SUPPLEMENTARY INFORMATION

FOR *PatternFormer: Learning Multiple Solution Patterns in Reaction–Diffusion Systems*

Contents

S1 Notations

Table S1 summarizes the symbols used throughout this Supplementary Information.

Table S1: Summary of notation.

Symbol	Meaning
p	partial-differential-equation (PDE) parameter (scalar or vector)
$\mathcal{S}(p)$	set of coexisting steady solutions at p
$K(p)$, K_t	number of coexisting solutions at p (true count)
K	fixed number of generated solutions (fixed-budget design, $K{=}24$)
$u_{(k)}$	k-th solution in canonical order
$z_k \in \mathbb{R}^{256}$	latent code of the k-th solution
$\hat{z}_k$	model-predicted latent
$\prec$	canonical ordering relation (monotone field functional)
$\mathbf{v}_{\mathrm{s}}, \mathbf{v}, \mathbf{v}_{\mathrm{p}}$	structural markers (block-start, solution, stop; learnable embeddings)
$\lambda_{\mathrm{mse}}, \lambda_{\mathrm{lat}}, \lambda_{\mathrm{pde}}, \lambda_{\mathrm{ce}}$	loss weights
ϵ	physics-residual hinge floor ($\mathrm{ReLU}(R-\epsilon)$)
$\sigma,\ \xi_k$	latent-noise scale and Gaussian noise vector ($\xi_k \sim \mathcal{N}(0, \mathbf{I})$)
r	low-rank adaptation (LoRA) rank

S2 Model and training details

The trained models described here are referred to collectively as PatternFormer (PF).

Pretrained model and adaptation. At its core is the pretrained Qwen2.5-7B model, adapted with low-rank adaptation (LoRA) of rank $r{=}16$ applied to the four attention projection matrices (query, key, value and output); all base weights are frozen. The base tokenizer and word-embedding table are not used: parameters and solutions enter as continuous vectors through the input projector g_{in}, and the structural markers as rows of a small learnable table.

Projectors and heads. The input projector g_{in} is a two-layer multilayer perceptron (MLP) with Gaussian-error-linear-unit (GELU) activations mapping $[z; p] \in \mathbb{R}^{256+d_p}$ to the model's hidden size, where d_p is the parameter dimension. The output projector g_{out} is a two-layer GELU MLP from the hidden size to $\mathbb{R}^{256}$. The dual head consists of a classification head (marker-type logits) and a regression head ($\to \mathbb{R}^{256}$); in the fixed-budget design the classification head is disabled.

Autoencoder. Each solution field is compressed to a 256-dimensional latent by a frozen autoencoder, pretrained per problem with a reconstruction objective and held fixed thereafter (1D UNet for the 1D problems, MLP for the 2D finite-element problem, convolutional for the Gray–Scott fields).

Autoencoder reconstruction floor. Because generation happens in this frozen latent space, the autoencoder's reconstruction error sets a floor on the *precision* of any generated field: no decoder can reconstruct a field more accurately than $\mathcal{D}(\mathcal{E}(u))$. Table S2 reports this floor, measured as the relative L^2 distance $\|\mathcal{D}(\mathcal{E}(u)) - u\|_2/\|u\|_2$ over the held-out split. The floors (2×10^{-3} to 2×10^{-2}) are of the same order as the direct-generation errors reported in the main text, so the moderate precision of the raw generations is largely set by this fixed autoencoder rather than by the sequence model. The autoencoder cannot itself discover or generate coexisting solutions; the sequence model's role is to place the correct set of coexisting states in latent space, while the subsequent Newton refinement closes the residual precision gap. Direct-generation precision is therefore a floor-limited quantity, not a measure of what the sequence model contributes.

S3 Hyperparameter settings

Common settings. All models use Qwen2.5-7B (hidden size 3584, bf16) with LoRA rank $r{=}16$, $\alpha{=}32$, dropout 0.05 on the four attention projections; the latent dimension is 256. The fixed-budget design uses a fixed generation budget $K{=}24$, no stop marker, and a disabled classification head ($\lambda_{\mathrm{ce}}{=}0$). The physics residual uses a finite-difference discretization consistent with the data generator; in the fixed-budget design it is hinged as $\mathrm{ReLU}(R{-}\epsilon)$, where R is the

Table S2: Frozen-autoencoder reconstruction error (relative L^2 of $\mathcal{D}(\mathcal{E}(u))$ versus u), per problem, on the held-out test split (validation split for Gray–Scott). Latent dimension 256 throughout.

Problem	**Autoencoder**	**Reconstruction rel. L^2**
1D single-parameter (Ex. 1)	1D UNet	1.0×10^{-2}
1D two-parameter (Ex. 2)	1D UNet	2.2×10^{-2}
2D single-parameter (Ex. 3)	MLP	2.0×10^{-3}
Gray–Scott, small D (main text)	convolutional	8.4×10^{-3}
Gray–Scott, large D (ablation)	convolutional	1.9×10^{-3}

Table S3: Per-problem dataset and discretization, from the released configuration and data-generation files. For the known-multiplicity problems the multiplicity is known a priori; for Gray–Scott the coexisting set is effectively unbounded and we report the mean canonicalized (D_4-reduced) training count.

Setting	**PDE**	**Parameter**	**Discret.**	**Mult. k**	**# params (tr/val/te)**
1D single	$-u'' + u^2(u^2 - p) = 0$	$p \in [0, 18]$	1024-pt grid	$\{1, 3, 5, 7\}$	18,001 (12.6k/2.7k/2.7k)
1D two	$-u'' + a_4 u^4 + a_2 u^2 = 0$	$a_4 \in [0.30, 0.80]$, $a_2 \in [-14.9, -2.0]$	1024-pt grid	$\{1, 3, 5\}$	26,381 (21.1k/2.6k/2.6k)
2D single	$-\Delta u - u^2 = -s \sin \pi x \sin \pi y$	$s \in [-138, 1600]$	145-node finite-element (ℓ=3)	$\{2, 4\}$	1,739 (1217/260/262)
GS (large D)	Gray–Scott, D_A=2.5×10^{-4}, D_S=5×10^{-4}	$\rho \in [0.037, 0.110]$, $\mu \in [0.052, 0.072]$	128×128	effectively unbounded (mean 9.5)	1,083 (758/162/163)
GS (small D)	Gray–Scott, D_A=6.25×10^{-5}, D_S=1.25×10^{-4}	(ρ, μ)	128×128	effectively unbounded (mean 9.3)	1,930 (1350/288/292)

Table S4: Training hyperparameters per problem, from the released configuration files. "Phases": AE = frozen-autoencoder pretraining; the Qwen stages follow. lr is given as (projector / LoRA). Effective batch = batch × grad-accumulation × GPUs. All Qwen stages use AdamW (weight decay 0.01, gradient clip 1.0, bf16) with ReduceLROnPlateau.

Setting	**Phases**	**Ep.**	**lr (proj/LoRA)**	**Batch (eff.)**	**Design**	**Loss weights**
1D single	AE + 2-stage (ctx → no-ctx)	100	10^{-4}/2×10^{-5} (then 2×10^{-5}/5×10^{-6})	8×2×8=128	stop-token	mse 1.0, ce 1.0, pde 0.05
1D two	AE + 2-stage (warm-start from 1D single)	60	10^{-4}/2×10^{-5} (then 2×10^{-5}/5×10^{-6})	4×2×8=64	stop-token	mse 1.0, ce 1.0, pde 0.05
2D single	AE + 2-stage (ctx → no-ctx)	1000	10^{-4}/2×10^{-5}	8	stop-token	mse 0.5, ce 1.0, pde 0.05
GS (large D)	AE (600 ep) + 2-stage (ctx → no-ctx), from stop-token ckpt	60	3×10^{-5}/10^{-5}	8×4×4=128	fixed-K	mse 0.5, lat 0.5, ce 0, pde 0.02, tail 0.01
GS (small D)	AE + 2-stage (ctx → no-ctx)	60	10^{-4}/2×10^{-5}	8	fixed-K	mse 0.5, lat 0.5, ce 0, pde 0.02, tail 0.01

field's mean-squared FDM residual and ϵ=10^{-3} is a floor, to prevent the trivial solution from dominating. Scheduled sampling (probability 0.25) is used in the fixed-budget design to mitigate exposure bias. The stop-token cross-entropy is class-weighted (stop marker 3.0, solution 1.0). Training also draws each parameter with probability proportional to $K_t^{1.5}$ (multiplicity-weighted sampling), so rare high-multiplicity parameters are not neglected, and uses distributed data parallelism with manual gradient averaging.

S4 Ground-truth data generation (classical solvers)

All ground-truth solution sets were produced by classical multiple-solution solvers developed by Hao and co-workers; the learned model never invokes a classical solver except for optional post-processing. We summarize each pipeline below.

1D single-parameter: companion-based multilevel FEM with continuation[1]. The equation $-u''+u^2(u^2-p) = 0$ on $(0, 1)$ with $u'(0) = 0$, $u(1) = 0$ is discretized by a second-order finite-difference (FDM) operator on a uniform 1024-point grid. Multiple solutions are seeded by the companion-based multilevel finite-element method (CBMFEM):

node-wise, the quartic nonlinearity reduces to a polynomial whose roots are obtained from the eigenvalues of a 4×4 companion matrix, yielding several coarse initial guesses that are refined by Newton's method across mesh levels. The full dataset is then traced by *parameter continuation*: starting from the solutions at $p = 18$, the parameter is decreased to $p = 0$ in steps of $\Delta p = 10^{-3}$, each branch warm-started from the previous step and refined by damped Newton to $\|F\| < 10^{-9}$; a branch is terminated at a fold when the Newton update jumps (infinity-norm change > 0.2). This yields the parameter-dependent count $(1, 3, 5, 7)$.

1D two-parameter: homotopy continuation with solution maps[2]. The equation $-u'' + a_4 u^4 + a_2 u^2 = 0$ on $(0, 1)$ with $u'(0) = 0$, $u(1) = 0$ is discretized by the same 1024-point FDM. The coexisting solutions over the (a_4, a_2) region are computed by homotopy continuation with mesh refinement, which produces one-parameter bifurcation diagrams and two-parameter solution maps that partition the parameter plane by the number of steady states. Every stored solution is Newton-refined to $\|F\| < 10^{-9}$ so that it is an exact root of the same FDM operator later used for post-processing.

2D single-parameter: companion seeding, marching, and symmetry reduction[1]. The equation $-\Delta u - u^2 = -s \sin(\pi x) \sin(\pi y)$ on $[0, 1]^2$ with $u = 0$ on $\partial\Omega$ is discretized by piecewise-linear finite elements on a triangular mesh (ℓ=3: 145 nodes, 256 elements, 113 free nodes). At an anchor value $s = 1600$, a seed search combined with companion-matrix root finding, augmented by the dihedral (D_4) symmetry of the square domain, produces the full set of solutions (10 before symmetry reduction). The parameter s is then marched downward, each solution warm-started from the previous s and refined by Newton's method to $\|F\|_\infty < 10^{-9}$. Solutions are reduced to D_4-orbit representatives, giving 2 or 4 coexisting solutions per s.

Gray–Scott: quasi-Newton tensor-product solver with continuation[3]. The steady Gray–Scott system is discretized by FDM on a 128×128 grid with Neumann boundary conditions. Each nonlinear solve uses an efficient graphics-processing-unit (GPU) quasi-Newton method that approximates the Jacobian by a linear Laplacian plus simplified reaction terms with a tensor-product (Kronecker) structure[3]. Multiple coexisting patterns at each (ρ, μ) are obtained by parameter continuation (a breadth-first sweep over the (ρ, μ) grid, warm-starting each cell from solutions found at inner neighbours) together with random polynomial perturbations of the warm starts to probe different basins of attraction. Homogeneous (trivial) states are dropped (field range below 0.05), and each parameter's solution set is reduced to D_4-orbit representatives. Across the held-out test parameters this continuation procedure yields on average about 9.7 distinct D_4-reduced coexisting patterns per parameter, and up to 25 at a single parameter. Crucially, continuation attains this richness by warm-starting each parameter from the solutions of its grid neighbours, i.e. by sweeping the entire (ρ, μ) plane; this neighbour information is unavailable when a single target parameter is queried in isolation, as it is for the trained model. PatternFormer and the random-start solver both operate from the target parameter alone, and on this equal-information footing PatternFormer (2.90 deterministic, 8.39 with noise) far exceeds the random-start solver (0.91); continuation is therefore used only as a rich reference set, not as a same-setting baseline. That reference itself remains a lower bound rather than the exhaustive coexisting set, which for Gray–Scott is effectively unbounded.

S5 Data splitting and weighted training

Parameter-level splits. All datasets are split at the *parameter* level. The entire solution set of a given parameter is assigned to a single split, so no parameter leaks across train/validation/test. The 1D single-parameter set uses 12,600/2,700/2,701 parameters (seed 42); the 1D two-parameter set 26,381 parameters (21,103/2,637/2,641, seed 42); the 2D set 1,739 parameters (1,217/260/262, seed 42); the Gray–Scott set 758/162/163 parameters at the larger diffusion coefficients (seed 42) and 1,350/288/292 parameters (total 1,930, seed 42) at the four-fold-smaller coefficients. Solutions are canonicalized (and, for Gray–Scott, D_4-reduced) before splitting.

Multiplicity-weighted sampling. The number of coexisting solutions is highly skewed across parameters (most parameters have few solutions; high-multiplicity parameters are rare). To prevent the model from neglecting the rare high-multiplicity parameters, training uses a multiplicity-weighted epoch sampler that draws each parameter with probability proportional to $K_t^{1.5}$, where K_t is its number of ground-truth solutions, with each target parameter repeated a fixed number of times per epoch.

Class-weighted termination (stop-token design). In the stop-token design, the cross-entropy on the classification head is class-weighted (weight 3.0 for the stop marker versus 1.0 for solution markers) so that the rare termination marker is not overwhelmed by the many solution markers; this is what enables the accurate ($<1\%$ missing) termination on the known-multiplicity problems.

Physics-residual penalty. A physics residual computed by the FDM operator of the data generator is added to the loss to assist training. It is hinged, $\mathrm{ReLU}(R-\epsilon)$ (with R the field's mean-squared FDM residual and floor $\epsilon = 10^{-3}$), with a small weight and a warmup ramp (epochs $5 \to 30$ for the known-multiplicity problems; from epoch 0 in the Gray–Scott fine-tune), so that it improves the physical accuracy of generated fields without letting the trivial solution dominate. The floor marks the residual level at which a generated field already counts as a valid approximate solution. Above it a field is still pushed toward lower residual, while below it the physics loss is flat, so once a field is a valid pattern the gradient switches off and the loss never drives it on into the trivial state's near-zero-residual basin. The floor must therefore lie below the residual of fields far from any steady state, which stay penalized, and well above the near-zero residual of the trivial state itself; 10^{-3} sits comfortably in this range. In the fixed-budget design this residual additionally supervises the tail solutions ($k \geq K_t$), driving the discovery of valid solutions beyond the training set. On these unsupervised tail slots the residual carries a smaller weight (0.01, versus 0.02 on the teacher-forced head slots), so that the exploratory slots are gently guided toward valid patterns without biasing the head or collapsing onto the trivial state.

S6 Post-processing: refinement, matching and counting

Known-multiplicity problems. Each generated field is refined by the same classical solver used for data generation (damped Newton with backtracking line search; minimum step 10^{-4}, contraction 0.5) to $\|F\| < 10^{-9}$ within at most 30 iterations. Predicted solutions are matched one-to-one to the ground-truth set by the Hungarian algorithm minimizing relative L^2 distance; a ground-truth solution is counted as recovered if the matched prediction reaches relative $L^2 < 10^{-2}$ after refinement. Duplicate predictions are removed at relative $L^2 < 10^{-3}$.

Gray–Scott. Each of the $K{=}24$ generated fields is refined by the quasi-Newton FDM solver (lenient settings: up to 8000 iterations, early exit at 3000 iterations / residual 0.05) and accepted if it converges to a non-trivial steady state (activator range above 0.05). Converged solutions are deduplicated up to the D_4 symmetry (relative $L^2 \leq 0.15$), and the number of distinct solutions is reported as the deterministic # solutions. Coverage is the fraction of the reference solution set matched, using the same 0.15 threshold: a converged field is counted as matching a reference solution when their D_4 relative L^2 falls below 0.15, and a converged solution that matches no reference solution for that parameter is counted as beyond-ground-truth.

S7 Evaluation protocol

Known-multiplicity problems. We report relative L^2 error per solution before and after refinement, the number of refinement (Newton) steps, and exact count matching against the complete ground-truth set, binned by the number of coexisting solutions.

Gray–Scott. We report the deterministic # solutions (number of distinct converged solutions per parameter), coverage of the reference set, and the fraction of generated solutions valid but beyond the training set, reported separately for the two diffusion settings. Distinct solutions are counted after Newton refinement (tolerance 10^{-9}) and deduplication up to discrete symmetry (relative L^2 threshold 0.15).

Headline numbers.

- **Known-multiplicity problems, direct vs. refined (median relative L^2):** 1D single-parameter $2.3 \times 10^{-2} \to 2.9 \times 10^{-8}$ (median 3 Newton steps, 99.9% converged, 2,701 test params); 1D two-parameter $2.3 \times 10^{-2} \to 2.6 \times 10^{-8}$ (median 3 steps, 99.8%); 2D $6.1 \times 10^{-3} \to 1.4 \times 10^{-3}$ (median 4 steps, 99.6%).
- **GS, $D_A{=}6.25 \times 10^{-5}$, $D_S{=}1.25 \times 10^{-4}$:** on 292 held-out test parameters and a common refine-and-deduplicate protocol, the deterministic fixed-K model returns a mean of 2.90 # solutions per parameter, versus 1.86 for the stop-token model and 0.91 for a traditional solver from random initial guesses. About 1.8 of the 2.90 are valid patterns beyond the training set, and 59% of parameters (172/292) yield at least one; up to 14 distinct appear at a single parameter. Injecting latent noise and taking the union of a few passes raises the mean to 8.39.
- **GS, $D_A{=}2.5 \times 10^{-4}$, $D_S{=}5 \times 10^{-4}$ (reported separately):** on 163 held-out test parameters, deterministic # solutions 5.18, versus 4.05 (stop-token, fixed-K) and 0.54 (a classical quasi-Newton solver from random initial guesses—distinct from the randomly-initialized-backbone model, which gives 1.44); the stop-token head under its own stopping rule under-generates further, to 2.53; about 1.93 of the 5.18 are valid patterns beyond the training set, 70% of parameters (114/163) yield at least one, and up to 17 distinct appear at a single parameter; with the noise knob, 9.16.

- **Known-multiplicity wall-clock speedup** (model-initialized refinement vs. the classical solver, both returning the same solutions at residual 10^{-9}): 1D single-parameter $\sim$4–9$\times$ at N=1024 rising to $\sim$51–57$\times$ at N=4096 (vs. a multi-scale solver); 1D two-parameter 14–16$\times$ (vs. multi-start Newton search); 2D 89–331$\times$ (vs. seed-search plus continuation).

Fixed-budget slot utilization. Of the K=24 fixed slots, a mean of 5.3 converge to valid non-trivial steady states before deduplication, which reduce to the 2.90 distinct patterns after D_4 reduction. This over-generate-then-reduce behaviour mirrors the classical continuation solver, which likewise returns many symmetry-equivalent duplicates before D_4 reduction. Because most parameters admit only a handful of coexisting patterns, the remaining slots reproduce symmetry-equivalent states or do not converge, rather than indicating unused capacity.

Deduplication threshold. The distinct-solution count is not an artifact of the deduplication threshold: tightening it three-fold, from 0.15 to 0.05, changes the mean distinct count by only $+15\%$, so the discovered patterns are genuinely separated rather than near-duplicates merged by a permissive threshold.

Canonical ordering. The canonical order is induced by the mean activator concentration $\langle A \rangle$, which yields a well-defined sequence whenever coexisting solutions have distinct $\langle A \rangle$. Across the 1,563 multi-solution parameters in the dataset, genuine near-degeneracies are rare: only 0.8% of coexisting solution pairs have $\langle A \rangle$ within 10^{-4} and 5.1% within 10^{-3} (median gap between adjacent solutions 4.6×10^{-3} over an $\langle A \rangle$ range of $\approx$0.38), so the ordering is well-posed for the large majority of solution sets.

Timing protocol and optimization. All timing experiments compare the two methods at the same accuracy (both driven to $\|F\| < 10^{-9}$), since a classical solver can otherwise trade accuracy for speed. Given only the target parameter, each method must return the full solution set: PatternFormer generates candidates in a single GPU forward pass and refines them with Newton's method, whereas the classical baseline—the same multiple-solution solver used to generate the reference data—discovers the branches from scratch by multi-start search or parameter continuation on CPU. The reported speedups are therefore end-to-end pipeline costs. Optimization uses AdamW with distributed training.

S8 Known-multiplicity problems: two-parameter and two-dimensional results

Figures S1 and S2 give the complete results for the 1D two-parameter and 2D single-parameter problems, with the same panel layout as the 1D single-parameter figure in the main text (Fig. 2): example solutions (direct and refined) against ground truth, per-multiplicity error/step/residual statistics, the solution map or bifurcation diagram, and the wall-clock speedup over the corresponding classical solver.

S9 Two-dimensional cases with elevated relative error

A small number of two-dimensional test parameters show a raised relative L^2 in the binned statistics of Fig. S2B. These are artefacts of the error metric, not different or missing solutions, as the following two diagnostics show.

The first is a near-zero-amplitude branch (Fig. S3). At small s one coexisting solution has a tiny field norm ($\|u\|_2 \approx 0.8$), so its relative $L^2 = \|\hat{u} - u\|_2 / \|u\|_2$ is inflated by the small denominator even though the absolute error is small; the other branch, of normal amplitude, is recovered to relative $L^2 \approx 0$.

The second is a set of near-degenerate branches (Fig. S4). Where several coexisting patterns are nearly identical, the one-to-one matcher cross-assigns them and reports an elevated relative L^2, yet the outputs do not merge: the smallest nearest-neighbour distance among the model's solutions (0.15) equals that among the ground-truth branches (0.15), so every branch stays distinct. This is the opposite of mode collapse.

S10 Extrapolation beyond the training parameter range

The main-text extrapolation results (Fig. 4) use the in-distribution evaluation pipeline with the test parameters replaced by values outside the training range. Ground truth at each extrapolated parameter is produced by the classical solver alone, by continuation of the in-distribution branches or a multi-start search, with no learned warm start.

Out of distribution a multi-solution model can fail in two ways that a forced one-to-one match would conflate, so we score by coverage. Every generated field is refined by the classical Newton solver at the extrapolated parameter and

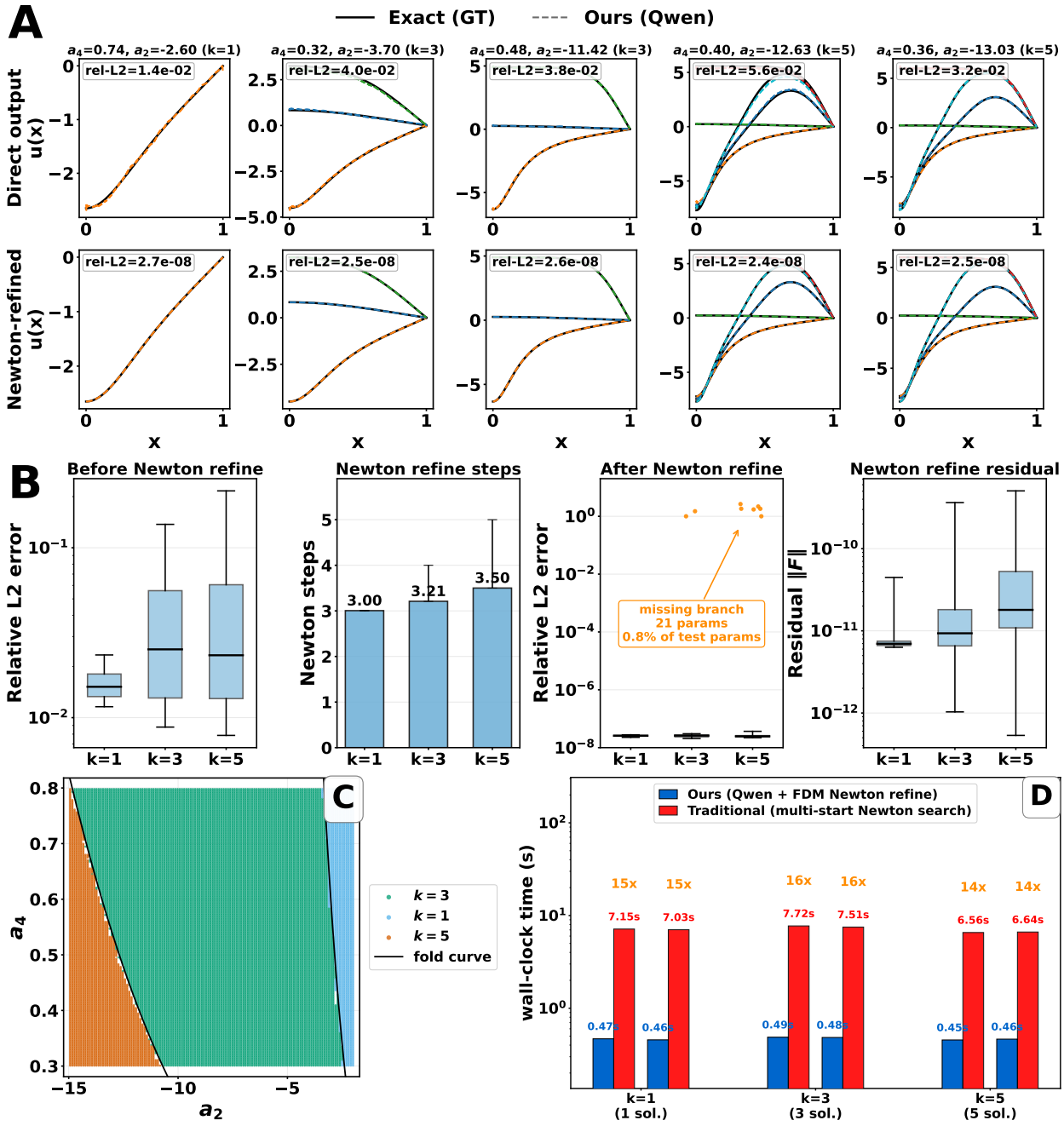

Figure S1: **Complete recovery of all coexisting solutions for the 1D two-parameter problem** $-u'' + a_4 u^4 + a_2 u^2 = 0$. **A**, At five representative (a_4, a_2) (with k coexisting solutions), predictions (dashed) over ground truth (solid); top, direct; bottom, after Newton refinement. **B**, Relative L^2 before and after refinement, Newton steps and residual $\|F\|$, binned by k (the full branch set is recovered for 99.2% of parameters). **C**, The (a_4, a_2) solution map, partitioned by the number of steady states ($k \in \{1, 3, 5\}$), with the fold curve. **D**, Wall-clock time versus a multi-start Newton search. Box plots, median and interquartile range; whiskers, 1.5×IQR.

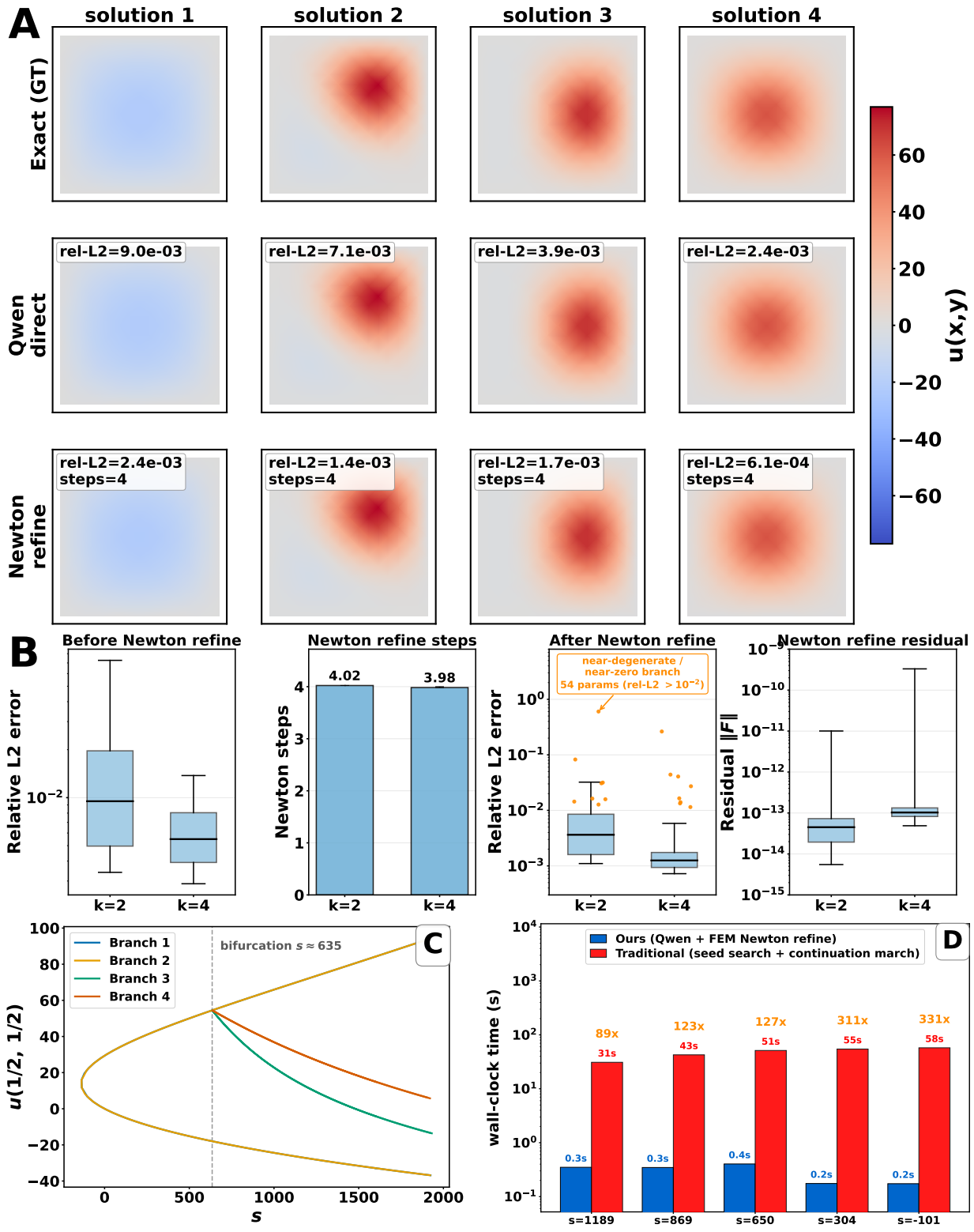

Figure S2: **Complete recovery of all coexisting solutions for the 2D problem $-\Delta u - u^2 = -s\sin(\pi x)\sin(\pi y)$. A**, For a representative s (four coexisting solutions): ground truth (top), direct output (middle), Newton-refined (bottom), shared colour scale. **B**, Relative L^2 before and after refinement, Newton steps and residual, binned by $k \in \{2, 4\}$. **C**, Bifurcation diagram, centre value $u(\frac{1}{2}, \frac{1}{2})$ versus s, with the secondary bifurcation near $s \approx 635$. **D**, Wall-clock time versus a seed-search-plus-continuation solver. Box plots, median and interquartile range; whiskers, $1.5\times$IQR.

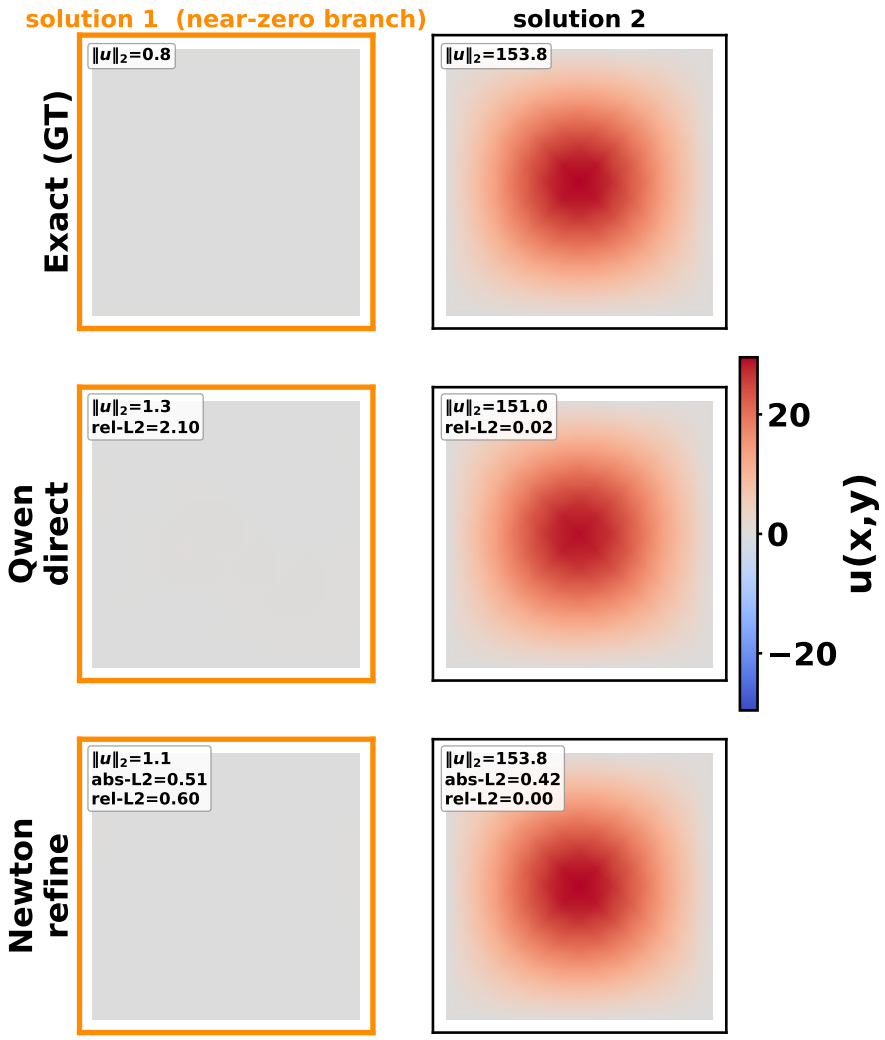

Figure S3: **A high relative L^2 can mark a near-zero-amplitude branch, not a model error.** The two coexisting solutions at s=3: ground truth (top), direct model output (middle) and Newton-refined output (bottom), sharing a colour scale, with the field norm $\|u\|_2$ annotated. Solution 1 (highlighted) has a tiny norm, so its relative L^2 is large despite a small absolute error (abs-L^2=0.51); solution 2, of normal amplitude, is recovered to relative $L^2 \approx 0$.

kept if it converges (residual below the data-generation tolerance); the converged fields are deduplicated to the distinct solutions the model actually produces. *Coverage* is the number of true branches recovered (relative $L^2 < 10^{-2}$, or $< 5\times10^{-2}$ in two dimensions) and *missing* $= n_{\mathrm{gt}} -$coverage; a distinct converged solution that matches no true branch is counted separately as a wrong solution (it refines to something that is not one of the true solutions there). Accuracy (relative L^2, Newton steps, residual) is reported only over the covered branches, so duplicates and non-converged outputs are not charged as unit error.

Wrong solutions are absent: the model does not invent spurious branches, and every branch it covers refines to solver precision (post-refinement relative $L^2 \approx 2.5 \times 10^{-8}$ in one dimension and $\approx 1.6 \times 10^{-3}$ in two dimensions, the latter being the ground truth's own residual-10^{-9} floor). The single outlying two-parameter point ($a_2{=}{-}15.5$) is most plausibly a true solution the reference continuation missed rather than a genuine model false positive. Out-of-distribution degradation is thus a drop in coverage—missing solutions, not wrong ones. Even far out, a raw output with direct relative $L^2 \gg 1$ often still lies in a true branch's Newton basin, so refinement pulls it back exactly.

Figure S5 shows the two-parameter coverage over the (a_2, a_4) plane for direct generation (left) and for single-jump seeding (right), each cell annotated with the recovered/existing branch counts. Direct generation is complete throughout the training rectangle and over a broad region beyond it, and falls where the quartic well deepens ($a_2 \lesssim -20$) or the branch count rises as $a_4 \to 0$. A single Newton jump from the nearest in-training output already recovers much of the missing coverage, and the stepped march of the main text (main-text Fig. 4C) closes the remainder, reaching full coverage at every extrapolated parameter.

The stepped march used for main-text Fig. 4C is the reference solver's own continuation, seeded from the model rather than from the dataset. We refine the model's raw output at the in-training cells to obtain a starting solution set, then propagate it outward across the (a_2, a_4) grid exactly as the ground-truth continuation does: each unsolved cell is resolved by Newton's method from the converged solution sets of its already-solved grid neighbours, sweeping outward until the grid is filled. Because the continuation advances into each cell from every solved direction rather than along a single straight line, it carries forward the branches that a one-directional march approaches from the wrong side and loses—in particular the fold-born branches at the $a_4 \to 0$ edge, where new solutions appear as the parameter crosses a saddle-node fold. Seeded this way from the model's boundary predictions, the continuation reproduces the full ground-truth branch set at every cell of the plane (coverage 1.0 at all 143 cells): the model supplies the in-distribution solution set and the reference continuation only carries it outward to the extrapolated parameter.

Under direct generation Example 1 collapses to a single branch by $p{=}20$ (main-text Fig. 4A), yet the full solution set stays reachable far beyond. Taking the model's raw solution set at the training boundary $p{=}18$ as seeds and marching them outward by stepped continuation ($\Delta p{=}1$, each step refined by damped Newton to $\|F\| < 10^{-9}$) recovers all seven coexisting branches at $p{=}50$ and again at $p{=}100$, more than five times the upper training value, in both cases matching the seven branches obtained by continuing the true $p{=}18$ set (Fig. S6). The branches persist out there. What degrades far out of distribution is the direct sampling, not the solution structure.

S11 Gray–Scott beyond the training range: direct generation versus seeding from the boundary

The main text reports that direct generation still returns non-trivial Gray–Scott patterns at a few parameters just outside the training band (main-text Fig. 3A). It is natural to ask the question raised by the known-multiplicity extrapolation experiments: does the alternative route there—seeding a classical solve from the model's solution set at the training boundary, rather than generating directly at the new parameter—also help on Gray–Scott. Because the Gray–Scott coexisting set has no exhaustive ground truth, we do not score coverage of a complete set; instead we count the distinct non-trivial steady states each route produces, and gauge how many patterns are present at all by continuing the true in-band solution set to the target parameter with the quasi-Newton solver.

We probe outward from twelve boundary parameters, along the diagonal that the pattern band follows in the (ρ, μ) plane, at three push distances each (36 out-of-distribution targets; larger diffusion coefficients). Two facts shape the outcome. First, the Gray–Scott pattern region is sharply bounded. At 24 of the 36 targets (67%), continuation of the true in-band set yields nothing, so no non-trivial steady state exists there for any method to recover. Second, at the twelve targets where patterns do persist (Table S5), the three routes are comparable and region-dependent rather than ordered. Averaged over those twelve, direct generation returns 5.7 distinct patterns, stepped continuation from the boundary 5.3, and a single Newton step from the boundary 4.3; direct generation is the best of the three at six targets, a seeded route at five, with one tie.

Two mechanisms explain why seeding offers no consistent advantage here, in contrast to one dimension. Unlike the one-dimensional problem, where direct generation collapses sharply out of distribution (main-text Fig. 4A), Gray–

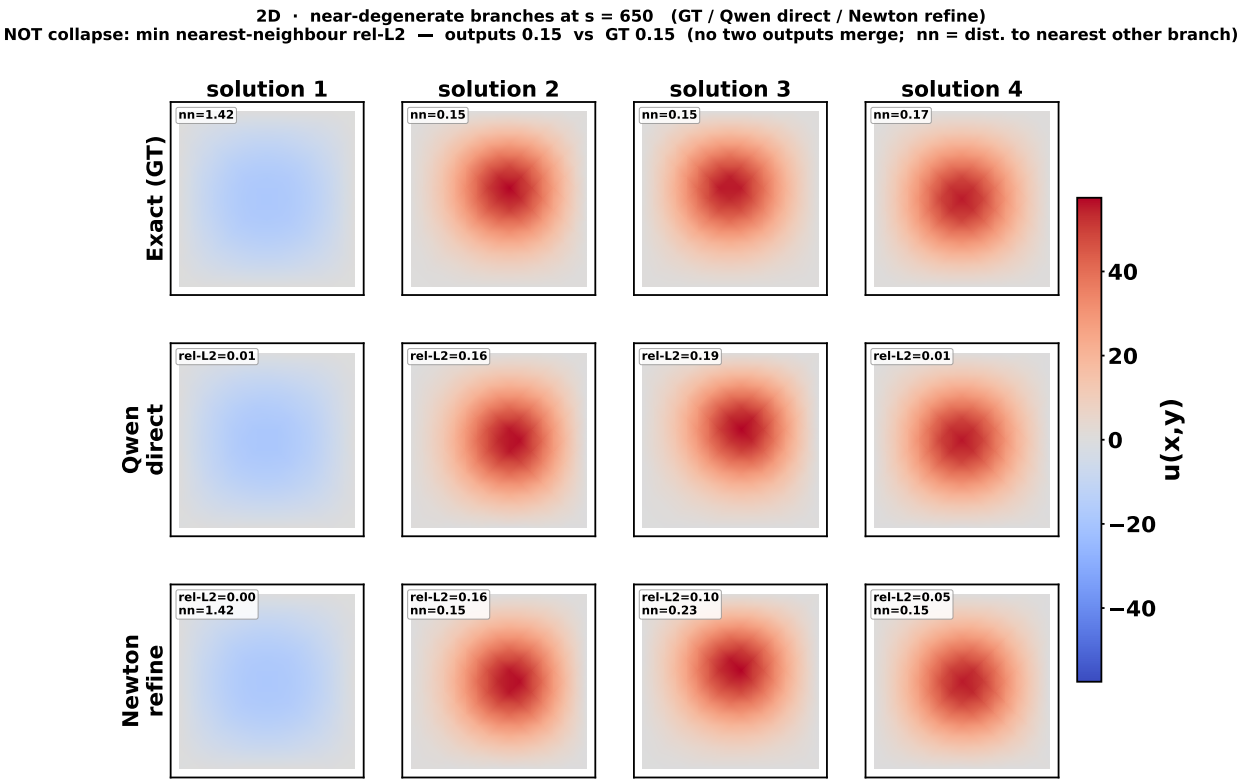

Figure S4: **Elevated relative L^2 among near-degenerate branches is not mode collapse.** The four coexisting solutions at s=650: ground truth (top), direct output (middle) and Newton-refined output (bottom). Several patterns are nearly identical, so the one-to-one matcher cross-assigns them and reports a raised relative L^2. The minimum nearest-neighbour distance (nn) among the model's outputs equals that among the ground-truth branches (0.15 in both), so no two outputs merge: all four remain distinct.

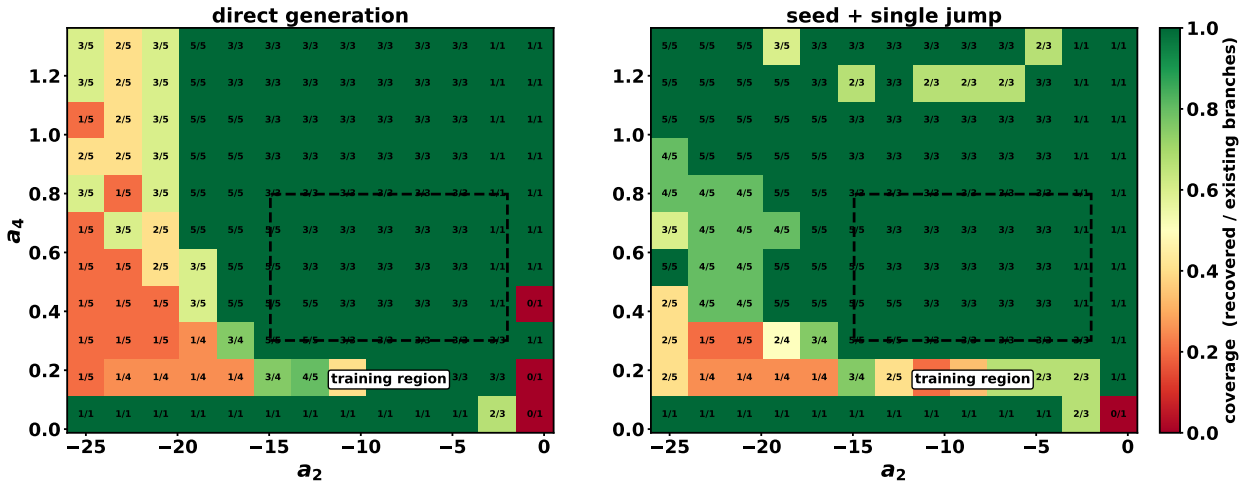

Figure S5: **Two-parameter extrapolation coverage: direct generation versus single-jump seeding.** Coverage (recovered/existing branches) over the (a_2, a_4) plane, each cell labelled with its recovered/existing branch counts. **Left**, direct generation at the target. **Right**, a single Newton jump at the target from the model's raw output at the nearest in-training parameter. The dashed box is the training region (ground truth: training data inside, classical continuation outside). The stepped march, which reaches full coverage across the whole plane, is in the main-text Fig. 4C.

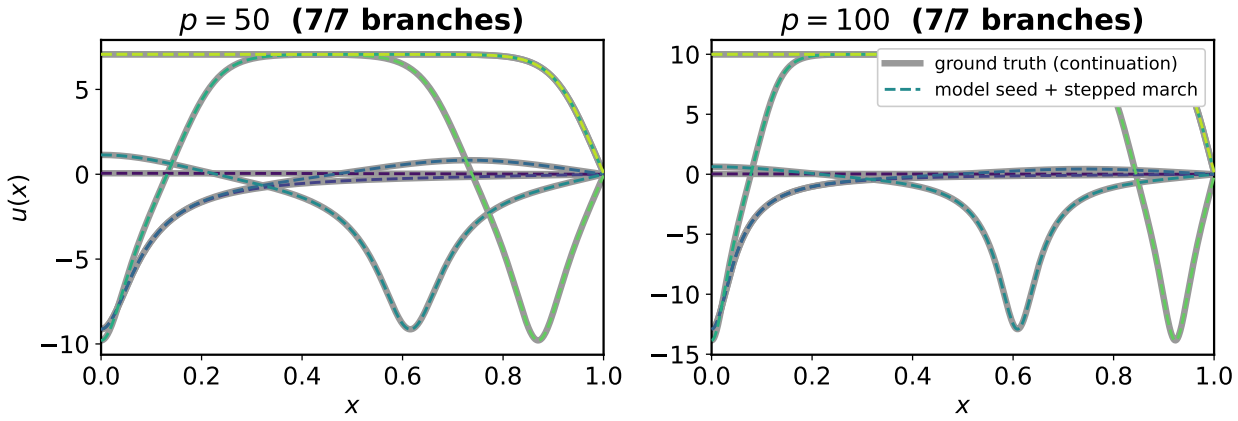

Figure S6: **Far extrapolation of Example 1 by the model-seeded stepped march.** Starting from the model's solution set at the training boundary p=18 and marching outward in steps of Δp=1 (damped Newton at each step), all seven coexisting branches are recovered at p=50 (left) and p=100 (right), more than five times the upper training value. Grey, the branches obtained by continuing the true p=18 set (ground truth); dashed colour, the model seed plus stepped march.

Scott direct generation does not collapse. The model already extrapolates enough to recover most surviving patterns on its own, leaving little for a seeded solve to add. And the single-Newton-step route that suffices in one dimension is unreliable here—Gray–Scott's tight Newton basins make a single jump across the parameter gap diverge, returning nothing at several targets—so only stepped continuation, which is exactly the procedure that generated the reference atlas, occasionally matches or exceeds direct generation. We therefore report direct generation as the out-of-distribution route for Gray–Scott in the main text, and note the boundary-seeding route here for completeness.

Table S5: Gray–Scott out-of-distribution targets where patterns persist (larger diffusion coefficients; the 24 further targets with no non-trivial steady state are omitted). For each target (ρ_*, μ_*) just outside the training band, "present" is the number of true in-band patterns that survive continuation to that parameter; the last three columns give the number of distinct non-trivial steady states recovered by direct generation, by a single Newton step from the model's boundary solutions, and by stepped continuation from them. A count can exceed "present" when a route also lands on valid patterns beyond the continued in-band set. All counts are after refinement and D_4 deduplication.

ρ_*	μ_*	present	direct	single step	continuation
0.067	0.070	5	6	3	3
0.070	0.069	5	6	0	3
0.073	0.069	5	7	0	3
0.075	0.068	7	6	6	7
0.078	0.067	6	3	0	6
0.099	0.060	3	5	4	4
0.113	0.051	9	3	6	6
0.113	0.053	9	8	7	7
0.116	0.050	7	8	6	5
0.116	0.052	9	5	7	7
0.119	0.050	6	6	6	5
0.119	0.052	9	5	7	7
mean		6.7	5.7	4.3	5.3

S12 Small-D atlas extrapolation (main-text Fig. 3A)

For the small-D regime shown in the main text, direct generation was queried at 144 out-of-distribution cells of the (ρ, μ) atlas lying beyond the training rectangle (the orange-framed cells of main-text Fig. 3A). Nine of these cells yield non-trivial coexisting patterns, 43 patterns in total and 1–9 per cell (Table S6); the remaining 135 cells lie in parameter regimes where continuation of the in-band solution set also finds no steady pattern, so their emptiness is a property of

the system rather than a failure of the model. Consistent with the one-dimensional extrapolation experiments, direct generation outside the training band produces valid steady states rather than spurious ones.

Table S6: Small-D Gray–Scott (D_A=6.25 $\times$ 10^{-5}, D_S=1.25 $\times$ 10^{-4}) out-of-distribution cells (main-text Fig. 3A, orange) that yield non-trivial patterns under direct generation. The other 135 of the 144 tested cells contain no steady pattern. Counts are after refinement and D_4 deduplication.

ρ_*	μ_*	# distinct patterns
0.121	0.049	9
0.113	0.051	8
0.106	0.055	6
0.113	0.053	5
0.121	0.051	5
0.106	0.056	4
0.113	0.055	3
0.106	0.053	2
0.050	0.071	1
total (9 cells)		43

S13 Gray–Scott at the four-fold-larger diffusion coefficients (D_A=2.5$\times$$10^{-4}$, D_S=5$\times$$10^{-4}$)

At the four-fold-larger diffusion coefficients the domain is effectively halved, producing coarser, sparser spots, stripes and labyrinths. Figure S7 repeats the main-text analysis (smaller diffusion coefficients) at the larger coefficients, on their own held-out test parameters; the two settings are two parameter regimes reported side by side rather than a difficulty comparison. The method ranks the same way against the baselines—traditional random-initial-guess solving (0.54 # solutions per parameter), the stop-token design (4.05), our deterministic fixed-K generator (5.18)—and the noise knob lifts the # solutions further to 9.16. The two settings are reported separately and never pooled.

S14 Generation beyond the training data at training parameters

The physics residual added to the open-ended Gray–Scott loss is meant to push generation past the solutions present in the training data. A direct test queries the model at parameters it was trained on, where the data already fixes a known list of coexisting solutions, and asks whether the model returns anything beyond that list. At each such parameter we pool the deterministic pass with a few noise-augmented passes, refine every output to an FDM residual below 10^{-9}, deduplicate up to the D_4 symmetry of the square, and label each distinct steady state as either reproducing a training-set solution or new beyond the data. Across the eight highest-multiplicity training parameters the model consistently returns many states absent from the training list. At the two parameters of Fig. S8, for instance, it returns 19 and 18 new states against the 5 that reproduce a training solution at each. The model therefore does more than reproduce the parameter-to-solution map in its data; the residual lets it find additional valid steady states even where the data is held fixed.

S15 Cross-PDE transfer

Each harder known-multiplicity problem is warm-started from the best checkpoint of a simpler one, carrying over the LoRA, projector and dual-head weights and re-initializing only the input projector when the parameter dimension changes. The two-parameter model is initialized from the single-parameter model, and the two-dimensional model from the two-parameter one. Because the governing equations differ across these settings, this warm-starting spans different equations. A controlled comparison against training the same model from scratch, holding architecture, data and budget fixed, isolates the effect of this cross-equation transfer for the hardest step, from the two-parameter one-dimensional problem to the two-dimensional one (main-text Fig. 6 E–H): warm-starting lowers the validation loss throughout training and roughly halves the test direct relative L^2 (median 6×10^{-3} versus 1.1×10^{-2}). This direct prediction is the appropriate point of comparison, because the downstream Newton refinement drives every converged initialization to the same final precision; the transfer benefit is thus properly measured on the raw generation, before refinement equalizes the endpoint. The Gray–Scott model is a deliberate exception: it replaces the generation design (stop-token to fixed-budget), so rather than crossing equations it is warm-started from its own stop-token Gray–Scott checkpoint.

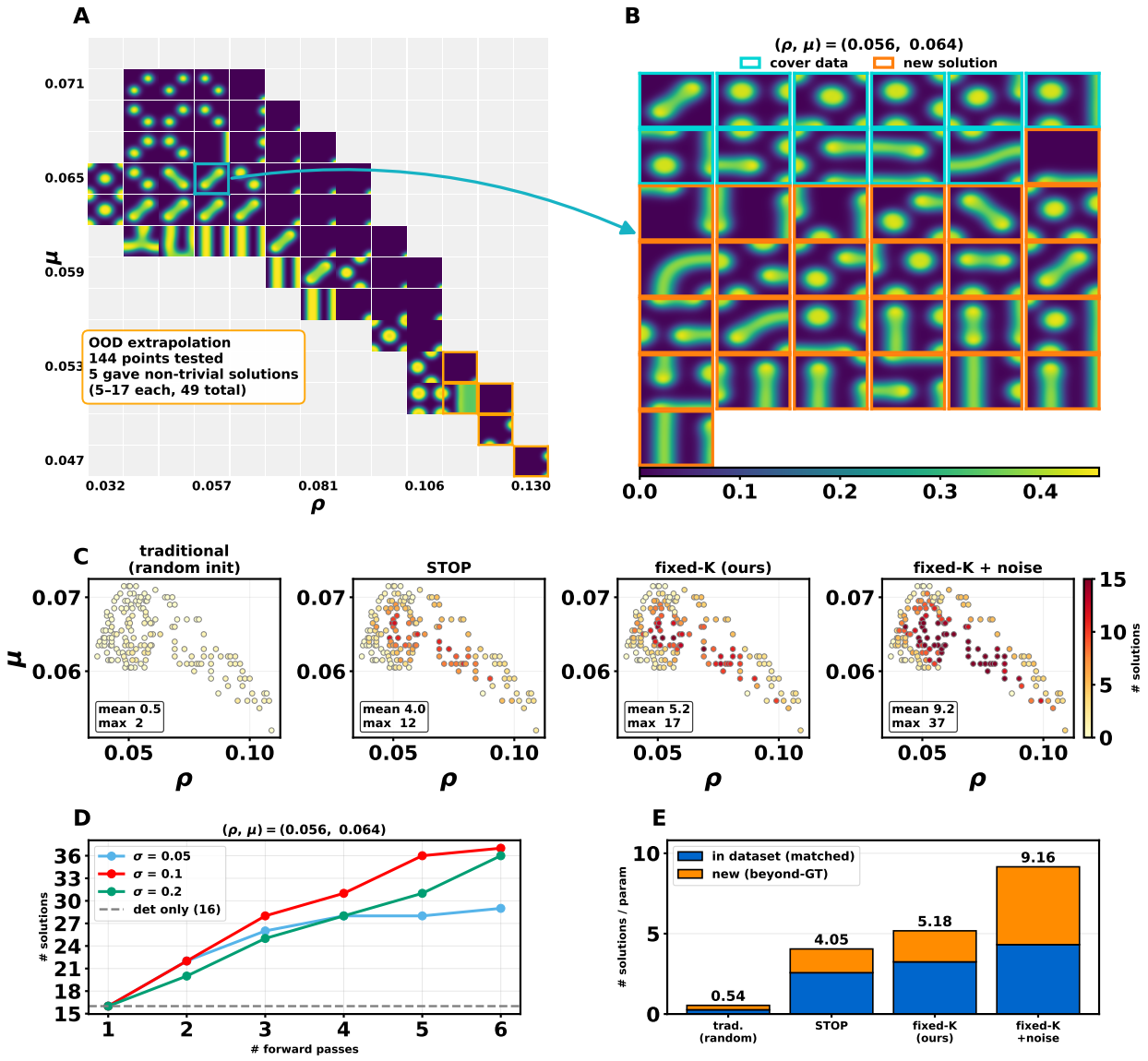

Figure S7: **Gray–Scott at D_A=2.5×10^{-4}, D_S=5×10^{-4} (D_A, D_S four times those of the main-text figure).** Same layout as the main-text figure (smaller diffusion coefficients), on these coefficients' held-out test parameters. **A**, Atlas over the (ρ, μ) plane with out-of-distribution cells marked. **B**, Distinct coexisting patterns the model returns at a representative parameter, pooled over the deterministic pass and a few noise-augmented passes (cyan = covers a training-set solution, orange = new). **C**, Per-parameter # solutions for the four methods. **D**, Noise sweep: cumulative # solutions versus number of forward passes for several σ. **E**, Mean # solutions per parameter, split into ground-truth-matched and new beyond-ground-truth.

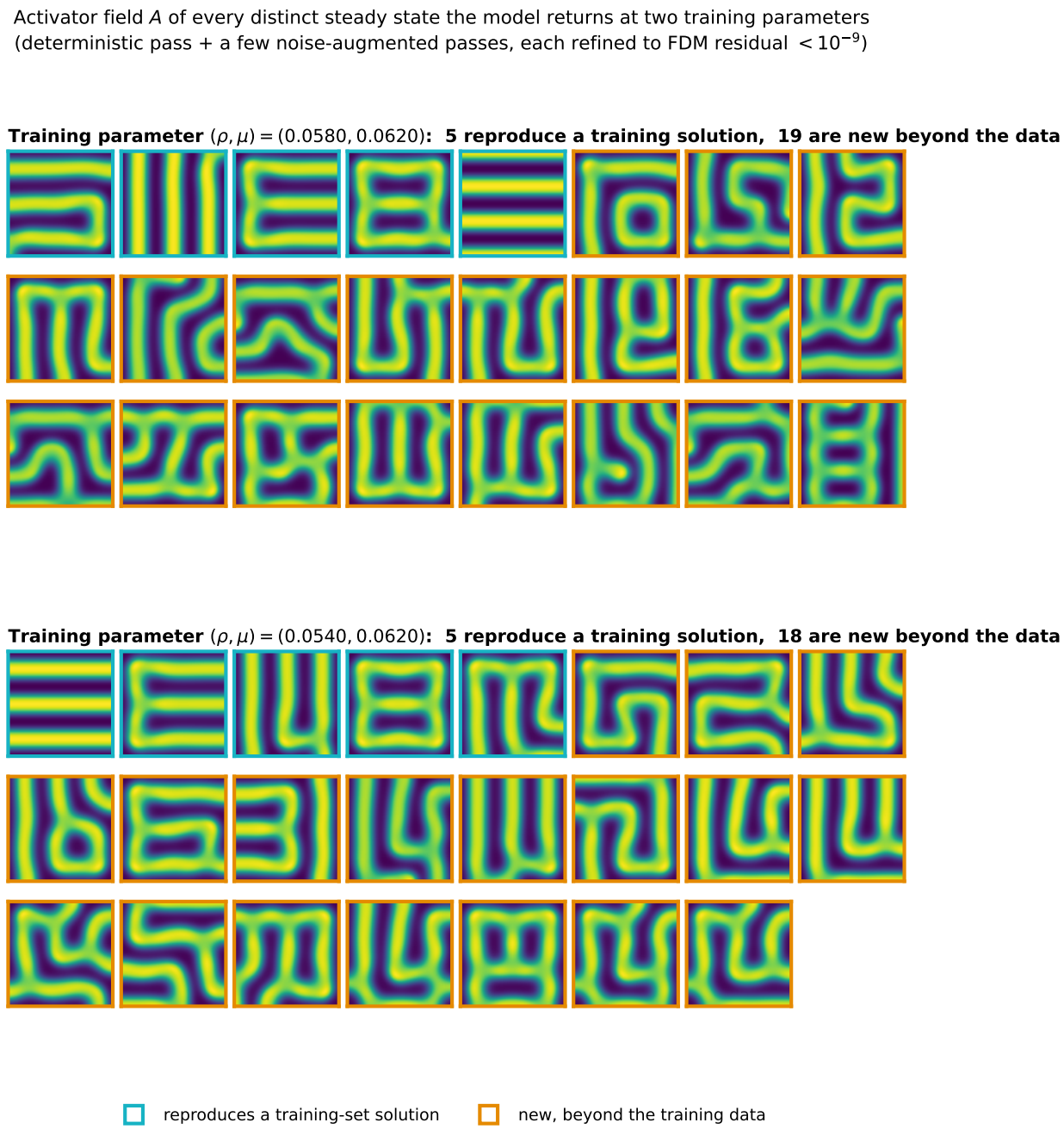

Figure S8: **Generation beyond the training data at training parameters (smaller diffusion coefficients).** For two training parameters, every distinct steady state the model returns, pooled over a deterministic pass and a few noise-augmented passes and refined to an FDM residual below 10^{-9} (activator field A shown; solutions deduplicated up to the D_4 symmetry of the square). Cyan border, the output reproduces a solution present in the training data; orange border, the output is a new coexisting state absent from the training data. At both parameters most returned states lie beyond the training list, showing that the physics residual drives the model past the data even where the parameter was seen in training.

Initialization ablation (numbers behind main-text Fig. 6I). The initialization ablation is a separate experiment from the method comparison above: it fixes the generation design (deterministic fixed-K) and varies only the starting weights, whereas the method comparison fixes the initialization and varies the generation strategy (traditional solver, stop-token, fixed-K, noise). Table S7 lists the deterministic Gray–Scott recovery under the three initializations of main-text Fig. 6I. For the elliptic transfer, warm-starting Example 3 from the Example 2 checkpoint lowers the median direct relative L^2 from 1.1×10^{-2} (a fresh adapter trained from the pretrained model) to 6×10^{-3}, while general LLM pretraining already accelerates convergence on Example 1 relative to a randomly initialized backbone.

Table S7: Initialization ablation on Gray–Scott (D_A=2.5×10^{-4}, D_S=5×10^{-4}), deterministic fixed-K generation on held-out test parameters, corresponding to main-text Fig. 6I. All three arms share architecture, data, optimizer and compute budget and differ only in the starting weights.

Initialization	**# solutions / param**	**Reference-set coverage**
Random-weight (LoRA only)	1.44	0.07
Pretrained Qwen2.5-7B	3.23	0.22
Warm-started PatternFormer	5.18	0.34

Robustness across random seeds. Repeating this deterministic ablation with a second random seed preserves the ordering. The randomly-initialized backbone recovers 1.30 distinct solutions per parameter (versus 1.44 for the first seed) and the pretrained backbone recovers 4.48 (versus 3.23), so the pretraining advantage widens rather than shrinks under the second seed; reference-set coverage follows the same pattern (0.07 and 0.31, versus 0.07 and 0.22). The two arms share architecture, data, optimizer and compute budget across seeds and differ only in the random initialization.

References

[1] Wenrui Hao, Sun Lee, and Young Ju Lee. Companion-based multi-level finite element method for computing multiple solutions of nonlinear differential equations. *Computers and Mathematics with Applications*, 168:162–173, 2024.

[2] Wenrui Hao and Chuan Xue. Spatial pattern formation in reaction–diffusion models: a computational approach. *Journal of Mathematical Biology*, 80(2):521–543, 2020.

[3] Wenrui Hao, Sun Lee, and Xiangxiong Zhang. An efficient quasi-Newton method with tensor product implementation for solving quasi-linear elliptic equations and systems. *Journal of Scientific Computing*, 103:89, 2025.