The density is

$$p(x) = 0.4 + 0.6(36 z_1^5 z_2^5). \tag{15}$$

This is a mixture of uniform points and independent $\mathrm{Beta}(6,1)$ outflow coordinates. Every bank is drawn at maximum width; each fit uses its specified prefix. The ladder contains widths 32, 48, 64, 96, 128, 160, 192, 256, 320, 384, 448, and 512.

If $WAD^{-1} = U\,\mathrm{diag}(s_i)V^\top$, the coefficient computation is

$$\alpha = D^{-1}V\,\mathrm{diag}(h_i)U^\top Wb, \qquad h_i = \begin{cases} s_i/(s_i^2+\lambda), & s_i > \max(10^{-12}s_{\max}, 10^{-30}), \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

The reported numerical rank counts retained singular values. Matrix and residual evaluations use chunks of 256 points. One CPU thread is used for the linear algebra.

Gate and search rule. For a PDE task τ, the forcing probe has $N_p = 48^2$ points. Let $\pi_i = |f(x_i)|/(\sum_r |f(x_r)| + 10^{-15})$ on that probe. Set $\kappa_f = N_p \sum_i \pi_i^2$ and $\mu = \sum_i \pi_i x_i$. Reflect μ_ℓ to $1-\mu_\ell$ when $\beta_\ell < 0$, then sort the two coordinates in descending order to obtain $\widetilde{\mu}$. The boundary statistic b_f sums forcing magnitude over all four width-$1/8$ strips and divides by twice the total forcing magnitude. Corners enter two strips. Roughness ρ_f is the root-mean-square (RMS) magnitude of the finite grid gradient of $|f|$, with unit grid-index spacing, divided by its RMS magnitude. The gate is

$$g(\tau) = \mathbf{1}\left\{ \begin{array}{c} \|\beta\|_2 \ge 0.5, \quad c \ge 0, \quad 40 \le \|\beta\|_2/|\epsilon_{\mathrm{ref}}| \le 160, \\ \kappa_f \ge 20, \quad \widetilde{\mu}_1 \ge 0.88, \quad \widetilde{\mu}_2 \ge 0.85, \quad b_f \ge 0.65, \quad 0.20 \le \rho_f \le 0.65 \end{array} \right\}. \tag{17}$$

All thresholds are fixed before test evaluation. The coordinate-specific ablation uses μ directly, without reflection or sorting.

In Table 1, $\mathrm{MAD}(S_A) = \mathrm{median}_{j\in A}\,|S_j - \mathrm{median}_{r\in A}\,S_r|$. The implementation adds 10^{-30} to the pruning threshold to handle numerical ties. Promotion refits the enlarged basis from scratch.

Comparator definitions. Best-of-K uses the largest common width for which K full fit-and-verify actions fit the budget; if none exists, it uses one smallest-width bank. Successive halving drops MAD pruning and returns the candidate with the best selection-sample score, without a separate acceptance charge. Hyperband divides the budget into two brackets with four and two initial banks, respectively. The four-bank bracket starts at minimum width; the two-bank bracket starts at its largest jointly affordable width. Both halve their active sets before promotion.

QR-pruned RF applies column-pivoted QR to the weighted, scaled matrix and retains columns whose diagonal pivots exceed $\max(10^{-10}|R_{11}|, 10^{-14})$, before SVD fitting. Weak-moment RF projects the weighted residual onto sine tests, using $\min(N_f/2, \max(2M, 64))$ modes. These controls add $2N_fM^2/64$ and $4N_fM^2/64$, respectively, to the action cost. Row-adaptive RF fits a first solution, evaluates residuals on a uniform pool of $4N_f$ points, and samples new rows with probability proportional to $0.1 + |r|/(\max|r| + 10^{-30})$. It charges the pool and second fit before refinement and uses importance weights for the discrete resampling law.

Broad tanh wide uses uniform anchors and $h = 12$; conditioned tanh wide uses the forcing-conditioned anchors above. Local and spectral RF replace the feature mixture with Gaussian and sine features, respectively. Each single-bank control chooses its largest affordable width. The strong score is the relative RMS residual; the training score is the weighted fitting residual; the hybrid score adds 0.03 times the strong score to the weak score. The random-choice ablation samples uniformly from the same completed candidate list. All other settings are held at their specified values except for the named intervention.

Manufactured tasks. The default coefficients are $\epsilon_{\mathrm{ref}} \sim U[0.014, 0.026]$, $\beta = a(\cos\vartheta, \sin\vartheta)$ with $a \sim U[1, 1.4]$, $\vartheta \sim U[0.45, 1.1]$, and $c \sim U[0, 0.8]$. A boundary envelope multiplies an exponential with coordinate rates drawn from $[20, 36] \times [15, 29]$. Each exponential component is normalized by its maximum magnitude on a 128^2 midpoint grid before components are mixed. Superposed-layer tasks add a second exponential with rates scaled independently by factors in $[0.6, 0.9]$ and amplitude in $[0.15, 0.45]$. Multiscale tasks add a sine mixture with amplitude in $[0.02, 0.15]$. The mixture is $\sin(\pi k_1 x_1)\sin(\pi k_2 x_2) + 0.2\sin(\pi(k_2+2)x_1)\sin(\pi(k_1+1)x_2)$, with independent k_1, k_2 uniform on $\{1,2,3,4,5\}$. Variable-diffusion tasks use periodic modulation of the reference diffusion, with modulation amplitude in $[0.1, 0.35]$ and coordinate frequencies $(2, 3)$. The exact field defines $f = \mathcal{L}u$ for evaluation; online algorithms receive only the operator and forcing.

Broad tasks are sampled without gate conditioning from the variable, multiscale, superposed, and smooth generators. Smooth tasks use a sine mixture, zero advection, diffusion in $[0.3, 1]$, and periodic diffusion modulation. Transformation tests use four square symmetries per base task, including the identity, with the operator and forcing transformed together. For independently forced tasks,

$$f(x) = \exp[-a\|x - z\|_2^2] + 0.3\sin(2\pi x_1)\sin(3\pi x_2), \quad z_\ell \sim U[0.3, 0.8], \quad a \sim U[15, 55]. \tag{18}$$

Their reference solution is constructed only during post-solve evaluation.

Table 5: Evaluation matrix for the reported linear-PDE study. Subset experiments reuse tasks; their instance counts must not be added to 136.

Experiment	Task instances	Evaluations
Default comparisons across four suites	136	45,696
Component ablations, including two controls	24	3,456
Gate transfer ablations	64	1,024
Acceptance size and dictionary size	12	864
Finite-difference resolution curves	136	544
Total		51,584

Pairing and inference. For methods A, B, define

$$d_\tau = \frac{1}{|\mathcal{P}_\tau|} \sum_{(b,s)\in\mathcal{P}_\tau} \log \frac{\mathcal{E}(A_{\tau,b,s})}{\mathcal{E}(B_{\tau,b,s})}, \qquad R_{A/B} = \exp\left(\frac{1}{n}\sum_{\tau=1}^{n} d_\tau\right). \tag{19}$$

Here n counts tasks, τ indexes a task, and $\mathcal{P}_\tau$ contains its paired budget indices b and seed indices s. Transformed tasks also average within base-PDE clusters before inference. Bootstrap intervals resample these independent task or cluster means. Sign-flip tests enumerate all sign patterns for at most 16 units and use 50,000 random sign patterns otherwise, with an add-one correction. Holm-adjusted p-values are retained from the complete frozen contrast family. Reported median errors and times aggregate individual runs and therefore need not order methods in the same way as $R_{A/B}$.

Component ablations use budgets $3, 12, 48$ million and seeds $8901, 8902, 8903$. Gate and verifier sensitivity use budgets $6, 24$ million, with paired seed sets $(8961, 8962)$ and $(8951, 8952)$, respectively. Default comparisons use seeds 8801–8804. The configuration fixes all task parameters, seed assignments, methods, and budgets. Missing or failed pairs invalidate the registered primary decision; none occurred. The reference-based oracle and fine-grid errors are computed after the online timer stops.

Reference resolution and resources. Manufactured references use analytic fields. For independently forced tasks, centered finite differences on grids with 128 and 256 intervals give the two-grid discrepancy; the finer solution is interpolated for evaluation. Separate upwind finite-difference runs use 16, 32, 64, and 128 intervals. These curves are retained in the artifact. Workers use one CPU and 2 GiB, with at most two concurrent workers, single-thread BLAS, and no GPU. Per-run records contain fitted candidate widths, costs, selected ranks, conditions, errors, and online times.

Representative solution fields. Figures 2 and 3 visualize completed primary-study solutions. The examples are selected without inspecting their spatial fields: at the 24-million work budget, within each primary structural family we take the task closest to the median GATE-RFM+/HYBRIDWIDE geometric error ratio, followed by the seed closest to that task's median ratio. The solvers are rerun and their 128^2 evaluation errors are checked against the reported results before plotting. Reference values enter only after both policies terminate.

B PROOFS AND ADDITIONAL BOUNDS

Proof of Theorem 1. Condition on the full search history, including fitting and adaptive decisions. Its candidate set is fixed and the acceptance points remain independent. Hoeffding's inequality gives, for each candidate and test,

$$\Pr\{|\widehat{m}_{\theta k} - m_{\theta k}| > t \mid \Theta\} \le 2\exp[-n_a t^2/(2R^2)]. \tag{20}$$

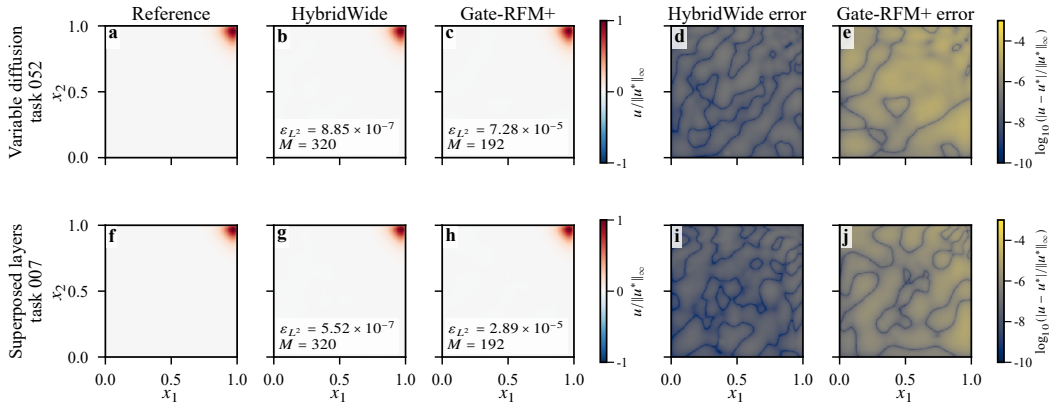

Figure 2: Representative primary-study solution fields at a work budget of 24×10^6. Rows show the variable-diffusion and superposed-layer cases selected by the deterministic median-ratio rule. The first three columns use a common normalized solution scale; the last two show pointwise absolute error on the same fixed logarithmic scale. Although both methods are visually close to the reference at field scale, the error maps expose the roughly two-order accuracy gap favoring HYBRIDWIDE. Insets report relative L^2 error and selected width M.

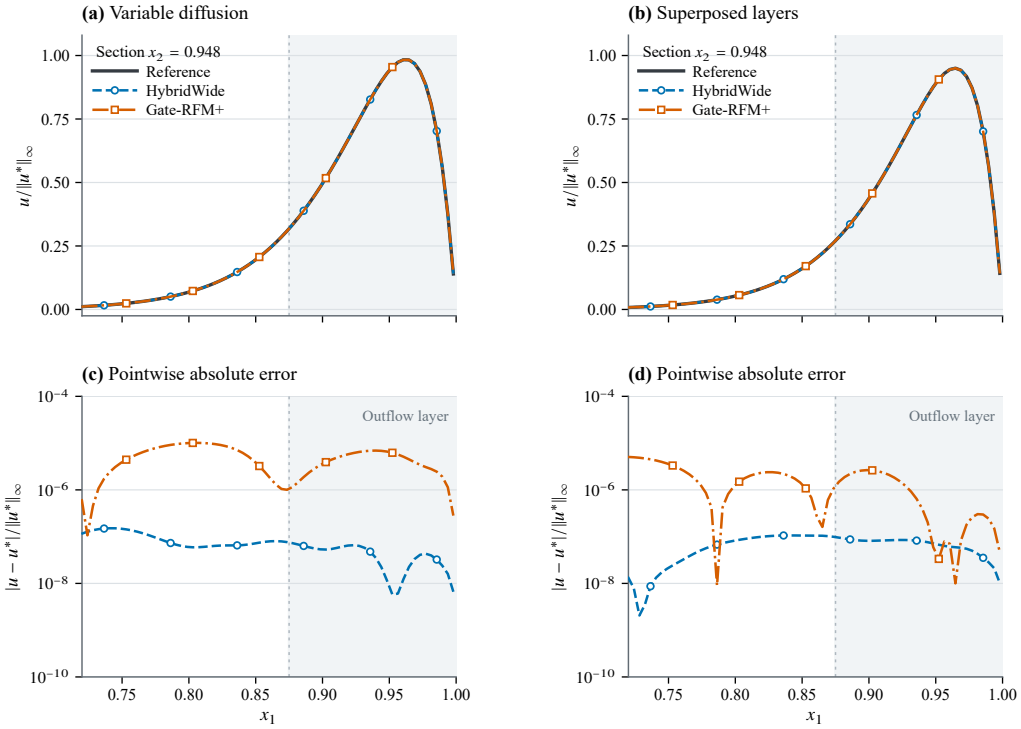

Figure 3: Boundary-layer cross-sections for the same two cases. Top panels overlay the normalized reference and numerical solutions at $x_2 \approx 0.95$; bottom panels separate their normalized pointwise errors on a logarithmic scale. The gray band marks the outer one-eighth of the domain used by the layer-error evaluation. The profiles resolve how the discrepancy varies through and outside the sharp outflow layer and show that it is systematically smaller for HYBRIDWIDE. These plots illustrate spatial errors; the PDE task remains the inferential unit.

A union bound over JK_v pairs, with $t = R\sqrt{2\log(2JK_v/\eta)/n_a}$, bounds the vector error by $\sqrt{K_v}t = \delta$. The reverse triangle inequality then bounds $|\widehat{V}_a - V|$. If θ_* minimizes V, the empirical minimizer satisfies

$$V(\widehat{\theta}_a) \le \widehat{V}_a(\widehat{\theta}_a) + \delta \le \widehat{V}_a(\theta_*) + \delta \le V(\theta_*) + 2\delta. \tag{21}$$

The forcing normalization in the score is common to all candidates, so it does not change the minimizer. Averaging the conditional failure probability over search histories gives the unconditional guarantee.

A challenger v improves on incumbent b in population weak score if $\widehat{V}_a(v) + 2\delta < \widehat{V}_a(b)$. If additionally $aV(w) \le \mathcal{E}(w) \le b_0 V(w)$ holds for both candidates, with $0 < a \le b_0$, then $b_0[\widehat{V}_a(v) + \delta] < a[\widehat{V}_a(b) - \delta]$ suffices for lower solution error. These comparison constants are assumptions; they are not supplied by moment concentration alone.

Proof of Theorem 2. On the unit square, $e_k(x) = 2\sin(\pi k_1 x_1)\sin(\pi k_2 x_2)$ and $\lambda_k = \nu\pi^2(k_1^2 + k_2^2) + c$. The weak equation gives $\langle v - u, e_k\rangle = \langle r_v, e_k\rangle/\lambda_k$. Parseval's identity yields

$$\|v - u\|_{L^2}^2 = V_I(v)^2 + \sum_{k \notin I} \frac{|\langle r_v, e_k\rangle|^2}{\lambda_k^2} \le V_I(v)^2 + T_I(v)^2. \tag{22}$$

Nonnegativity of the omitted sum gives the lower bound. For an error-minimizing candidate v_*, the simultaneous sampling bound implies $V_I(\widehat{v}) \le V_I(v_*) + 2\delta \le \|v_* - u\|_{L^2} + 2\delta$. Combining this with $\sqrt{a^2 + b^2} \le a + b$ proves the selection bound.

For $K = |I|$ spectral tests and J candidates, if $|r_v(x)e_k(x)/\lambda_k| \le R$ uniformly, the same conditional argument as above gives $\delta = R\sqrt{2K\log(2JK/\eta)/N}$ from N acceptance points. The residual tail uses the population L^2 norm. A sampled residual norm requires its own error bound before it can replace that quantity.

Stability and the finite dictionary. For a bounded Lipschitz domain with Poincaré constant C_P, diffusion $\epsilon \ge \epsilon_0 > 0$, and $c - \frac{1}{2}\nabla \cdot \beta \ge 0$, let $e = v - u \in H_0^1$ and $r = \mathcal{L}v - f \in H^{-1}$. Testing $\mathcal{L}e = r$ with e gives

$$\langle r, e\rangle = \int_\Omega \epsilon|\nabla e|^2 + \int_\Omega (c - \tfrac{1}{2}\nabla \cdot \beta)e^2 \ge \epsilon_0 \|\nabla e\|_{L^2}^2. \tag{23}$$

Duality gives $\langle r, e\rangle \le \|r\|_{H^{-1}}\|\nabla e\|_{L^2}$. Combining this with the preceding estimate and $\|e\|_{L^2} \le C_P\|\nabla e\|_{L^2}$ proves Eq. (12).

Fallback and deployment risk. With a closed gate, GATE-RFM+ and HYBRIDWIDE use the same bank seed, fitting points, matrix, and coefficient computation. Their numerical outputs and charged solve work are identical; the gate can still add measured runtime. Thus $\Pr(H_\epsilon \mid g = 0) = 0$, and the law of total probability gives Eq. (13). For known π, m independent gate-activated evaluation tasks, and h harms, a one-sided confidence bound of level $1 - \alpha$ is $\pi\ \mathrm{Beta}^{-1}(1 - \alpha; h + 1, m - h)$ for $h < m$, where Beta^{-1} is the beta-distribution quantile and $0 < \alpha < 1$; the conditional upper bound is one when $h = m$. This follows from the binomial tail bound for q_ϵ. Dependence within task, including repeated seeds and transformed copies, must be handled at the task level.

C SUPPLEMENTARY COMPARISONS, SENSITIVITY, AND TRANSFER

Paired method comparisons. Table 6 complements the absolute summaries in the main text with paired effects and uncertainty. Each ratio divides the error of GATE-RFM+ by that of the listed comparator, using the primary task aggregation.

Verification sensitivity. The sensitivity grid changes N_a and K_v jointly with their charged costs. Its 12-task subset, two budgets, and two seeds differ from the primary matrix, so its ratios should be compared within this grid.

Transfer results. Tables 8–10 report all default algorithms on the three additional suites. Table 11 uses the separate gate-ablation budgets and seeds; its ratios differ from the default-suite ratios for that reason.

Table 6: Prespecified paired contrasts. Primary HybridWide contrast uses the raw two-sided task sign-flip p; other rows use global Holm correction. Ratios below one favor Gate-RFM+.

Comparator	Gate/comparator	95% CI	Task wins	p
HybridWide	9.318	7.92–10.9	0/60	2e-05
Conditioned tanh wide	0.0001733	0.00015–0.000199	60/60	0.00274
Best-of-2	2.373	2.12–2.66	2/60	0.00274
Best-of-4	0.5537	0.509–0.604	58/60	0.00274
Best-of-8	0.2154	0.194–0.24	60/60	0.00274
Successive halving	1.465	1.4–1.53	3/60	0.00274
Hyperband variant	0.4437	0.398–0.499	57/60	0.00274
QR-pruned RF	4.295	3.62–5.06	2/60	0.00274
Weak-moment RF	0.3212	0.238–0.425	54/60	0.00274
Row-adaptive RF	0.8048	0.677–0.957	37/60	0.34

Table 7: Acceptance and dictionary sensitivity. Ratios compare Gate-RFM+ with HybridWide on the same subset and budgets.

N_a	K_v	Search/wide	95% CI	p_{Holm}
128	16	36.2	23.3–58.5	0.0327
128	32	38.5	22.9–67.9	0.0327
128	64	43.1	25.8–75.5	0.0327
512	16	36.5	23.1–57.9	0.0327
512	32	38.0	23.3–65.8	0.0327
512	64	42.2	25.4–72.0	0.0327
2048	16	89.0	58.2–142	0.0327
2048	32	93.9	58.6–161	0.0327
2048	64	103	65.5–172	0.0327

Table 8: Broad tasks: relative L^2 error, median online time, and fraction reaching 1% error. All runs completed.

Method	Median error	q95 error	Time (s)	Fraction $\leq 1\%$	Failed
Broad tanh wide	2.55	5.41	0.0887	0.0638	0
Best-of-4	0.0218	0.884	0.246	0.441	0
Row-adaptive RF	0.0109	0.633	0.293	0.495	0
Sine spectral RF	2.98	5.03	0.173	0.25	0
HybridWide	0.000364	0.523	0.11	0.654	0
Successive halving	0.00425	0.797	0.259	0.548	0
Weak-moment RF	0.00974	1.67	0.124	0.501	0
Upwind finite differences	0.148	0.635	0.0131	0.164	0
Conditioned tanh wide	2.12	4.18	0.0913	0.0273	0
Best-of-8	0.0387	1.08	0.323	0.367	0
QR-pruned RF	0.001	0.542	0.116	0.622	0
Gate-RFM+	0.000804	0.523	0.216	0.608	0
Best-of-2	0.0038	0.72	0.237	0.547	0
Hyperband variant	0.0289	0.933	0.357	0.424	0
Local Gaussian RF	2.52	5.41	0.123	0	0

D EXTENSION DESIGN AND NONLINEAR FITTING

The additional experiments examine three nonlinear PDE settings and a linear suite with curved domains and near-threshold square problems. They use boundary-conforming tanh solvers and their associated allocation policies. We call their single-bank baseline TANHWIDE to distinguish it from the hybrid-feature HYBRIDWIDE in the main comparison. The extensions test whether the allocation question remains useful under different operators, boundary conditions, and coefficient solvers; they do not reuse the final hybrid solver unchanged.

Each nonlinear suite contains eight challenging tasks and eight smooth controls. The linear extension contains twelve disk layers, twelve square tasks near the gate boundary, six smooth disk controls, and six variable-diffusion square controls. All methods use seeds 8981–8983. Together with the 51,584 evaluations described in Appendix A, these account for all 57,776 evaluations in the combined study.

Table 9: Transformed tasks: relative L^2 error, median online time, and fraction reaching 1% error. All runs completed.

Method	Median error	q95 error	Time (s)	Fraction $\leq 1\%$	Failed
Broad tanh wide	2.79	5.98	0.0878	0	0
Best-of-4	0.000655	0.0578	0.247	0.749	0
Row-adaptive RF	0.000422	0.0594	0.257	0.81	0
Sine spectral RF	2.94	5.42	0.163	0	0
HybridWide	4.62e-05	0.0119	0.108	0.943	0
Successive halving	0.000338	0.0417	0.245	0.842	0
Weak-moment RF	0.000772	0.0791	0.12	0.793	0
Upwind finite differences	0.217	0.637	0.0125	0	0
Conditioned tanh wide	2.21	3.99	0.09	0	0
Best-of-8	0.00237	0.114	0.302	0.639	0
QR-pruned RF	9.28e-05	0.0119	0.119	0.943	0
Gate-RFM+	0.000355	0.0679	0.307	0.819	0
Best-of-2	0.000142	0.035	0.223	0.861	0
Hyperband variant	0.000694	0.0687	0.346	0.743	0
Local Gaussian RF	2.55	5.47	0.121	0	0

Table 10: Independently forced tasks: relative L^2 error, median online time, and fraction reaching 1% error. All runs completed.

Method	Median error	q95 error	Time (s)	Fraction $\leq 1\%$	Failed
Broad tanh wide	0.498	0.752	0.0834	0	0
Best-of-4	0.101	0.41	0.221	0.132	0
Row-adaptive RF	0.0564	0.355	0.242	0.181	0
Sine spectral RF	0.726	0.888	0.166	0	0
HybridWide	0.0162	0.178	0.103	0.41	0
Successive halving	0.0623	0.356	0.231	0.198	0
Weak-moment RF	0.118	3.4	0.126	0.128	0
Upwind finite differences	0.108	0.277	0.012	0	0
Conditioned tanh wide	0.544	0.798	0.0846	0	0
Best-of-8	0.144	0.544	0.284	0.0833	0
QR-pruned RF	0.0255	0.178	0.116	0.326	0
Gate-RFM+	0.0162	0.178	0.104	0.41	0
Best-of-2	0.0474	0.293	0.203	0.271	0
Hyperband variant	0.103	0.518	0.3	0.142	0
Local Gaussian RF	0.72	0.994	0.115	0	0

Table 11: Gate variants on the separate transfer-ablation matrix. Wide denotes HybridWide; Gate denotes Gate-RFM+. Ratios are method/baseline.

Suite	Method	Baseline	Ratio	95% CI
Broad	Coordinate-specific gate	Wide	3.62	2.19–6.09
Broad	Always search	Wide	109	33.1–412
Broad	Gate-RFM+	Wide	3.62	2.21–6.07
Transformed	Coordinate-specific gate	Wide	2.28	1.92–2.61
Transformed	Always search	Wide	23.6	16.8–32.7
Transformed	Gate-RFM+	Wide	23.6	17.0–32.5
Broad	Coordinate-specific gate	Gate	1	1–1
Broad	Always search	Gate	30.1	6.88–153
Transformed	Coordinate-specific gate	Gate	0.0964	0.0763–0.122
Transformed	Always search	Gate	1	1–1

The task instances in the extensions are additional to the 136 linear-study instances; repeated budgets, seeds, and policies remain repeated measurements of the same task.

Shared representation within an extension. A trial function has the form $q_\Omega(z)\tanh(w_j^\top z + b_j)$, where q_Ω enforces the relevant homogeneous constraints. The broad proposal mixes scales from $\{1, 3, 10, 30, 60, 100\}$ with log-normal magnitude jitter of standard deviation 0.22, Gaussian

Table 12: Extension evaluation matrix. All suites use three paired seeds. Budgets are in millions of the extension work units.

Suite	Tasks	Budgets	Policies	Runs/policy	Total
Steady vector Burgers	16	6, 24	8	96	768
Space–time Burgers	16	3, 12	8	96	768
Homogeneous Cole–Hopf	16	3, 12	8	96	768
Curved/near-threshold linear	36	0.6, 1.2, 2.4, 4.8	9	432	3,888
Total	84				6,192

directions normalized to unit length, and bias jitter of standard deviation 0.15. Conditioned proposals change scale and anchor placement using observable equation and forcing summaries. They do not change the reference field or use its interior values for model selection. For the homogeneous Burgers problem, the supplied initial and boundary data determine a lift before the correction is fitted.

Allocation controls. TANHWIDE allocates the available solve budget to one broad bank; conditioned TANHWIDE changes the proposal while retaining one-bank allocation. Uniform bank search distributes work across banks, and holdout halving promotes banks ranked by independent residual evaluations. Row adaptation changes fitting points. Row–column adaptation changes both the points and trial space. Guarded nonlinear search preserves a wide incumbent and accepts a challenger only through its held-out comparison. The linear extension additionally compares physics-based and geometry-based gates and an incumbent portfolio.

The extension tables report every evaluated policy, using pooled absolute errors, upper-tail errors, online times, and accuracy fractions. These are descriptive summaries of the specified suite and budget matrix. In particular, a ratio of two tabulated median errors is not the task-paired effect used in the primary analysis. The tables retain capped nonlinear iterates: coefficient stationarity and small PDE residual are assessed separately from completion of the allotted iterations.

Nonlinear fitting and stopping criteria. For a fixed bank, let a collect the coefficients, $r_M(a) \in \mathbb{R}^n$ be the full collocation residual, and $J_M(a)$ its analytic Jacobian. The nonlinear coefficient problem is

$$F_M(a) = \frac{1}{2n}\|r_M(a)\|_2^2, \qquad \nabla F_M(a) = \frac{1}{n} J_M(a)^\top r_M(a). \tag{24}$$

Here n counts scalar residual entries, including both equation components for the vector system. The solver uses a column-scaled Levenberg–Marquardt step. With $D_{jj} = \max(\|J_{:j}\|_2/\sqrt{n}, 10^{-12})$, it solves the augmented least-squares problem equivalent to

$$(J^\top J + \lambda D^\top D)p = -J^\top r, \qquad \lambda > 0, \tag{25}$$

where p is the coefficient step and λ is damping. The trial coefficients are accepted only if the residual norm decreases.

Proposition 2 (Descent direction)**.** *If $J^\top r \neq 0$ and D is nonsingular, the step in Eq. (25) satisfies $(J^\top r)^\top p < 0$.*

Proof. The coefficient matrix is positive definite, and $p \neq 0$. Multiplication by $p^\top$ gives $(J^\top r)^\top p = -\|Jp\|_2^2 - \lambda\|Dp\|_2^2 < 0$. □

Descent concerns the coefficient objective; it guarantees neither a global minimizer nor an accurate trial space. Damping increases after a rejected step and decreases after an accepted step.

The normalized stationarity measure is

$$\chi_M(a) = \frac{\|J_M(a)^\top r_M(a)/n\|_\infty}{(\|J_M(a)\|_F/\sqrt{n})(\|r_M(a)\|_2/\sqrt{n}) + 10^{-15}}. \tag{26}$$

Here $\|\cdot\|_F$ is the Frobenius norm and $\|\cdot\|_\infty$ is the largest absolute vector entry. The measure is recomputed at the returned coefficients with the normalized residual. A small final coefficient update is also a stopping condition, but is not substituted for χ_M in the reported stationarity counts.

Table 13: Common nonlinear solver settings in the three extension suites.

Setting	Value
Maximum coefficient iterations	30
Initial damping	10^{-2}
Damping increase / decrease factors	10 / 0.3
Maximum damping trials per iteration	10
Normalized residual tolerance	10^{-6}
Normalized stationarity tolerance	10^{-5}
Relative coefficient-step tolerance	10^{-8}
Ridge floor	10^{-10}

To distinguish approximation and optimization, define $\rho_M = \inf_a \|r_M(a)\|_2/(\sqrt{n}\,\sigma)$ for a fixed positive residual normalization σ. If a residual threshold is below ρ_M, no coefficient algorithm can satisfy it in that bank. Conversely, $J_M(a)^\top r_M(a) = 0$ may hold with nonzero residual. Reporting the final residual and χ_M therefore shows whether the coefficients are stationary and whether the PDE residual is small, without treating either as a replacement for the post-solve solution error.

E Nonlinear and geometric extensions: results and analysis

Steady vector Burgers. The steady extension solves the two-component system

$$-\nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + c\mathbf{u} = \mathbf{f} \quad \text{in } (0,1)^2, \qquad \mathbf{u}|_{\partial\Omega} = \mathbf{0}. \tag{27}$$

Both components share the same bank, $u_i(x) = \sum_{j=1}^{M} a_{ji}\phi_j(x)$, with $\phi_j = x_1(1-x_1)x_2(1-x_2)\tanh(w_j^\top x + b_j)$. The boundary condition holds for every coefficient vector. The two components have separate coefficients and are coupled through convection.

For component indices $i, q \in \{1,2\}$ and coefficient a_{jq}, differentiation with Kronecker delta δ_{iq} gives

$$\frac{\partial r_i}{\partial a_{jq}} = \delta_{iq}\big[-\nu\Delta\phi_j + \mathbf{u}\cdot\nabla\phi_j + c\phi_j\big] + \phi_j\,\partial_q u_i. \tag{28}$$

Thus changing one velocity component also changes the transport of the other. The coefficient solver uses the full coupled Jacobian, with $2M$ unknowns and two residual equations per fitting point. The work proxy accounts for both components and the prescribed nonlinear iteration cap.

The sixteen tasks comprise eight cross-layer fields and eight smooth controls. Manufactured fields define $\mathbf{f}$ through Eq. (27); the online solver receives the resulting operator, forcing, and boundary conditions. Reference velocity fields are used only in post-solve evaluation. The error uses the joint discrete L^2 norm of both components, so a good fit to only one component is insufficient.

Table 14: Steady vector Burgers: all evaluated policies. Errors and times summarize individual runs; the success fraction uses relative L^2 error $\le 1\%$.

Method	Median error	q95 error	Time (s)	Success	Failed
TanhWide	0.222	1.12	0.115	0.0312	0
Conditioned TanhWide	0.281	1.06	0.0921	0	0
Uniform bank search	0.239	0.979	0.978	0	0
Conditioned bank search	0.384	1.37	1.17	0	0
Holdout halving	0.304	1.14	0.472	0	0
Guarded nonlinear search	0.22	0.804	0.236	0.0312	0
Row adaptation	0.327	1.36	0.317	0	0
Row–column adaptation	0.436	1.55	0.368	0	0

The guarded nonlinear policy and TanhWide have similar median errors, 0.220 and 0.222. The guarded policy has a smaller pooled 95th-percentile error, 0.804 versus 1.12, but its median online time is 0.236 rather than 0.115 seconds. Both reach 1% error in only 3.12% of these runs. Conditioned one-bank fitting and the unguarded search controls do not improve the pooled median over the broad wide baseline. The results therefore describe a difficult approximation regime, rather than high-accuracy solutions across the suite.

This comparison is also sensitive to what is called a successful solve. The guarded policy satisfies the stationarity threshold in 94 of 96 selected solutions, compared with 91 of 96 for TANHWIDE; neither passes the separate 10^{-6} residual threshold. The median error comparison alone would conceal that distinction. Table 18 reports these convergence measures for all policies, including less favorable stationarity outcomes.

Space–time Burgers. The scalar space–time problem is

$$u_t - \nu u_{xx} + uu_x + cu = f, \qquad (x,t) \in (0,1)^2, \quad u(0,t) = u(1,t) = u(x,0) = 0. \tag{29}$$

The coordinate t is time, not a second spatial direction. A bank uses

$$\phi_j(x,t) = x(1-x)t\,\tanh(w_{jx}x + w_{jt}t + b_j). \tag{30}$$

These functions impose the initial and spatial boundary conditions exactly while leaving the terminal-time edge unconstrained. The method fits the whole space–time strip in one collocation problem; it is not a sequential time-stepping scheme.

For $u = \sum_j a_j\phi_j$, the analytic Jacobian column is

$$J_{:j} = \partial_t\phi_j - \nu\partial_{xx}\phi_j + u\,\partial_x\phi_j + (u_x + c)\phi_j. \tag{31}$$

The nonlinear term consequently depends on both the feature value and its spatial derivative. Selection and acceptance evaluate the complete space–time residual, while reference errors integrate over space and time. The sixteen tasks again contain eight layer problems and eight smooth controls, evaluated at two budgets and three paired seeds.

Table 15: Space–time Burgers: all evaluated policies. Errors and times summarize individual runs; the success fraction uses relative L^2 error $\leq 1\%$.

Method	Median error	q95 error	Time (s)	Success	Failed
TanhWide	0.058	0.585	0.0942	0.156	0
Conditioned TanhWide	0.0562	0.407	0.0728	0.188	0
Uniform bank search	0.235	1.03	0.642	0.0833	0
Conditioned bank search	0.264	1.23	0.722	0	0
Holdout halving	0.114	0.744	0.269	0.115	0
Guarded nonlinear search	0.0593	0.585	0.213	0.156	0
Row adaptation	0.167	1.24	0.254	0.0625	0
Row–column adaptation	0.274	1.49	0.232	0.0625	0

Conditioned TANHWIDE has the lowest pooled median error, 0.0562, and reaches 1% error in 18.8% of runs. Broad TANHWIDE has median 0.0580, with 15.6% of runs below 1% error. Guarded nonlinear search has median 0.0593 and the same accuracy fraction, with more than twice the broad baseline's median online time. Uniform and conditioned bank search have substantially larger pooled median errors, 0.235 and 0.264.

Fitting several banks divides computation among nonlinear solves before residual comparison. Conditioning a single trial space can have a different effect from repeatedly fitting and selecting among conditioned spaces. That distinction is visible without assuming that every change in median error is a statistically significant task-level improvement.

The stationarity counts further distinguish optimization from allocation. All 96 conditioned-wide selections meet the stationarity threshold, as do 95 broad-wide selections and 94 guarded selections. None meets the strict residual threshold. The analysis below gives the residual stability statement for this equation and identifies the derivative assumptions needed to turn a full residual bound into solution-error control.

Residual stability for nonlinear PDEs.

Theorem 3 (Space–time residual stability)**.** *Let u solve Eq. (29) with $\nu > 0$ and $c \geq 0$, and let v satisfy the same initial and spatial boundary conditions. Assume sufficient regularity for the energy identity and $u_x, v_x \in L^\infty((0,1)^2)$. Put $r = v_t - \nu v_{xx} + vv_x + cv - f$ and $M_* = \|u_x\|_\infty + \frac{1}{2}\|v_x\|_\infty$. Then*

$$\|v(t) - u(t)\|_{L^2_x}^2 \leq e^{(2M_*+1)t}\int_0^t \|r(s)\|_{L^2_x}^2\,ds, \qquad 0 \leq t \leq 1. \tag{32}$$

Proof. Set $e = v - u$. The convection difference is $vv_x - uu_x = ve_x + eu_x$. Subtract the two PDEs, multiply by e, and integrate in space. The homogeneous error boundary condition removes the boundary terms, giving

$$\frac{1}{2}\frac{d}{dt}\|e\|_2^2 + \nu\|e_x\|_2^2 + c\|e\|_2^2 = \langle r, e\rangle + \frac{1}{2}\int_0^1 v_x e^2\,dx - \int_0^1 u_x e^2\,dx. \tag{33}$$

Cauchy–Schwarz and Young's inequality imply

$$\frac{d}{dt}\|e\|_2^2 \le (2M_* + 1)\|e\|_2^2 + \|r\|_2^2. \tag{34}$$

Since $e(0) = 0$, Grönwall's inequality gives the convolution with $e^{(2M_*+1)(t-s)}$, bounded above by Eq. (32). □

This estimate controls solution error through the full space–time residual and a derivative-dependent stability factor. Narrow layers can make that factor large. A finite weak score is not the full residual norm, and its nonlinear symbol weight may depend on the candidate and evaluation sample. Consequently the finite-set theorem in the main text is not automatically a certification theorem for these nonlinear implementations. The estimate instead specifies what a residual-based certificate would need to control.

What preserving an incumbent guarantees. Let b be the stored wide incumbent and v the challenger. Guarded nonlinear search accepts the challenger on $A = \{S_a(v) \le 0.8S_a(b)\}$, using its independent acceptance sample. For any deterministic post-solve loss L,

$$L(\widehat{v}) = \mathbf{1}_A L(v) + \mathbf{1}_{A^c} L(b), \qquad \{L(\widehat{v}) > L(b)\} \subseteq A. \tag{35}$$

The identity follows directly from which stored candidate is returned. Off gate, the method calls the incumbent action with the same seed. A rejected challenger therefore cannot change the numerical solution, although generating and checking it can increase runtime.

Acceptance does not by itself certify improvement in L^2 error: the acceptance event A uses a residual score, whereas improvement in L^2 error requires comparison with the reference solution. This distinction also prevents exact fallback from being misread as a guarantee on accepted challengers. The extension results record numerical accuracy, coefficient stationarity, and runtime together so that each of these properties can be assessed on its own terms.

Homogeneous Cole–Hopf Burgers. This suite removes manufactured interior forcing. On $(x, t) \in (0, 1)^2$, define the positive heat potential and velocity

$$\Phi(x,t) = 1 + ae^{-\nu\pi^2 t}\cos(\pi x), \qquad u(x,t) = -2\nu\,\frac{\Phi_x(x,t)}{\Phi(x,t)}, \qquad 0 < a < 1. \tag{36}$$

Because $\Phi_t = \nu\Phi_{xx}$ and $\Phi > 0$, the Cole–Hopf map gives $u_t - \nu u_{xx} + uu_x = 0$. The spatial boundary values vanish. The supplied initial condition is $u_0(x) = 2\nu a\pi\sin(\pi x)/(1 + a\cos(\pi x))$. The solution changes shape as it decays, rather than translating as a fixed travelling wave.

The eight challenging parameter pairs span $\nu = 0.020$–0.047 and $a = 0.62$–0.88; the eight smooth controls span $\nu = 0.067$–0.160 and $a = 0.44$–0.72. Pairings follow the fixed task definitions rather than a Cartesian parameter grid. The solver uses the initial data in a lift $\ell(x, t) = u_0(x)$ and fits a homogeneous correction $w = v - \ell$ using Eq. (30). Its residual is

$$r = w_t - \nu w_{xx} + (\ell + w)(\ell_x + w_x) + \ell_t - \nu\ell_{xx}. \tag{37}$$

A nonzero right-hand side for this correction equation is the residual of the prescribed lift; the physical PDE still has zero forcing.

Conditioned TANHWIDE is the strongest method by pooled median error in this suite: 2.43×10^{-3}, compared with 9.68×10^{-3} for the broad wide bank and 9.25×10^{-3} for guarded nonlinear search. It reaches 1% error in 94.8% of runs, versus 52.1% and 54.2%, respectively. Its 95th-percentile error is also below 1%. Holdout halving improves on the broad wide median, but is slower than either one-bank method.

Table 16: Homogeneous Cole–Hopf Burgers: all evaluated policies. Errors and times summarize individual runs; the success fraction uses relative L^2 error $\leq 1\%$.

Method	Median error	q95 error	Time (s)	Success	Failed
TanhWide	0.00968	0.0301	0.0839	0.521	0
Conditioned TanhWide	0.00243	0.00988	0.0753	0.948	0
Uniform bank search	0.0248	0.0537	0.916	0	0
Conditioned bank search	0.0359	0.0688	0.983	0	0
Holdout halving	0.00476	0.0235	0.325	0.75	0
Guarded nonlinear search	0.00925	0.0301	0.182	0.542	0
Row adaptation	0.015	0.0515	0.2	0.354	0
Row–column adaptation	0.0129	0.0606	0.284	0.427	0

These outcomes preserve the main interpretive distinction in a nonlinear problem: useful proposal conditioning need not imply useful searched allocation. All selected solutions from the four policies just discussed meet the stationarity threshold, while none meets the strict residual threshold. Their accuracy differences therefore persist among stationary coefficient fits. The analytic solution serves as an independent evaluation reference; it is not used by candidate selection or as an interior training label.

Curved domains and gate transfer. The geometry suite includes linear advection–diffusion–reaction problems on the disk

$$\Omega_D = \{x : \|x - (1/2, 1/2)\|_2 < 0.48\}, \qquad q_D(x) = 0.48^2 - \|x - (1/2, 1/2)\|_2^2. \tag{38}$$

Features $q_D(x)\tanh(w_j^\top x + b_j)$ satisfy homogeneous Dirichlet data on the circular boundary. Their derivatives use $\nabla q_D = -2(x - (1/2, 1/2))$ and $\Delta q_D = -4$, so the operator can be evaluated without approximating the boundary by a polygon. Square tasks retain the square boundary envelope.

The 36-task matrix combines twelve disk boundary layers, twelve square tasks near the coordinate-gate thresholds, and twelve controls. This separates a change of geometry from a change of activation near a fixed threshold. Four budgets and three seeds give 432 runs per method. The absolute summaries in Table 17 pool these strata; they must not be read as disk-only estimates.

Table 17: Curved and near-threshold linear tasks: all evaluated policies. Errors and times summarize individual runs; the success fraction uses relative L^2 error $\leq 1\%$.

Method	Median error	q95 error	Time (s)	Success	Failed
TanhWide	1.12	4.96	0.0759	0.157	0
Geometry-guarded halving	1.16	3.1	0.121	0.157	0
Conditioned TanhWide	1.37	4.1	0.0769	0.00926	0
Guarded portfolio	1.1	4.73	0.0781	0.157	0
Conditioned bank search	1.29	2.97	0.303	0	0
Row adaptation	1.53	5.78	0.234	0.116	0
Holdout halving	1.17	2.89	0.188	0	0
Row–column adaptation	1.63	4.49	0.165	0	0
Physics-guarded halving	1.07	4.87	0.0795	0.157	0

The geometry-aware gate uses a point-cloud forcing centroid, boundary mass, concentration, and roughness. On a disk, boundary mass is the fraction of forcing magnitude at radius at least 0.36. Roughness is the standard deviation of forcing magnitude divided by its RMS. Outflow location is measured by projecting the centered forcing centroid onto the advection direction, avoiding square-specific coordinate thresholds. The configured gate requires Péclet proxy 40–160, concentration at least 10, outflow projection at least 0.30, boundary mass at least 0.85, and roughness between 0.85 and 1.05.

The physics-guarded policy has pooled median error 1.07, compared with 1.12 for TANHWIDE, 1.10 for the guarded portfolio, and 1.16 for geometry-guarded halving. The common 15.7% fraction of runs below 1% error for these guarded policies and the wide baseline gives a different view from their upper-tail errors. Conditioned one-bank fitting reaches 1% error in only 0.926% of runs. Changing proposal placement or admitting a new geometry therefore does not provide uniform accuracy across the suite.

Boundary-conforming features enforce boundary conditions exactly but may fail to resolve interior advective layers. Likewise, a coordinate-independent gate changes where search is attempted, not the resulting approximation space. Absolute errors therefore complement gate activation and compute measurements.

Nonlinear convergence across policies. Table 18 reports coefficient stationarity and residual convergence for all nonlinear policies. Every nonlinear suite has 96 selected solutions per method. Counts use χ_M at the returned coefficients, rather than completion of the iteration loop. The separate residual threshold remains unmet in every listed group.

Table 18: Optimization diagnostics for every nonlinear policy. Each row contains 96 selected solutions. Stationary counts satisfy $\chi_M \leq 10^{-5}$; residual passes satisfy the separate normalized 10^{-6} threshold.

Suite	Method	Stationary	Median χ_M	Residual passes
Steady	TanhWide	91/96	6.77e-06	0/96
Steady	Conditioned TanhWide	91/96	6.53e-06	0/96
Steady	Uniform bank search	87/96	6.21e-06	0/96
Steady	Conditioned bank search	78/96	7.72e-06	0/96
Steady	Holdout halving	85/96	6.84e-06	0/96
Steady	Guarded nonlinear search	94/96	6.83e-06	0/96
Steady	Row adaptation	83/96	6.73e-06	0/96
Steady	Row–column adaptation	80/96	8.15e-06	0/96
Space–time	TanhWide	95/96	5.49e-06	0/96
Space–time	Conditioned TanhWide	96/96	5.40e-06	0/96
Space–time	Uniform bank search	92/96	4.88e-06	0/96
Space–time	Conditioned bank search	87/96	4.95e-06	0/96
Space–time	Holdout halving	95/96	4.92e-06	0/96
Space–time	Guarded nonlinear search	94/96	5.52e-06	0/96
Space–time	Row adaptation	94/96	5.95e-06	0/96
Space–time	Row–column adaptation	92/96	5.72e-06	0/96
Cole–Hopf	TanhWide	96/96	5.05e-06	0/96
Cole–Hopf	Conditioned TanhWide	96/96	6.33e-06	0/96
Cole–Hopf	Uniform bank search	96/96	3.22e-06	0/96
Cole–Hopf	Conditioned bank search	96/96	1.62e-06	0/96
Cole–Hopf	Holdout halving	96/96	4.99e-06	0/96
Cole–Hopf	Guarded nonlinear search	96/96	5.28e-06	0/96
Cole–Hopf	Row adaptation	96/96	4.50e-06	0/96
Cole–Hopf	Row–column adaptation	96/96	4.35e-06	0/96

Compute accounting. The extension work model uses $N_f = \max(64, 4M)$ fitting points. For d fields, its dense-fit term charges

$$C_{\text{fact}} = \left\lceil \frac{T\,[2(dN_f)(dM)^2 - \frac{2}{3}(dM)^3]}{64} \right\rceil, \tag{39}$$

where $T = 30$ is the iteration cap for the nonlinear suites. Proposal and operator evaluations are added separately. This fixed charge reserves the prescribed nonlinear fit cost even when a coefficient solve stops early; measured online time captures the actual iterations and rejected damping trials. These units differ from the hybrid solver's work model and support within-suite comparisons, not equivalent floating-point operation (FLOP) counts across solvers.

Evaluation protocol. Reference errors use 4,096 test points after online timing stops. All fits from the fixed tasks, budgets in Table 12, and seeds 8981–8983 enter the tables, including nonstationary selections. Residual convergence, coefficient stationarity, solution error, and runtime are reported separately.