RESIDUAL GEOMETRY SHAPES DOWNSTREAM CORRECTION WORK FOR LEARNED WARM STARTS IN NONSMOOTH PDE SOLVERS

Anonymous authors
Paper under double-blind review

ABSTRACT

A learned warm start hands a solver a proposal that needs correction. We study how the residual metric used in training changes that downstream correction work in nonsmooth PDEs. In two-phase enthalpy diffusion, a network proposes a window of implicit steps, a linear decoder enforces balance and semismooth Newton corrects them; only the residual weighting changes. Branch structure selects the candidate metric: for phase-change diffusion the inverse stiffness K^{-1} is, up to scale, the only quadratic metric under which no fixed-branch step amplifies differences, whereas a reaction term admits the identity but not K^{-1}. K^{-1} weighting cuts Newton outer iterations by 37% at 4,000 updates, and the advantage persists at the longest tested budget. All four prospective placement tests where both metrics accept every request favor the predicted metric, and a preregistered obstacle test favors Euclidean weighting. Residual size cannot stand in for correction work: starts with smaller residuals in every vector norm can require more exact Newton steps, and reference-assisted repair of wrong-phase sites closes most of the outer-work gap. K^{-1}-trained starts help Newton but not a fixed-point corrector; continued on solver-accepted states, the learned route beats tuned FISTA and a coarse-grid classical route, while a training-free temperature-form FISTA is faster still. These results support selecting candidate metrics from branch structure and judging them by the deployed solver's work.

1 INTRODUCTION

Repeated high-fidelity simulation, including digital-twin workloads, requires solving nonlinear or nonsmooth partial differential equations (PDEs) across changing conditions while retaining the numerical solver's acceptance criteria. Learned warm starts offer a natural hybrid strategy: the network proposes an initial trajectory, while the numerical solver retains the final acceptance test. The training objective therefore plays two roles: it measures equation mismatch, and it shapes the starting point from which the solver proceeds. Where the nonlinearity has branches, as in phase change (solid, phase change, liquid) and contact, the two roles can disagree: the remaining work depends on which branch each grid site of the start occupies. We show how this structure selects a candidate residual metric for training the starts, in two-phase (Stefan-type) diffusion in enthalpy form. For each request (initial state, forcing), a network proposes a whole window of implicit steps, each a box-constrained strictly convex quadratic program, and semismooth Newton corrects them.

Classical squared dual and negative-order (H^{-1}-type) residual norms, such as $r^{\mathsf{T}}K^{-1}r$ with K the stiffness matrix, provide natural residual geometries, while energy-based training objectives are also well established (Section 2). For separated zero-slope branches, we characterize which quadratic metrics are compatible with the entire admissible branch family and measure how that choice changes full-order correction work. With branches held fixed, an implicit step is a linear response; we seek a quadratic metric in which no such response amplifies input differences. For the zero-slope branch of phase change this metric is unique up to scale: the inverse stiffness K^{-1} (Proposition 4.1). When the nonlinearity is placed in a pointwise reaction term instead of the diffusion term, the criterion, imposed at every step size, rejects K^{-1} and admits the identity, so the candidate can point away from the energy norm or toward 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.

We keep the network, decoder, corrector and data fixed and change only this metric. At 4,000 updates, inverse-stiffness weighting (Inverse) reduces Newton outer iterations by **37.0%** relative to Euclidean weighting $r^\top r$, the identity metric (Table 1). The advantage narrows with training; Inverse still uses 7.5% less outer work at 128,000 updates (Figure 1b; Section 4.1). Without labels, solver unrolling or a teacher, Inverse is within 2.5% in median outer iterations of five of eight solver-aware objectives and cheaper than the other three (Table 2). Together, these results make residual geometry a solver-facing design variable rather than merely a loss-design detail. Our contributions:

Branch structure selects a candidate residual metric, and the choice reverses with nonlinear placement. For separated zero-slope branches (Proposition 4.1), K^{-1} is, up to scale, the only quadratic metric in which every fixed-branch response is nonexpansive; for a reaction term the identity is admitted and K^{-1} is not. In four of five prospective placement tests, both metrics accept every request and all four favor the predicted metric; the fifth has no verdict because Euclidean misses one request per seed. A preregistered obstacle test also favors Euclidean (Sections 4.2–4.3).

Correction work is branch-sensitive, not residual-size ordered. Of two constructed starts, the one with the smaller residual in every vector norm needs more exact Newton steps, so minimizing the residual and minimizing correction work are different objectives; learned checkpoints show the same residual/work misordering across training budgets. Reference-assisted repair of a Euclidean start's wrong-phase sites (sites whose phase differs from the reference solution's) recovers 97.6% of the Euclidean–Inverse gap in outer work, against 38.6% for a displacement control (Section 5).

Warm-start value is solver-relative. Continuing training on solver-accepted states uses 31.2–34.9% less outer work in later windows than continuing on the original inputs (Section 6.1). The resulting learned route is 23–24% faster than tuned per-step FISTA in the liquid fraction (q-form) and 47–48% faster than a coarse-grid classical route on a benchmark generated after every route was fixed; a training-free temperature-form FISTA is faster still (Section 6.2). Inverse starts help Newton but not Picard (Section 5.3).

2 Related Work

2.1 Learned Initializers and Solver-Aware Training

Jin et al. (2025) train implicit-residual warm starts and analyze Newton savings, Chevalier et al. (2022) learn inverse-Jacobian updates, and neural-operator initial guesses speed up several solvers

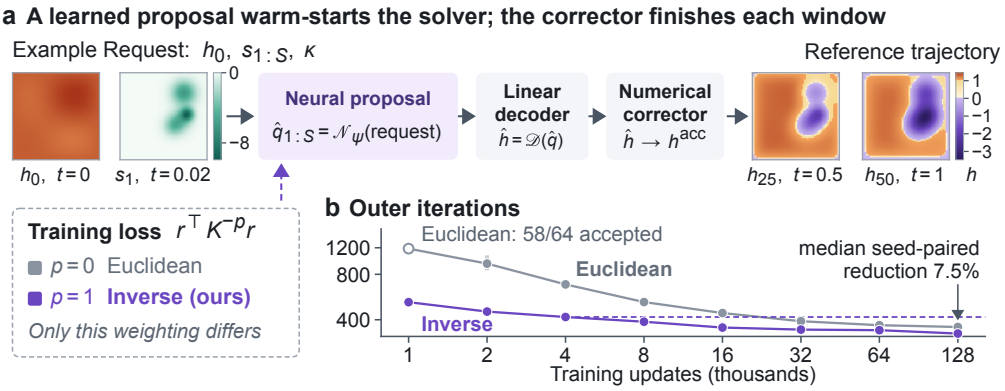

Figure 1: **Residual metric choice changes the correction work of learned warm starts.** (a) A network maps a request to a whole-window proposal; a linear decoder enforces balance and a numerical corrector finishes each step. Training minimizes $r^\top K^{-p} r$: Euclidean ($p = 0$) or Inverse ($p = 1$). (b) Held-out outer iterations, summed over 64 requests, against training updates; each point is a separately completed training schedule (three-seed medians and ranges). Dashed: Inverse at 4,000 updates; at 128,000 updates the median seed-paired reduction is 7.5% (Section 4.1).

(Taghikhani et al., 2026; Zhou et al., 2025; Eshaghi et al., 2026; Aghili et al., 2025). Quadratic-program warm starts are learned by unrolling iterations or predicting active sets (Sambharya et al., 2023; 2024; Klaučo et al., 2019; Schmidtobreick et al., 2026), and networks learn search directions or corrections inside linear solvers (Kaneda et al., 2023; Hsieh et al., 2019; Zhang et al., 2024; Wu et al., 2026). Training on a model's own rollouts mitigates distribution shift (Ross et al., 2011; Brandstetter et al., 2022; Li et al., 2025), whereas we continue training on solver-accepted states.

2.2 Residual Metrics and Nonsmooth Structure

For symmetric positive-definite (SPD) K, $r^\mathsf{T} K^{-1} r$ is the squared discrete dual (H^{-1}-type) norm of the residual, and for a linear system $Ke = r$ it equals the squared energy error $e^\mathsf{T} K e$ (Kirby, 2010; Mardal & Winther, 2011); least-squares finite elements minimize such negative-order norms (Bramble et al., 1997). Ritz, dual-norm and negative-Sobolev objectives train in such norms for accuracy and conditioning (E & Yu, 2018; Rojas et al., 2024; Bachmayr et al., 2025; Cao et al., 2025), and De Ryck et al. (2024) tie the difficulty of physics-informed residual training to the conditioning of a differential operator. Neural preconditioned Born series (Wang et al., 2026) match a learned preconditioner's training metric to its inference geometry, for Helmholtz problems and Newton linear systems. Oh et al. (2026) restore a definite Jacobian for neural-operator Newton warm starts by label-free energy fine-tuning at three-dimensional scale. Ezzehi et al. (2026) embed network corrections in a hyper-reduced semismooth Newton method for parametric variational inequalities. Here the network proposes a full-order trajectory before a fixed nonsmooth corrector; we choose the residual metric from the branch structure, where the candidate reverses with the placement of the nonlinearity, and measure the consequences for correction work.

A smaller residual or a closer warm start can also need more solver work in smooth problems (Um et al., 2020; Lei et al., 2026; Bose et al., 2026; Oh et al., 2026; Taheri & Molzahn, 2026); here the Jacobians stay nonsingular and the difference comes from branches.

3 Problem Setup and Training

3.1 Nonsmooth PDEs and Whole-Window Warm Starts

The network maps a request to a whole-window warm start. Training minimizes a residual surrogate, whereas evaluation measures the numerical correction that remains. The question is which geometry the surrogate should have. A request supplies an initial enthalpy h_0, a source sequence $s_{1:S}$ and coefficients over S steps. In the two-phase model the liquid fraction is $P(h) = \mathrm{clip}(h, 0, 1)$ and the temperature is $\beta(h) = h - P(h)$, piecewise linear with slopes 1, 0 and 1 on its solid, phase-change and liquid branches. Let K be the five-point Dirichlet stiffness matrix at unit spacing on an $N \times N$ interior grid, $c = \Delta t \kappa (N+1)^2$ the diffusion number for diffusivity κ and f_n the resulting time-integrated forcing; the residual of step n is

$$r_n(h) = h_n - h_{n-1} + cK\beta(h_n) - f_n, \qquad n = 1, \dots, S. \tag{1}$$

A causal space–time U-Net predicts the latent liquid fraction $\widehat{q}_{1:S}$ of the whole window, with $\widehat{q}_0 = P(h_0)$. A fixed linear decoder turns it into a start that satisfies linear balance (Figure 1a), so that its residual comes essentially from constitutive mismatch alone:

$$(I + cK)\widehat{T}_n = \widehat{T}_{n-1} + \widehat{q}_{n-1} - \widehat{q}_n + f_n, \quad \widehat{T}_0 = \beta(h_0), \quad \widehat{h}_n = \widehat{T}_n + \widehat{q}_n. \tag{2}$$

The decoded trajectory $\widehat{H} = (\widehat{h}_1, \dots, \widehat{h}_S)$ is the initial guess for numerical correction; the decoder uses no reference solution, and $\widehat{T} = \beta(\widehat{h})$ need not hold.

3.2 Numerical Correction and Acceptance

Semismooth Newton, the primal–dual active-set method standard for such steps (Qi & Sun, 1993; Hintermüller et al., 2002), corrects all S steps of the window together, with right-preconditioned BiCGSTAB (van der Vorst, 1992), inexact forcing and true-residual Armijo backtracking (Dembo et al., 1982; Eisenstat & Walker, 1996). An accepted path is finite, preserves h_0 and satisfies

$$R(h) = \sum\nolimits_{n=1}^{S} \|r_n(h)\|_1 \le \varepsilon. \tag{3}$$

Table 1: **Inverse weighting cuts Newton work; matching its eigenvalues alone does not.**

Comparison (n)	Outer↓	[seed range]	Matvec↓
Inverse (ours) vs. Euclidean			
Main, 4k (6)	-37.0	$[-41.7, -32.9]$	-39.4
Matched residual (6)	-35.2	$[-38.3, -32.3]$	—
Matched field error (6)	-35.7	$[-39.1, -32.9]$	-40.6
Longer, 32k (3)	-12.3	$[-15.9, -10.4]$	-13.9
Scale, diagonal, V-cycle or full K^{-1} vs. Euclidean			
Trace scalar (3)	-0.7	$[-6.1, -0.3]$	-1.7
Initial-loss scalar (3)	$+0.4$	$[-2.3, +1.6]$	-0.7
Inverse diagonal (3)	-2.1	$[-4.7, -2.0]$	-1.5
Full inverse, same seeds (3)	-38.9	$[-41.7, -37.0]$	-41.5
One multigrid V-cycle (3)	-38.9	$[-42.4, -37.6]$	-41.5
Sign-masked K^{-1} vs. paired Inverse			
Random sign mask M2 (6)	$+63.3$	$[+55.3, +69.7]$	—
Random sign mask M3 (3)	$+63.1$	$[+62.5, +68.7]$	—

Table 2: **Static Inverse (shaded) is within 2.5% of five of eight solver-aware objectives.**

Method	Acc.↑	Outer↓	Matvec↓
No labels, unrolling or teacher			
Euclidean	64	702	4,815
Inverse (ours)	64	435	2,746
Energy/BT	64	425	2,711
Unrolled solver steps or teacher			
NSD3	64	446	2,769
JSN2	64	507	3,560
L2WS-FP15	64	430	2,682
L2WS-FP60	64	622	3,820
SIP-Newton	64	438	2,746
Solution labels			
L2WS-Reg1	64	432	2,675
L2WS-Reg60	61	826	5,053

Table 1: paired changes (%), seed medians; separately trained models; all requests accepted. Scalars: $\mathrm{tr}(K^{-1})/N^2$ (trace) or a constant from initial losses. Matched rows count work to a matched residual or field error $E_\infty \le 10^{-4}$. M2, M3: $\Sigma K^{-1}\Sigma$ for two fixed random sign masks $\Sigma = \mathrm{diag}(\pm 1)$ (M1: three of six runs invalid); —: not reported. **Table 2**: trained without continuation (static) on a second held-out set of 64 requests; three-seed median totals; Acc. = accepted of 64, minimum over seeds; failed attempts charged (App. C.1, C.2, G.2). Energy/BT (introduced here): K-energy of the constitutive defect, $2c(d_n^\mathsf{T} K d_n)^{1/2}$, plus the ℓ^1 balance defect. NSD3, JSN2: displacement of three damped or two Jacobian-smoothed Newton steps. L2WS-FPk, Reg$k$ (Sambharya et al., 2024): k unrolled fixed-point steps, residual (FP) or label distance (Reg). SIP-Newton (Holl et al., 2022): regression to a stopped-gradient Newton teacher.

Newton outer iterations and active matrix–vector products (matvecs: per-request Krylov applications, failed attempts included) are reported separately. Together they measure Newton work. Relaxed Picard (Section 5.3) iterates $\widehat{q} \leftarrow (1-\omega)\widehat{q} + \omega P(\widehat{h})$ on the decoded $\widehat{h}$ (ω per request; App. F.1).

3.3 Residual Weighting and Training Protocols

For a batch of M requests, the two training losses average space–time sums of a weighted residual,

$$\mathcal{L}_p = \tfrac{1}{M} \sum_{i=1}^{M} \sum_{n=1}^{S} r_{i,n}^\mathsf{T} K^{-p} r_{i,n}, \quad p = 0 \text{ (Euclidean)}, \quad p = 1 \text{ (Inverse)}, \tag{4}$$

and the controlled comparison changes only this weighting. Split the residual of a decoded start into a balance defect $b_n = \widehat{h}_n - \widehat{h}_{n-1} + cK\widehat{T}_n - f_n$ and a constitutive defect $d_n = \widehat{q}_n - P(\widehat{h}_n)$, so that $r_n = b_n + cKd_n$. The decoder (2) leaves b_n at rounding level (Appendix C.1), and then

$$r_n^\mathsf{T} r_n = c^2 d_n^\mathsf{T} K^2 d_n, \qquad r_n^\mathsf{T} K^{-1} r_n = c^2 d_n^\mathsf{T} K d_n, \tag{5}$$

and in general $r_n^\mathsf{T} K^{-p} r_n = c^2 d_n^\mathsf{T} K^{2-p} d_n$. The two losses are thus two operator weightings of the same constitutive defect: relative to Inverse, Euclidean multiplies each eigenmode of d_n by its eigenvalue of K, stressing rough over smooth defects.

Setup at a glance (fixed across all comparisons unless a section states otherwise).

Problem	Two-phase enthalpy diffusion (1); 64^2 grid, $S = 50$, $\Delta t = 0.02$, $c \in [0.4225, 1.69]$; 128 training / 32 development / 64 held-out requests, shared by every seed and loss.
Model	Causal space–time U-Net for $\widehat{q}_{1:S}$; fixed linear decoder (2); AdamW, learning rate 10^{-3}, cosine schedule, batch 8, 4,000 updates; three seeds, six where stated (Appendices A.1, C.1).
Reporting	Acceptance (3) with $\varepsilon = 4 \cdot 10^{-4}$ at 64^2, scaled with the node count on other grids; effects are seed medians of paired changes (%) in summed held-out counts against Euclidean (negative: less work). *Registered*: methods, seeds and decision rule fixed before the runs they judge, placement tests before data generation (protocol and amendments: Appendix E.1).
Families	Laws $\beta(h)$ named by their branch slopes (two-phase: 1/0/1; Table 3, Appendix D.1). A variable-coefficient K_a and a parabolic obstacle complete the set. *Placement* puts a law in the diffusion term, as in (1), or in a reaction term, $r_n = h_n - h_{n-1} + cKh_n + 4c\beta(h_n) - f_n$.

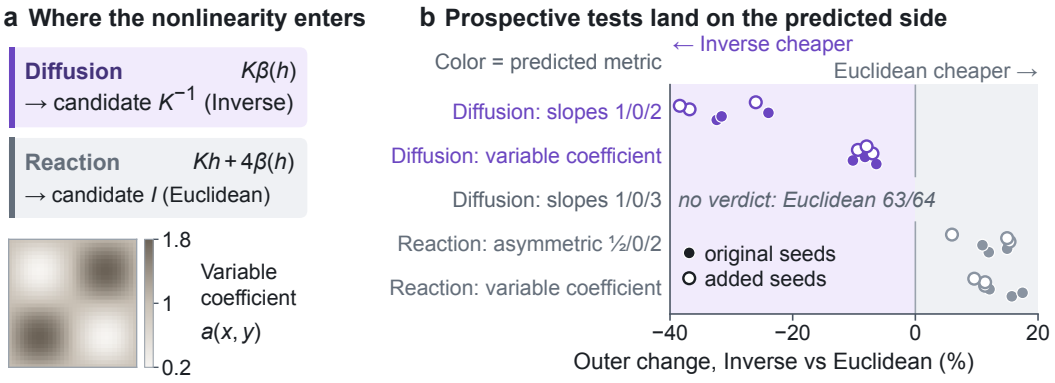

Figure 2: **Nonlinear placement changes the candidate metric, and prospective comparisons with complete acceptance follow the predicted direction.** (a) Nonlinear diffusion points to K^{-1}, nonlinear reaction to I; below: coefficient $a(x, y)$ of K_a. (b) Seed-paired effects of placement tests fixed before training; slopes 1/0/3 has no verdict (Euclidean 63/64).

4 Residual Geometry and Correction Work

4.1 Residual Metric Choice Changes Newton Work

Headline and budget. Changing only the residual metric to inverse stiffness reduces outer work (held-out Newton iterations) by a six-seed median of 37.0% at 4,000 updates, and matvecs by 39.4% (Table 1). In a budget sweep Inverse needs less outer work at every budget and seed; the advantage narrows with training (55.0% at 1,000 updates, 7.5% at 128,000) but persists at the longest tested budget and meets the registered last-doubling criterion (ratios of the 128,000- to 64,000-update reductions: 0.88 and 0.78, against the 0.75 threshold; Table C.2).

What the metric must contain. Under AdamW a constant loss factor leaves the updates essentially unchanged; accordingly, rescaling the Euclidean loss by a scalar or by the slowly varying diagonal of K^{-1} moves median outer work by at most 2.1%. Sign-masked metrics $\Sigma K^{-1}\Sigma$ keep the eigenvalues and diagonal of K^{-1} but reweight frequencies; they cost 63% more outer work than Inverse, within 2.2% of Euclidean (Table 1; App. C.2). Among spectrally equivalent inverses, a single geometric multigrid V-cycle retained the inverse-metric saving, and on the variable-coefficient K_a the constant-coefficient K^{-1} matched K_a^{-1} to within 2.0% at every seed (Appendix E.1).

Solver-aware objectives. Eight solver-aware objectives (Table 2; the table notes define each) share the network, decoder and corrector. Inverse, which uses no labels, solver unrolling or teacher, is within 2.5% in median outer work of five of them and needs 14–47% less than the other three (failed attempts charged). Of those five, three need slightly less: our Energy/BT objective (the unsquared form of Inverse's metric when $b_n \approx 0$), L2WS-FP15 (fifteen unrolled fixed-point steps) and L2WS-Reg1 (solution labels). Labels help Euclidean far more than Inverse, and Inverse without labels beats Euclidean with them at every seed. With labels in both losses the metric saves 6.8%, below the 10% threshold fixed in advance (App. G.2).

4.2 Nonlinear Placement Changes the Candidate Metric

With branches fixed, $\beta(h) = Dh + o$ (diagonal slopes D, offset o) and the implicit step $h + cK\beta(h) = g$, with $g = h_{n-1} + f_n$, responds to g through $(I + cKD)^{-1}$. We seek a metric in which none amplifies differences. An SPD matrix W is *compatible* if $\mathrm{sym}(WKD) \succeq 0$ for every slope matrix $D \succeq 0$ the law admits, where $\mathrm{sym}(A) = (A + A^{\mathsf{T}})/2$.

Proposition 4.1 (Separated zero-slope branches)**.** *Let SPD K have a connected off-diagonal graph, and suppose that for every ordered pair $j \neq k$ some admissible D has $D_{jj} = 0 < D_{kk}$. For SPD W the following are equivalent: (i) $\mathrm{sym}(WKD) \succeq 0$ for every admissible D; (ii) for one (and hence every) $c > 0$, every $(I + cKD)^{-1}$ is nonexpansive in $\|\cdot\|_W$; (iii) $W = \alpha K^{-1}$ for some $\alpha > 0$.*

Table 3: **Where a verdict exists, measured work favors the predicted metric.** Outer change of Inverse vs. Euclidean (%), seed median; 32^2 unless marked 64^2. Reg.: registered before training (App. E.1); *several compatible metrics; no prediction; ‡physical-state step (App. B.3).

Diffusion placement: predicted K^{-1}			Reaction placement and obstacle: predicted I		
Law	Seeds	Outer	Law	Seeds	Outer
Two-phase, 64^2	6	-37.0	Two-phase	6	$+11.9$
Two-phase, 32^2	6	-5.5	Eutectic A, 1/0/1/0/1	6	$+7.4$
Slopes 1/0/2 (reg.)	6	-32.0	Asymmetric, $\frac{1}{2}$/0/2 (reg.)	6	$+13.5$
Variable coefficient (reg.)	6	-8.1	Variable coefficient (reg.)	6	$+11.7$
Eutectic A, 1/0/1/0/1	3	-20.7	Parabolic obstacle, 64^2 (reg.)‡	6	$+22.9$
Asymmetric, $\frac{1}{2}$/0/2	3	-22.8	Diffusion, not established: one-phase, -2.5 (Inverse		
Richards-type, 5/1/5*	3	-48.8	cheaper at 2 of 3 seeds); slopes 1/0/3, no verdict (Fig. 2b).		

With the previous state fixed, compatibility bounds the exact correction $\delta_n = -(I + cKD_n)^{-1}r_n$ of step n by $\|\delta_n\|_W \leq \|r_n\|_W$. A coupled window also propagates earlier corrections, $(I + cKD_n)\delta_n = \delta_{n-1} - r_n$ with $\delta_0 = 0$. This motivates the metric through stable frozen responses; it bounds neither the number of branch changes nor the numerical iterations. For reaction placement the fixed-branch response involves $K + 4D$ instead of KD: the identity makes every $K + 4D$ monotone, whereas K^{-1} fails this all-pattern monotonicity test (App. B.3). The criterion proposes Inverse for diffusion and Euclidean for reaction. Direction (iii)⇒(ii) is the discrete H^{-1} contraction of monotone resolvents (Brézis, 1971; Magenes et al., 1987); new here are the converse, K^{-1}'s incompatibility under reaction placement, the mixed-placement boundary and the measured Newton work.

Placement tests fixed before training. Five placement tests, fixed before data generation (App. E.1), cover new laws and a variable-coefficient operator. Of the four with complete acceptance, two place the nonlinearity in diffusion (predicted Inverse) and two in reaction (predicted Euclidean), and all four land on the predicted side at every seed (Figure 2b). The fifth (slopes 1/0/3) has no verdict because Euclidean misses one request per seed. A 64^2 parabolic obstacle problem, fixed before training like the placement tests but a variational inequality rather than a phase-change law, favors Euclidean as registered ($+22.9\%$; Table 3). Its physical-state step admits I, whereas for the auxiliary residual used in training Proposition 4.1 singles out K^{-1} (App. B.3).

4.3 Mixed Placement Separates Compatibility from Work

Exponents from $p = \frac{1}{2}$ to $\frac{3}{2}$ all keep most of the saving (31.4–38.9%). At $p = 2$ the loss is $c^2 d^{\mathsf{T}} d$, the best-conditioned member of the family in d, yet it saves only 20.4% on the main problem and costs more than Euclidean on the low-diffusion and eutectic-A problems (Table E.3). Extending the reaction exponents to $p = -1$ meets neither registered outcome ($p = 0$, or a negative exponent, cheapest at every seed in both counters) but leaves the preference near the identity: $p = -\frac{1}{2}$ is within 1.3% of $p = 0$ at every seed and $p = -1$ costs more (App. E.2). Mixed placement weights the constitutive term by $B_\theta = \theta K + 4(1-\theta)I$ and linear diffusion by $1 - \theta$, from reaction ($\theta = 0$) to diffusion ($\theta = 1$). The Euclidean metric is compatible exactly for $\theta \leq t_N$, with $t_{64} \approx 0.6677$ (Proposition B.1). The measured preference switches between $\theta = 0.94$ and 0.99, well beyond t_{64}: under mixed placement, compatibility and measured work separate (App. E.2). The matched metric B_θ^{-1}, compatible for $\theta < 1$, was trained at three of the five interior θ: at each it is cheaper than Inverse, yet at none is it cheaper than both I and K^{-1} at every seed in both counts (App. E.2).

Laws, grids and backbones. The saving grows with the diffusion number c: on each grid's own requests (three seeds; six in Table 3) it is 4.9% at 32^2 and 38.9% at 64^2, but only 10.3% at 64^2 with c in the 32^2 range (Table D.4). From scratch under the shared recipe, Euclidean collapses to a near-constant proposal at 128^2 and 256^2, whereas Inverse collapses on neither grid and is accepted for 63–64 of 64 requests at 128^2 (six seeds). At 256^2, a stress test, Inverse is accepted for 17–25 requests against 2, and neither arm reaches the acceptance target of 60 of 64 (Appendix D.2). At 256^2 with the diffusion number matched to the 64^2 range, Euclidean still collapses to a constant

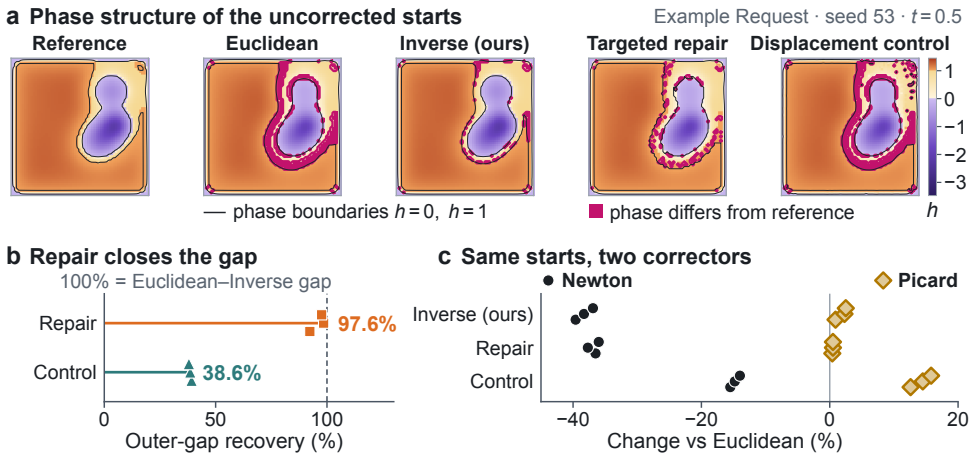

Figure 3: **Reference-assisted repair of wrong-phase sites recovers most of the Euclidean–Inverse gap.** (a) One request's replayed starts and a same-seed accepted reference; magenta: wrong-phase sites. (b) Outer-gap recovery by repair and control (five-draw medians, 3 seeds). (c) Change vs. Euclidean under Newton (outer iterations) and Picard (iterations to acceptance).

proposal yet is accepted on 62 of 64 requests per seed; Inverse does not collapse and uses about 57% less outer work on jointly accepted requests (App. D.2). On a whole-window Fourier neural operator, the diffusion preference transfers, whereas the reaction-side reversal is not reproduced (App. D.2).

5 Branch-Sensitive Correction Work

5.1 Smaller Residuals Can Require More Exact Newton Work

No fixed norm of the residual orders exact Newton work for all starts (steps $-J_F^{-1}F$, J_F a generalized Jacobian of the window residual F). For one two-phase window we construct two decoded starts with liquid fractions in $(0, 1)$ and $F(H_{\text{good}}) = -2F(H_{\text{trap}}) \neq 0$, so the trap's residual is smaller in every vector norm. The good start lies on the solution's affine branch and needs one solve. The trap sits on the plateau and sheds one ring of nodes per update (Papadopoulos & Hintermüller, 2026; Fig. B.1). Exact rational arithmetic at 64^2 certifies one versus four steps, and a family of decoded starts needs one versus $m + 2$ updates for every $m \geq 2$, with grid, data and tolerance depending on m (App. B.5). Lower residual values do not by themselves explain the work advantage: Euclidean starts at 16,000 updates have a smaller summed per-step K^{-1} residual norm S_1 than Inverse starts at 4,000 (medians 13.2–13.8 against 15.9–17.2) and smaller ℓ^1 and Euclidean residuals, yet need 6.2–7.4% more outer work at every seed (Table C.3). At equal budgets, by contrast, the smaller ℓ^1 residual has the smaller work in all 24 same-seed pairs. Across budgets, in 84 same-seed pairs of a longer-trained Euclidean and a shorter-trained Inverse, the ℓ^1 residual orders work in 48, the defect norm $S_2 = \sum_n \|K^{-1}r_n\|_2$ in 73 and the wrong-phase fraction of the starts, measured against an accepted reference, in 79 (retrospective; App. C.1).

5.2 Reference-Assisted Wrong-Phase Repair Recovers Most of the Gap

An exact, undamped Newton step solves its current affine system, $(I + cKD)h^+ = g - cKo$; if that root stays on the same branches, no further update is needed. We therefore test branch structure by intervening on the wrong-phase sites of Euclidean starts, those whose phase differs from the same-seed accepted reference (Figure 3a), for three seeds on 63 new requests not inspected beforehand. *Targeted repair* projects a random subset of these sites into the reference's phase intervals, without copying reference values, leaving as many as the same-seed Inverse start has. The *displacement control* applies the same displacements to as many phase-correct sites, matching the perturbation but adding phase errors. Repair recovers **97.6%** of the same-seed Euclidean–Inverse gap in outer work and the control 38.6% (five-draw medians; Figure 3b). The control recovers less than half of

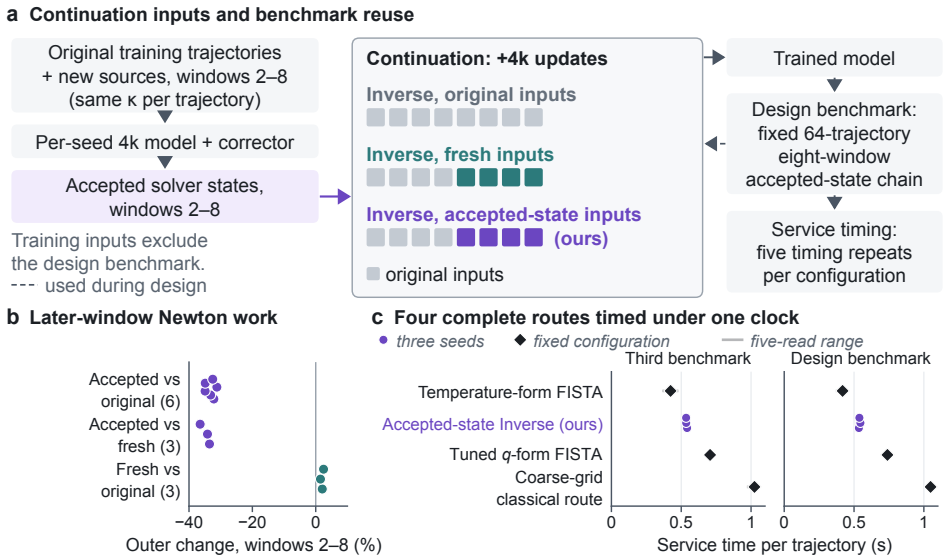

Figure 4: **Accepted-state continuation improves the learned route; solver formulation changes the complete-route ranking.** (a) Continuation inputs. (b) Outer-work change over windows 2–8 (seed pairs in parentheses). (c) Service time per eight-window trajectory in one quiet-host session (App. H.3): four complete routes on the third (left) and reused design (right) benchmarks; five-read medians, bars their range; accepted-state Inverse at three seeds, each classical route one fixed configuration; every read completes 64/64 without fallback; training and preparation excluded.

repair's outer-work gain but more than half of its median matvec gain (App. F). Only the recovery beyond the control distinguishes repair from an equal displacement.

5.3 The Same Starts Help Newton but Not Picard

The three Euclidean and Inverse model pairs give the same starts to Newton and to relaxed Picard on 64 new requests. Inverse starts reduce Newton outer work by 36.9–39.3% (Table F.1), and similarly with other inner solvers (Table F.4). They raise Picard's iterations to acceptance by 0.9–2.5%. Under Picard, repair changes work by less than 1%. FISTA (Section 6.2) does better from its own previous-state start than from either learned start (Appendix F.1): the value of a start depends on its corrector (a retrospective affine-response diagnostic agrees; App. F.2).

6 Solver-Relative Warm-Start Value

6.1 Accepted-State Continuation Reduces Later-Window Work

Repeated solves chain windows into trajectories that feed the network unseen states; the *design benchmark*, used in design and reused for evaluation (Section 6.2 adds one generated later), has 64 eight-window trajectories, each window starting from the last accepted state. From each seed's 4,000-update Inverse model, three continuations add 4,000 updates and differ only in their inputs (Figure 4a). The *accepted-state* continuation adds accepted states from windows 2–8 of the training trajectories, the *original-input* one keeps the original inputs, and the *fresh-input* one adds as many new first-window requests. Over windows 2–8, the accepted-state continuation uses 31.2–34.9% less outer work than the original-input one (six seed pairs) and 34.2% less than the fresh-input one (three seeds; Figure 4b). The design benchmark supplies no continuation inputs (App. H.1).

6.2 Solver Formulation Changes Complete-Route Ranking

Service time is GPU-synchronized wall time per trajectory, charging failed attempts and fallback but not training, one session per comparison (App. H.4). Comparators, defined in App. H.2, step from

the previous accepted state: FISTA (Beck & Teboulle, 2009) on the quadratic program in q (q-form), in both default and development-tuned configurations; the same FISTA in temperature form; and semismooth Newton. A coarse-grid classical route runs FISTA at 32^2 before fine-grid Newton.

Table 4: **Continuation on solver-accepted states lowers service time for both weightings; with these inputs, Inverse runs faster than default q-form FISTA at every seed and Euclidean runs slower.** Seconds per eight-window trajectory, design benchmark (five-repeat medians, all complete 64/64; only Euclidean continuations fall back; default per-step q-form FISTA 0.712 s; App. H.4).

Continuation input	Inverse (ours)			Euclidean		
	s=53	101	202	s=53	101	202
original	0.687	0.709	0.719	1.011	1.032	1.046
fresh first-window	0.694	0.731	0.738		—	
solver-accepted	**0.517**	**0.531**	**0.520**	0.908	0.879	0.907

Under the same continuation-input policy, the Inverse configurations are 31–43% faster than their Euclidean counterparts (Table 4; reductions $1 - t_1/t_2$). Relative to original-input continuation, accepted-state continuation reduces service time by 25–28% for Inverse and 10–15% for Euclidean; each weighting generates its own accepted-state pool (Appendix I). On the design benchmark each static model is slower than FISTA (Table H.2).

One quiet-host session times the routes under one clock on a third benchmark generated after every route was fixed; the reused design benchmark gives the same order. The learned route (accepted-state Inverse) is **23–24%** faster than tuned q-form FISTA and 47–48% faster than the coarse-grid classical route. A change of solver formulation can outweigh initializer gains: FISTA in temperature form needs no training, applies the stiffness stencil instead of an inverse of $I + cK$ at each gradient step, and is 21–22% faster than the learned route (Table H.4; Figure 4c). With labels, supervised continuation was 3–8% faster than accepted-state Inverse in an earlier matched session (Appendix H.3). Under the conditional run-wall accounting of Appendix I, cumulative savings over default q-form FISTA reach the recorded continuation time after 5,472–7,063 trajectories and the total recorded preparation time after 10,190–16,708 (six seeds).

7 Discussion and Conclusion

Scope of the evidence. We test two-dimensional enthalpy diffusion across constitutive laws, nonlinear placements, variable coefficients and 32^2–256^2 grids, together with a parabolic obstacle problem and a second backbone. The diffusion-side preference holds wherever a verdict was reached; the reaction-side reversal is not reproduced by the second backbone, and the high-diffusion 256^2 stress test remains below its acceptance target. The criterion identifies a candidate metric rather than ranking all compatible metrics or locating the mixed-placement switch. Following concerns about weak numerical baselines (McGreivy & Hakim, 2024), the learned route is benchmarked against a strong classical reformulation. Extending the analysis to higher-dimensional problems and to specialized nonlinear multigrid (Kornhuber, 1994; Gräser & Sander, 2019) is a natural next step.

The branch structure of a nonsmooth step supplies a candidate residual metric for training, not a work-optimal one. For the monotone piecewise-linear families studied here, nonlinear diffusion points to an inverse-stiffness metric, whereas the reaction formulations and the physical-state obstacle step point toward the identity. Residual size and correction work are different quantities: a fixed norm measures equation mismatch, while branch information captures aspects of correction work that residual magnitude misses. Warm-start value is solver-relative: continued on solver-accepted states, the learned route beats q-form FISTA and a coarse-grid classical route, while a training-free temperature-form FISTA is faster still. Branch analysis thus gives a principled starting point for choosing the training metric, whose value is then measured by the correction work of the deployed solver and the cost of the complete route. Because the solver applies the same acceptance test whatever the start, this design targets the cost of reaching solver-accepted states without relaxing that test, a useful property for repeated high-fidelity simulation and digital-twin workloads.

AI USE STATEMENT

Generative AI tools assisted research planning; formulation and refinement of hypotheses, experimental designs and controls; mathematical derivations and proof-checking; research-code implementation; organization and analysis of numerical records; literature discovery and source comparison; internal critical review; and drafting, revision, figure design and typesetting. These uses included assistance in developing and checking the theoretical arguments, not only language editing. No data set in this work was generated by a generative AI model; all data are outputs of deterministic numerical solvers. We have reviewed all AI-assisted work: the residual-ordering construction and the 64^2 witness are verified by exact rational certificates, and the reported thresholds and spectral quantities were recomputed independently. The manuscript identifies prospective comparisons, retrospective diagnostics and reuse of the deployment benchmark. Figures distinguish stored numerical results, analytically prescribed illustrations and protocol-consistent visualization replays; no generated image is presented as an experimental measurement or microscopy observation. The authors retain responsibility for the code, analyses, proofs, references, figures and scientific claims.

ETHICS STATEMENT

The current study uses synthetic diffusion problems. No human-subject study or personal data are involved.

REPRODUCIBILITY STATEMENT

Appendix B gives the proofs and assumptions; Appendix A.1 specifies the network, finite request distributions, training and numerical acceptance; Appendix D.1 gives the model problem equations and decoders. Subsequent appendices provide the controls, prospective tests, interventions, comparison summaries and cost definitions, with detailed tables retained in the accompanying supplementary archive. Appendix A distinguishes the typesetting and figure-reproduction files from the numerical run records required to rerun experiments. An anonymized minimal reproduction package (training losses, decoder, corrector, two checkpoints, and scripts that reproduce the headline counts on a subset and the certificate of Section 5.1) is included in the supplementary material. Upon publication, we will release the training and evaluation code, request generators, the trained checkpoints behind the reported comparisons, per-request work records and a timestamped index of the registrations that separate prospective from retrospective analyses.

REFERENCES

Joubine Aghili, Emmanuel Franck, Romain Hild, Victor Michel-Dansac, and Vincent Vigon. Accelerating the convergence of Newton's method for nonlinear elliptic PDEs using Fourier neural operators. *Communications in Nonlinear Science and Numerical Simulation*, 140:108434, 2025. doi: 10.1016/j.cnsns.2024.108434.

Markus Bachmayr, Wolfgang Dahmen, and Mathias Oster. Variationally correct neural residual regression for parametric PDEs: on the viability of controlled accuracy. *IMA Journal of Numerical Analysis*, 2025. doi: 10.1093/imanum/draf073. Advance article.

Amir Beck and Marc Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. doi: 10.1137/080716542.

Shourya Bose, Helgi Hilmarsson, and Dhruv Suri. Newton's Lantern: A Reinforcement Learning Framework for Finetuning AC Power Flow Warm Start Models. arXiv:2605.11102v1, 2026. URL `https://arxiv.org/abs/2605.11102v1`. Preprint.

S. Boyd and Q. Yang. Structured and simultaneous Lyapunov functions for system stability problems. *International Journal of Control*, 49(6):2215–2240, 1989. doi: 10.1080/00207178908559769.

James H. Bramble, Raytcho D. Lazarov, and Joseph E. Pasciak. A least-squares approach based on a discrete minus one inner product for first order systems. *Mathematics of Computation*, 66(219):935–955, 1997. doi: 10.1090/S0025-5718-97-00848-X.

Johannes Brandstetter, Daniel Worrall, and Max Welling. Message Passing Neural PDE Solvers. In *International Conference on Learning Representations*, 2022.

H. Brézis. Monotonicity methods in Hilbert spaces and some applications to nonlinear partial differential equations. In E. H. Zarantonello (ed.), *Contributions to Nonlinear Functional Analysis*, pp. 101–156. Academic Press, New York, 1971. doi: 10.1016/B978-0-12-775850-3.50009-1.

Luigi Brugnano and Alessandra Sestini. Iterative solution of piecewise linear systems for the numerical solution of obstacle problems. *Journal of Numerical Analysis, Industrial and Applied Mathematics*, 6(3–4):67–82, 2011. Preprint arXiv:0912.3222.

Shuhao Cao, Francesco Brarda, Ruipeng Li, and Yuanzhe Xi. Spectral-Refiner: Accurate Fine-Tuning of Spatiotemporal Fourier Neural Operator for Turbulent Flows. In *International Conference on Learning Representations*, 2025.

Samuel Chevalier, Jochen Stiasny, and Spyros Chatzivasileiadis. Accelerating Dynamical System Simulations with Contracting and Physics-Projected Neural-Newton Solvers. In Roya Firoozi, Negar Mehr, Esen Yel, Rika Antonova, Jeannette Bohg, Mac Schwager, and Mykel Kochenderfer (eds.), *Proceedings of The 4th Annual Learning for Dynamics and Control Conference*, volume 168 of *Proceedings of Machine Learning Research*, pp. 803–816. PMLR, 23–24 Jun 2022. URL https://proceedings.mlr.press/v168/chevalier22a.html.

N. G. de Bruijn. Remarks on Hermitian matrices. *Linear Algebra and its Applications*, 32:201–208, 1980. doi: 10.1016/0024-3795(80)90017-8. URL https://pure.tue.nl/ws/files/2165089/597564.pdf.

Tim De Ryck, Florent Bonnet, Siddhartha Mishra, and Emmanuel de Bézenac. An operator preconditioning perspective on training in physics-informed machine learning. In *International Conference on Learning Representations*, 2024.

Ron S. Dembo, Stanley C. Eisenstat, and Trond Steihaug. Inexact Newton Methods. *SIAM Journal on Numerical Analysis*, 19(2):400–408, 1982. doi: 10.1137/0719025.

Jean-Pierre Dussault, Mathieu Frappier, and Jean Charles Gilbert. A lower bound on the iterative complexity of the Harker and Pang globalization technique of the Newton-min algorithm for solving the linear complementarity problem. *EURO Journal on Computational Optimization*, 7 (4):359–380, 2019. doi: 10.1007/s13675-019-00116-6.

Weinan E and Bing Yu. The Deep Ritz Method: A Deep Learning-Based Numerical Algorithm for Solving Variational Problems. *Communications in Mathematics and Statistics*, 6(1):1–12, 2018. doi: 10.1007/s40304-018-0127-z.

Stanley C. Eisenstat and Homer F. Walker. Choosing the Forcing Terms in an Inexact Newton Method. *SIAM Journal on Scientific Computing*, 17(1):16–32, 1996. doi: 10.1137/0917003.

Mohammad Sadegh Eshaghi, Cosmin Anitescu, Navid Valizadeh, Yizheng Wang, Xiaoying Zhuang, and Timon Rabczuk. NOWS: Neural Operator Warm Starts for accelerating iterative solvers. *Computer Methods in Applied Mechanics and Engineering*, 458:118989, 2026. doi: 10.1016/j.cma.2026.118989.

Sofiane Ezzehi, Virginie Ehrlacher, Guillaume Enchéry, and Thibault Faney. A Hyper-Reduced Neural Network-Augmented Semi-smooth Newton Method for Nonlinear Parametric Variational Inequalities. arXiv:2607.21242v2, 2026. URL https://arxiv.org/abs/2607.21242v2. Preprint.

Carsten Gräser and Oliver Sander. Truncated nonsmooth Newton multigrid methods for block-separable minimization problems. *IMA Journal of Numerical Analysis*, 39(1):454–481, 2019. doi: 10.1093/imanum/dry073.

Michael Hintermüller, Kazufumi Ito, and Karl Kunisch. The Primal-Dual Active Set Strategy as a Semismooth Newton Method. *SIAM Journal on Optimization*, 13(3):865–888, 2002. doi: 10.1137/S1052623401383558.

Philipp Holl, Vladlen Koltun, and Nils Thuerey. Scale-invariant Learning by Physics Inversion. In *Advances in Neural Information Processing Systems*, volume 35, pp. 5390–5403, 2022.

Jun-Ting Hsieh, Shengjia Zhao, Stephan Eismann, Lucia Mirabella, and Stefano Ermon. Learning Neural PDE Solvers with Convergence Guarantees. In *International Conference on Learning Representations*, 2019.

Tianyu Jin, Georg Maierhofer, Katharina Schratz, and Yang Xiang. A fast neural hybrid Newton solver adapted to implicit methods for nonlinear dynamics. *Journal of Computational Physics*, 529:113869, 2025. doi: 10.1016/j.jcp.2025.113869.

Ayano Kaneda, Osman Akar, Jingyu Chen, Victoria Kala, David Hyde, and Joseph Teran. A Deep Conjugate Direction Method for Iteratively Solving Linear Systems. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15720–15736, 2023.

Michael G. Kapteyn, Jacob V. R. Pretorius, and Karen E. Willcox. A probabilistic graphical model foundation for enabling predictive digital twins at scale. *Nature Computational Science*, 1:337–347, 2021. doi: 10.1038/s43588-021-00069-0.

Robert C. Kirby. From Functional Analysis to Iterative Methods. *SIAM Review*, 52(2):269–293, 2010. doi: 10.1137/070706914.

Martin Klaučo, Martin Kalúz, and Michal Kvasnica. Machine learning-based warm starting of active set methods in embedded model predictive control. *Engineering Applications of Artificial Intelligence*, 77:1–8, 2019. doi: 10.1016/j.engappai.2018.09.014.

Ralf Kornhuber. Monotone multigrid methods for elliptic variational inequalities I. *Numerische Mathematik*, 69(2):167–184, 1994. doi: 10.1007/BF03325426.

Mingcheng Lei, Weishao Tang, Yufei Zhang, and Haixin Chen. Reliable and efficient steady CFD from surrogate predictions through Newton-Krylov correction. arXiv:2608.04400v1, 2026. URL https://arxiv.org/abs/2608.04400v1. Preprint.

Zhaoxin Li, Xiaoke Wang, Letian Chen, Rohan Paleja, Subramanya Nageshrao, and Matthew Gombolay. Faster Model Predictive Control via Self-Supervised Initialization Learning. arXiv:2408.03394v3, 2025. URL https://arxiv.org/abs/2408.03394v3. Version 3.

Enrico Magenes, Ricardo H. Nochetto, and Claudio Verdi. Energy error estimates for a linear scheme to approximate nonlinear parabolic problems. *ESAIM: Mathematical Modelling and Numerical Analysis (M2AN)*, 21(4):655–678, 1987. URL http://www.numdam.org/item?id=M2AN_1987__21_4_655_0.

Kent-Andre Mardal and Ragnar Winther. Preconditioning discretizations of systems of partial differential equations. *Numerical Linear Algebra with Applications*, 18(1):1–40, 2011. doi: 10.1002/nla.716.

Nick McGreivy and Ammar Hakim. Weak baselines and reporting biases lead to overoptimism in machine learning for fluid-related partial differential equations. *Nature Machine Intelligence*, 6 (10):1256–1269, 2024. doi: 10.1038/s42256-024-00897-5.

Jaemin Oh, Youngkyu Lee, Jerome Darbon, and George Em Karniadakis. Spectrally Safe Neural Operator Warm-Starts for Large-Scale Newton Solvers. arXiv:2606.21828v2, 2026. URL https://arxiv.org/abs/2606.21828v2. Preprint.

Ioannis P. A. Papadopoulos and Michael Hintermüller. Mesh-dependent iteration count growth in primal-dual active set strategies. arXiv:2607.26622v2, 2026. URL https://arxiv.org/abs/2607.26622v2.

Liqun Qi and Jie Sun. A nonsmooth version of Newton's method. *Mathematical Programming*, 58: 353–367, 1993. doi: 10.1007/BF01581275.

Sergio Rojas, Paweł Maczuga, Judit Muñoz-Matute, David Pardo, and Maciej Paszyński. Robust Variational Physics-Informed Neural Networks. *Computer Methods in Applied Mechanics and Engineering*, 425:116904, 2024. doi: 10.1016/j.cma.2024.116904.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635. PMLR, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Rajiv Sambharya, Georgina Hall, Brandon Amos, and Bartolomeo Stellato. End-to-end learning to warm-start for real-time quadratic optimization. In Nikolai Matni, Manfred Morari, and George J. Pappas (eds.), *Proceedings of The 5th Annual Learning for Dynamics and Control Conference*, volume 211 of *Proceedings of Machine Learning Research*, pp. 220–234. PMLR, 2023. URL https://proceedings.mlr.press/v211/sambharya23a.html.

Rajiv Sambharya, Georgina Hall, Brandon Amos, and Bartolomeo Stellato. Learning to warm-start fixed-point optimization algorithms. *Journal of Machine Learning Research*, 25(166):1–46, 2024.

Ella J. Schmidtobreick, Daniel Arnström, Paul Häusner, and Jens Sjölund. Warm-starting active-set solvers using graph neural networks. In Gaurav Sukhatme, Lars Lindemann, Stephen Tu, Adam Wierman, and Nikolay Atanasov (eds.), *Proceedings of The 8th Annual Learning for Dynamics and Control Conference*, volume 331 of *Proceedings of Machine Learning Research*, pp. 662–677. PMLR, 2026. URL https://proceedings.mlr.press/v331/schmidtobreick26a.html.

Guodong Shao, Joe Hightower, and William Schindel. Credibility consideration for digital twins in manufacturing. *Manufacturing Letters*, 35:24–28, 2023. doi: 10.1016/j.mfglet.2022.11.009.

Kianoosh Taghikhani, Yusuke Yamazaki, Jerry Paul Varghese, Markus Apel, Reza Najian Asl, and Shahed Rezaei. Neural-Initialized Newton: Accelerating Nonlinear Finite Elements via Operator Learning. *International Journal for Numerical Methods in Engineering*, 127(14):e70380, 2026. doi: 10.1002/nme.70380.

Babak Taheri and Daniel K. Molzahn. Not All Warm Starts Help: Benchmarking Primal-Dual Initializations for ACOPF Algorithms. arXiv:2606.08984, 2026.

Kiwon Um, Robert Brand, Yun Raymond Fei, Philipp Holl, and Nils Thuerey. Solver-in-the-Loop: Learning from Differentiable Physics to Interact with Iterative PDE-Solvers. In *Advances in Neural Information Processing Systems*, 2020.

H. A. van der Vorst. Bi-CGSTAB: A Fast and Smoothly Converging Variant of Bi-CG for the Solution of Nonsymmetric Linear Systems. *SIAM Journal on Scientific and Statistical Computing*, 13(2):631–644, 1992. doi: 10.1137/0913035.

Juntao Wang, Jiwei Jia, and Xinliang Liu. Neural Preconditioned Born Series: A Metric-Matched Framework for Learning-based Preconditioners. arXiv:2603.18527v4, 2026.

Yuhan Wu, Jan Willem van Beek, Victorita Dolean, and Alexander Heinlein. Are Deep Learning Based Hybrid PDE Solvers Reliable? Why Training Paradigms and Update Strategies Matter. *Computing in Science & Engineering*, 2026. doi: 10.1109/MCSE.2026.3696587.

Moslem Zamani and Milan Hladík. Error bounds and a condition number for the absolute value equations. *Mathematical Programming*, 198:85–113, 2023. doi: 10.1007/s10107-021-01756-6.

Enrui Zhang, Adar Kahana, Alena Kopaničáková, Eli Turkel, Rishikesh Ranade, Jay Pathak, and George Em Karniadakis. Blending neural operators and relaxation methods in PDE numerical solvers. *Nature Machine Intelligence*, 6:1303–1313, 2024. doi: 10.1038/s42256-024-00910-x.

Xu-Hui Zhou, Jiequn Han, Muhammad I. Zafar, Eric M. Wolf, Christopher R. Schrock, Christopher J. Roy, and Heng Xiao. Neural operator-based super-fidelity: A warm-start approach for accelerating steady-state simulations. *Journal of Computational Physics*, 529:113871, 2025. doi: 10.1016/j.jcp.2025.113871.

A PROBLEM, DATA AND NUMERICAL CONTRACTS

Appendix contents. The appendices develop the main text, organized by the question a reader wants answered; Table A.1 maps them to the main-text claims, and the PDF bookmark panel lists every section and subsection. Numbers and page numbers are links.

Table A.1: Appendix map: where each main-text claim is developed. The last column links to the page on which the named evidence starts.

Claim or setting	Main text	Appendix; tables	Page
Network, data, training, corrector and acceptance	§3	A.1	14
Inverse weighting reduces Newton work	§4.1	C.1; C.1	24
Scale, diagonal and eigenvalues do not reproduce it; a V-cycle does	§4.1	C.1, C.2	24
The saving persists at the longest tested budget	§4.1	C.1; C.2	25
Matched labels and solver-aware baselines	§4.1	G.2; G.1, G.2	35
Decoder balance holds numerically	§3	C.1	23
Only multiples of K^{-1} are compatible; finite-grid boundary	§4.2, 4.3	B.1–B.3; B.1	16
Nominated placement preferences	§4.2	E.1; E.1	30
Laws, grids, coefficients, backbone	§4.3	D.1, D.2; D.1–D.4	27
Exponent and mixed placement	§4.3	E.2; E.3, E.4	32
Residual size can misorder work	§5.1	B.4, B.5, C.1; B.2, C.3	19
Branch repair reduces Newton work	§5.2	F.1; F.2, F.3	34
Starts serve Newton and Picard differently; inner-solver variants	§5.3	F.1; F.1, F.4	33
Frozen-branch response tracks work	§5.3	F.2; F.5	35
Accepted-state inputs help later windows	§6.1	H.1; H.1	37
Continued Inverse beats default and tuned q-form FISTA and a coarse-grid classical route; temperature-form FISTA is faster; labels help further	§6.2	H.2–H.4; H.2–H.4	37
Clocks and read qualification behind every service time	§6.2	H.4	40
Preparation cost and amortization	§6.2	I	41

Appendix guide and evidence access. Table E.2 states, for each registered comparison behind a main-text claim, when its rule was fixed and where its outcome is reported. Process records (timing-quality checks, overlap checks between timing repeats, per-request counts, histories of earlier experiments) and values at their original precision are in the supplementary details, and the supplementary archive holds the exact certificate values, full exponent tables, model-problem ranges and the comparison matrix; the evidence ledger, original experiment modules, mixed-placement tables, baseline counts and the files that rebuild the typesetting and figures from the recorded fields and results will be released upon publication, whereas rerunning training additionally requires the code, requests, checkpoints and registrations. Material moved out of this appendix (the amended entry test, the full registration index, further precedents, an archived training arm, per-session service times and an earlier deployment campaign) is in Part 3 of the supplementary details, Sections S3.1–S3.6. Percentages here have one decimal unless a source table carries two, and seconds three.

A.1 IMPLEMENTATION AND NUMERICAL CONTRACTS

Network, inputs and finite requests. The main network uses tensor order (request, channel, time, y, x). A $7 \to 16$ pointwise lift feeds causal blocks at widths 16, 32 and 64, with two spatial downsamplings: kernel $(1, 4, 4)$, stride $(1, 2, 2)$, padding $(0, 1, 1)$. Each encoder scale has one dilation-one block; the coarse trunk has six blocks with temporal dilations $1, 2, 4, 8, 16, 32$. Two nearest-neighbor upsamplings concatenate skips, mix through pointwise convolutions ($64+32 \to 32$, $32 + 16 \to 16$), and each add a dilation-one block. A $16 \to 1$ sigmoid head predicts 50 future latent frames; $q_0 = P(h_0)$ is exact. There are 162,753 parameters (162,736 trunk, 17 head).

A width-w, dilation-d residual block contains channel LayerNorm at each site, depthwise spatial kernel $(1, 3, 3)$ with $(0, 1, 1)$ padding, temporal left padding $2d$, depthwise kernel $(3, 1, 1)$ with dilation $(d, 1, 1)$, pointwise $w \to 2w$, GELU and pointwise $2w \to w$. No dropout or BatchNorm is used. The

temporal receptive-field span is 135 positions; there is no temporal downsampling or future-frame mixing. The input channels are $(2h_0 - 1,\ 2P(h_0) - 1,\ s_n/7,\ \log(\kappa/0.01),\ n\Delta t,\ \eta_y,\ \eta_x)$, where η uses `linspace(-1,1,N)`, and h_0, κ repeat across frames. Decoder boundary temperature is -0.2, charged once per adjacent boundary edge. The DST-I eigenvalues are $\lambda_{jk} = 4 - 2\cos(\pi j/(N+1)) - 2\cos(\pi k/(N+1))$. Training uses a sequential spectral scan; evaluation uses an equivalent causal convolution with a zero-padded length-$2S$ temporal FFT.

Physical coordinates are $x_i = i/(N+1)$, $t_n = 0.02n$. Draws occur in the following order: $\kappa \sim \exp U(\log .005, \log .02)$; $J \sim U\{1,2,3,4\}$; Gaussian centers $\mu_j \sim U(.15,.85)^2$; widths $\sigma_j \sim U(.05,.12)$; amplitudes $a_j \sim U(3,7)$; four quarter multipliers $v_{j,\ell} \sim U(.5,1.5)$; initial bias $b \sim U(-.4,1.4)$; and coefficients $a^{(0)}_{kl} \sim U(-.15,.15)$ for ordered modes $(1,1),(1,2),(2,1),(2,2)$. Then

$$h_0 = b + \sum_{kl} a^{(0)}_{kl} \sin(k\pi x)\sin(l\pi y), \quad s_n = \sum_j \zeta_{\ell(n)} a_j v_{j,\ell(n)} \exp[-\|(x,y) - \mu_j\|^2/(2\sigma_j^2)].$$

Here $\ell(n) = \min(\lfloor t_n/.25 \rfloor, 3)$. The source signs $(\zeta_0, \ldots, \zeta_3)$ over the four quarters of the window follow one of four programs, $++++$, $----$, $++--$ or $--++$, assigned by request index modulo four. Initial enthalpy is not clipped; requests are not filtered by phase or difficulty. Split seeds 9141601/9141602/9141603 generate 128/32/64 requests, with one spawned child seed per request. Training repeatedly permutes the fixed 128. The separate 64-request accuracy study uses seed 9170601; transfer evaluates the same 16 physical programs at both grids.

Training and budget calibration. AdamW uses explicit learning rate 10^{-3} and weight decay 10^{-4}; resolved defaults are betas $(.9, .999)$, epsilon 10^{-8}, decoupled decay, and no AMSGrad. Cosine decay spans the entire update budget and ends at zero, with one scheduler step per optimizer step. Batch size is eight; the global gradient norm is clipped at one after the backward pass, and pre-clip norms are recorded. Initialization seeds are 53/101/202, and `default_rng(53)` fixes the common minibatch order. The final update-4,000 model is used; the diagnostics at update 1 and every 1,000 updates do not select a checkpoint. Losses sum time and spatial entries and average requests, without a factor $1/2$ or division by N^2S.

LayerNorm (epsilon 10^{-5}, affine) and exact-erf GELU are used; training/network arithmetic is float32 and deployed physics float64, with deterministic cuDNN; library versions and determinism settings are in the supplementary details. Bitwise reproducibility across devices was not tested.

Calibrated updates are $U = \mathrm{round}(4000 w_{\mathrm{BT}}/w)$ (ties-to-even), with a cosine schedule of length U. Prices are steady-state update intervals on a dedicated RTX 4090 after 20 warm-up updates, measured over 60-update prefixes for the solver-in-the-loop methods ($w_{\mathrm{BT}} = 0.185$ s) and over 120-update prefixes with three rotated cycles for the external ones ($w_{\mathrm{BT}} = 0.191$ s); the resulting U are in Table G.1, and the individual prices in the supplementary details. Startup, data loading, diagnostics and saving are excluded from these prices but retained in recorded training walls, so calibration equalizes prices, not realized training time. The main learning rate was inherited from earlier development runs.

Newton, acceptance and explicit recovery. Newton uses merit $\phi(h) = \sum_n \|r_n(h)\|_1$, exact h_0, outer/inner caps 40/80 and Armijo coefficient 10^{-4}. At each iteration, classify solid $h \le 0$, liquid $h \ge 1$ and mixed otherwise. Let M mask mixed future sites and m count them. Set

$$\eta_k = \min(.5, \sqrt{\phi_k/\phi_0}), \quad F_k = \max(\eta_k \phi_k, .1\varepsilon), \quad \tau_k = F_k/(16c\sqrt{\max(m,1)}).$$

The fixed-phase linear solve acts on the mixed sites through the decoder's temperature map $T = T_{\mathrm{known}} + \Gamma q$, $\Gamma = -(B_t + cK)^{-1} B_t$ (B_t the backward time difference), with liquid sites fixed at $q = 1$; right-preconditioned BiCGSTAB uses $\mathcal{P}v = M[Mv + \mathrm{cumsum}_n(cK(Mv))]$, which restricts the inverse of the unrestricted map $-\Gamma$ to free sites and is exact when all sites are free (recovery formulas in the supplementary details). BiCGSTAB uses per-request products, a fixed shadow residual and absolute target τ_k; zero denominators or nonfinite scalars end that request without restart. A candidate must first have finite current-branch affine merit at most F_k. Backtracking tries $\alpha = 2^{-j}$, $j = 0, \ldots, 20$, accepting the first finite candidate with

$$\phi(h_k + \alpha d) \le [1 - 10^{-4}\alpha(1 - F_k/\phi_k)]\phi_k \quad \text{or} \quad \phi \le \varepsilon.$$

A failed search keeps the last accepted path; no other numerical rescue is applied. Evaluation independently checks finite state, prescribed h_0 and the full physical residual. Outer attempts, inner iterations, active matvecs and backtracks count different work; failed attempts remain charged. Active matvecs count executed per-request Krylov operator actions (one per Krylov-solve initialization, up to two per BiCGSTAB iteration), retained after failure; preconditioner applications are counted separately.

For stepwise FISTA set $b = h_{n-1} + \Delta t\, s_n + c g_\partial$ and $G = (I + cK)^{-1}$. Solve $\min_{0\le q\le 1} \frac{1}{2}(q - b)^{\mathsf{T}} G(q - b)$, recover $T = G(b - q)$ and $h = q + T$. This strictly convex program is equivalent to the implicit step $h + cK\beta(h) = b$ (its optimality conditions are the two-phase branches of β). With $L = (\min(1 + c\lambda))^{-1}$, $\mu = (\max(1 + c\lambda))^{-1}$ and $\theta = (\sqrt{L} - \sqrt{\mu})/(\sqrt{L} + \sqrt{\mu})$, the iteration is $q^+ = \mathrm{clip}(y + G(b - y)/L, 0, 1)$, $y^+ = q^+ + \theta(q^+ - q)$. Its condition number $L/\mu = \max(1 + c\lambda)/\min(1 + c\lambda)$ stays below $1 + 8c$: at most 14.4 at 64^2, where $c \le 1.69$, but $c = \Delta t \kappa (N + 1)^2$ grows with the grid, so for the same diffusivities the condition number reaches 54 at 128^2 and 211 at 256^2. Specialized solvers for this problem class, such as monotone multigrid (Kornhuber, 1994) and truncated nonsmooth Newton multigrid (Gräser & Sander, 2019), are outside the benchmark; a fair comparison would need the same discretization, acceptance test, tuning effort and timing protocol. Initialize $q = y = P(h_{n-1})$. Use DST-I inverses and true step-residual target ε/S, checked at iteration zero, every five iterations and at the 2,000 cap. A batch stops when all pass or at the cap; the returned state seeds the next step. FISTA counters are shared-batch quantities. In closed-fallback deployment, failed Newton requests restart this solver from the same window input, in ordered batches of at most eight. A finite fallback replaces the output; an unaccepted state remains contaminated. A nonfinite fallback retains a finite Newton state, or breaks the chain if neither is finite. Recovery work and completion are reported separately.

This finite-state recovery applied to an earlier deployment campaign (Section S3.6 of the supplementary details). The service timing uses a separate strict-acceptance contract: only an accepted primary or fallback output advances the parent trajectory; a rejected fallback ends that parent and leaves later windows unattempted (Appendix H.4).

Grid and reference contracts. All base tests use 50 steps of .02. At $N = 32, 64, 128$, $\varepsilon_N = 10^{-4}(N/32)^2$ preserves $\varepsilon_N/N^2 = 9.765625 \cdot 10^{-8}$, a per-node time-summed budget, not a field or continuum error bound. Eutectic/one-phase use the projected-free corrector; asymmetric/Richards use slope-free; reaction uses reaction-free; the obstacle uses contact-restricted. Their 32^2 threshold is 10^{-4}. The 48^2 test uses $2.25 \cdot 10^{-4}$. Single-window batches are eight; the chain has 64 trajectories of eight windows with $4 \cdot 10^{-4}$ per window. Transfer uses 16 requests, batch one and $1.6 \cdot 10^{-3}$ on the fine grid. The diagnostic tolerance sweep uses $4 \cdot 10^{-k}$, $k = 2, \ldots, 6$, on its 64 requests (or first eight unbatched with $k = 2, 4, 6$); uncorrected outputs do not become accepted by this ladder.

Stored 64^2 references target $4 \cdot 10^{-7}$ full-window residual ($8 \cdot 10^{-9}$ per step; FISTA cap 4,000 with Newton fallback), nothing is filtered, and the accuracy study excludes thresholds below their measured residual floor. Transfer references come from two independently initialized fine-grid FISTA passes that must both accept, preserve h_0 and agree within field $L_\infty \le 10^{-5}$; the stricter pass supplies field error, and agreement is not a mesh-convergence proof. Pass settings are in the supplementary details.

B Proofs and restricted correction certificates

B.1 Separated branches: proof of Proposition 4.1

Sufficiency of inverse stiffness is the discrete counterpart of the H^{-1} contraction of monotone resolvents (Brézis, 1971); the finite-grid boundary of Proposition B.1 uses Lyapunov and Jordan-product bounds (Boyd & Yang, 1989; de Bruijn, 1980).

Let $A_D = I + cKD$. It is invertible: KD is similar to the positive semidefinite matrix $K^{1/2} D K^{1/2}$, so A_D has positive eigenvalues. If $\mathrm{sym}(WKD) \succeq 0$, choose an admissible $D = \mathrm{diag}(d)$ with $d_j = 0 < d_k$. The (j, j) entry of this symmetric matrix is zero. Its (j, k) entry is $(WK)_{jk} d_k/2$, which must vanish since a positive semidefinite matrix with zero diagonal has the corresponding row zero. Thus $(WK)_{jk} = 0$ for every distinct ordered pair. Hence $WK = \Lambda$ is diagonal.

For the single-c assertion, nonexpansiveness is equivalent to $A_D^{-T} W A_D^{-1} \preceq W$, or, by congruence,

$$Q_D = A_D^T W A_D - W = c(DKW + WKD) + c^2 DKWKD \succeq 0. \tag{6}$$

Again $d_j = 0$ makes $(Q_D)_{jj} = 0$ and $(Q_D)_{jk} = c(WK)_{jk} d_k$. The same row argument gives $WK = \Lambda$. Now $W = \Lambda K^{-1}$ and symmetry imply $\Lambda K = K\Lambda$. For each nonzero off-diagonal K_{jk}, $\Lambda_{jj} = \Lambda_{kk}$. Connectedness implies $\Lambda = \alpha I$; positive definiteness makes $\alpha > 0$.

Conversely, $W = \alpha K^{-1}$ gives $WKD = \alpha D \succeq 0$, so condition (i) holds. For any x, y and $u = A_D^{-1} x$, $v = A_D^{-1} y$, monotonicity gives

$$\langle u - v, x - y \rangle_W = \|u - v\|_W^2 + c\langle u - v, KD(u - v) \rangle_W \geq \|u - v\|_W^2. \tag{7}$$

Cauchy–Schwarz proves $\|u - v\|_W \leq \|x - y\|_W$ for every $c > 0$. This proves all equivalences. □

B.2 Finite-grid Euclidean boundary: statement and proof

Let $B_\theta = \theta K + 4(1 - \theta) I$ and $J_D = (1 - \theta) K + B_\theta D$ interpolate reaction ($\theta = 0$) and diffusion ($\theta = 1$) placement. This spatial branch family is not the full-window Newton Jacobian J_F.

Proposition B.1 (Finite-grid Euclidean boundary). *For the $N^2 \times N^2$ five-point Dirichlet matrix K and $0 \leq \theta \leq 1$, $\mathrm{sym}(J_D) \succeq 0$ for every $D = \mathrm{diag}(d)$, $d \in [0,1]^{N^2}$, if and only if $\theta \leq t_N = (64 - 2\rho_N^2)/(64 - \rho_N^2)$ with $\rho_N = 4\cos(\pi/(N+1))$. Strict positive definiteness for every D holds exactly when $\theta < t_N$; $t_{64} \approx 0.6677$.*

Proof. For $N \geq 2$, write $a_i = 1 - \theta + d_i$ and $b_{ij} = 1 - \theta + \theta(d_i + d_j)/2$ on every grid edge. The symmetric part has diagonal $4a_i$ and edge entries $-b_{ij}$. Let

$$\phi_{p,q} = \sin \frac{\pi p}{N+1} \sin \frac{\pi q}{N+1} > 0, \qquad \eta = 4/\rho_N. \tag{8}$$

The adjacency satisfies $\sum_{j \sim i} \phi_j = \rho_N \phi_i$. Therefore the exact identity is

$$\mathrm{sym}(J_D) = \sum_{\{i,j\} \in E} \mathrm{embed}_{ij} \begin{bmatrix} \eta a_i \phi_j / \phi_i & -b_{ij} \\ -b_{ij} & \eta a_j \phi_i / \phi_j \end{bmatrix}. \tag{9}$$

The Perron ratios reproduce the full Dirichlet diagonal; there is no extra boundary diagonal. At binary slopes, the pure 0–0 and 1–1 blocks are positive definite for $\theta < 1$. For a mixed edge, nonnegative determinant is equivalent to

$$16(1 - \theta)(2 - \theta) - \rho_N^2 (1 - \theta/2)^2 = \frac{2 - \theta}{4} [64 - 2\rho_N^2 - (64 - \rho_N^2)\theta] \geq 0, \tag{10}$$

that is, $\theta \leq t_N$. Every node belongs to an edge, so below t_N the sum is positive definite; at t_N it is positive semidefinite. Affinity in d expresses every nonbinary case as a convex combination of binary-vertex matrices, extending the conclusion to the cube.

For necessity, a checkerboard 0–1 pattern attains

$$\lambda_{\min}(\mathrm{sym}(J_D)) = 6 - 4\theta - \sqrt{4 + (1 - \theta/2)^2 \rho_N^2}. \tag{11}$$

It vanishes at t_N and is negative above it. This also rules out strict definiteness at the boundary. For $N = 1$, $K = [4]$ and $J_D = [4(1 - \theta + d)]$; $t_1 = 1$ gives both assertions directly without division by $\rho_1 = 0$. Since t_N decreases in ρ_N^2 and $\rho_N \uparrow 4$, $t_N \downarrow 2/3$. □

Compatibility of B_θ^{-1}. For $\theta < 1$, B_θ is symmetric positive definite and $B_\theta^{-1} J_D = (1 - \theta) B_\theta^{-1} K + D$. Since B_θ is a polynomial in K, it commutes with K, so $B_\theta^{-1} K$ is symmetric positive semidefinite, and $\mathrm{sym}(B_\theta^{-1} J_D) = (1 - \theta) B_\theta^{-1} K + D \succeq 0$ for every $D = \mathrm{diag}(d)$, $d \in [0,1]^{N^2}$: B_θ^{-1} is compatible with the whole family.

B.3 ALL-PATTERN COMPATIBILITY CLASSES

The classification states which fixed metrics are compatible with every admissible branch pattern of each model problem, not only with those a trained trajectory visits. Let K be the connected, symmetric positive-definite, unit-spacing Dirichlet matrix, $c > 0$, and $\|x\|_W^2 = x^T W x$ for one fixed $W = W^T \succ 0$. For a finite matrix family $\mathcal{A}$, define

$$\mathcal{M}(\mathcal{A}) = \{W \succ 0 : \mathrm{sym}(WA) \succeq 0 \; \forall A \in \mathcal{A}\}, \qquad \mathcal{S}(\mathcal{A}) = \{W \succ 0 : WA = A^T W \; \forall A \in \mathcal{A}\},$$
$$\mathcal{R}_c(\mathcal{A}) = \{W \succ 0 : Q_c(W, A) \succeq 0 \; \forall A \in \mathcal{A}\}, \qquad Q_c(W, A) = 2c \; \mathrm{sym}(WA) + c^2 A^T W A.$$

$\mathcal{M}$ is monotonicity, $\mathcal{R}_c$ is nonexpansiveness of $(I + cA)^{-1}$ in the W-norm at step c, and $\mathcal{S}$ is self-adjointness. $\mathcal{M} \subseteq \mathcal{R}_c$ for every c, and $\mathcal{R}_c$ along $c \downarrow 0$ implies $\mathcal{M}$; a single c generally does not. $\mathcal{S}$ is a separate requirement. For a finite family, $\mathcal{M}$ and $\mathcal{R}_c$ are by definition intersections of finitely many LMIs, so the sets in Table B.1 are exact sets, not sampled searches. The closed forms in the first two rows and the Richards $\mathcal{S}$ entry are proved below; the other Richards entries and the obstacle and reaction rows are supported by proof sketches only.

Table B.1: Compatibility sets over all admissible branch patterns. $D = \mathrm{diag}(d)$, $\alpha > 0$, all sets restricted to SPD W. Rows 1–2 and the Richards $\mathcal{S}$ entry are proved; the other Richards entries and the obstacle and reaction rows rest on proof sketches, not complete proofs.

Model problem; branch family	$\mathcal{M}$	$\mathcal{R}_c$	$\mathcal{S}$
Two-phase, eutectic A, one-phase diffusion; $A = KD$, $d_i \in \{0, 1\}$	αK^{-1}	αK^{-1}, any single $c > 0$	αK^{-1}
Asymmetric B diffusion; $A = KD$, $d_i \in \{0, \frac{1}{2}, 2\}$	αK^{-1}	αK^{-1}, any single $c > 0$	αK^{-1}
Richards-type diffusion; $\{K(I + 4E) : E_{ii} \in \{0, 1\}\}$	nonunique, interior around K^{-1}	$\supseteq \mathcal{M}$; nonunique	αK^{-1}
Physical obstacle state; $cK + N_C$, $C = \{u \geq \psi\}$	diagonal $\Lambda \succ 0$, $\Lambda K + K\Lambda \succeq 0$	$\{W : G_F^T W G_F \preceq W \; \forall F\}$, strictly larger than $\mathcal{M}$ in dimension > 1	αI
Two-phase, eutectic A reaction; $cK + aD$, $a = 4c$, $d_i \in \{0, 1\}$	$\mathcal{M}(\{K + 4D\})$, nonunique, interior around I	$\supseteq \mathcal{M}$	αI

For the obstacle, N_C is the Euclidean normal cone and $G_F = E_F(I + cK_{FF})^{-1} E_F^T$ for free set F with coordinate embedding E_F, $G_\varnothing = 0$. Here $\mathcal{M}$ means monotonicity of $cK + N_C$ for every obstacle ψ, while $\mathcal{R}_c$ tests the resolvent $(I + cK + N_C)^{-1}$. The row classifies the physical-state step. The registration of the 64^2 obstacle comparison (Appendix D.1) named no coordinate chart; it predicted Euclidean on the ground that the projection nonlinearity does not enter the diffusion operator. The training residual and Newton block of Appendix D.1 use the auxiliary y chart, in which the projection, with slopes in $\{0, 1\}$, enters the diffusion operator; there Proposition 4.1 applies and singles out αK^{-1}. The measured preference for Euclidean ($+22.9\%$, Table 3) supports the registered direction and agrees with the physical-state classification. It does not validate applying the criterion directly to the implemented training residual.

Proofs for rows 1–2 and the Richards $\mathcal{S}$ entry. The row argument in the proof of Proposition 4.1 uses only patterns with $d_j = 0 < d_k$, which both slope sets contain, and applies to both $\mathrm{sym}(WKD)$ and $Q_c(W, KD)$; its converse holds for every $D \succeq 0$. For $\mathcal{S}$, two patterns A, A' that differ only at coordinate j (in D or E) give $WK\mathbf{1}_{\{j\}}\mathbf{1}_{\{j\}}^T = \mathbf{1}_{\{j\}}\mathbf{1}_{\{j\}}^T KW$ after subtracting $WA' = A'^T W$ from $WA = A^T W$ and dividing by the nonzero slope difference. The left side is supported in column j and the right side in row j, so $(WK)_{ij} = 0$ for $i \neq j$; symmetry and connectedness then give $W = \alpha K^{-1}$ as in that proof. Conversely, αK^{-1} makes $WKD = \alpha D$ and $WK(I + 4E) = \alpha(I + 4E)$ symmetric.

Inverse stiffness is not compatible with the reaction family. Take one active interior site k, $D = e_k e_k^\mathsf{T}$, and $g = K^{-1} e_k$. Then $\mathrm{sym}(K^{-1}(K + 4D)) = I + 2(g e_k^\mathsf{T} + e_k g^\mathsf{T})$; on $\mathrm{span}\{e_k, g\}$ the

rank-two part has eigenvalues $G_{kk} \pm \|g\|_2$, so the smallest eigenvalue is $1 + 2(G_{kk} - \|g\|_2)$. At the center site it is -4.69 at 32^2 and -11.36 at 64^2, so $K^{-1} \notin \mathcal{M}(\{K + 4D\})$.

Proof sketches for the remaining entries. *Richards:* $\mathcal{M}$ is the displayed finite LMI intersection; $W = K^{-1} + B \succ 0$ with $B = B^T$, $\|BK\|_2 \le 1/\sqrt{5}$ is compatible, because $D^{-1/2}\,\mathrm{sym}(WKD)D^{-1/2} \succeq (1 - \sqrt{5}\|BK\|_2)I$; strict inequality gives an open neighborhood, not the whole boundary. *Obstacle:* an arbitrarily large normal at one active coordinate, with the others free, forces W diagonal; interior pairs then require the Lyapunov inequality, and coordinate-wise normal-cone monotonicity proves sufficiency. At fixed c, $\|G_F\|_2 \le (1 + c\lambda_{\min}(K))^{-1}$ and $\mathrm{cond}_2(W) \le (1 + c\lambda_{\min}(K))^2$ certify $\mathcal{R}_c$, including nondiagonal W near I. Singleton free sets force a self-adjoint W diagonal and the full free set forces commutation with K, giving αI. *Reaction:* the step derivative is $I + c(K + 4D)$, not $I + cKD$; at $W = I$ every $K + 4D$ is positive definite, and finiteness gives an open $\mathcal{M}$-neighborhood; $\mathcal{S}$ follows as for the obstacle. The diagonal-only classification for unbounded reaction slopes does not describe this bounded family.

For these piecewise-affine laws, integrating branch derivatives along a segment makes the $\mathcal{R}_c$ tests equivalent to nonexpansiveness of the nonlinear step; $\mathcal{M}$ extends from slope vertices to their convex hull by affinity, which is not a vertex rule for the quadratic Q_c. These stability classes do not order training or correction work (Section 4).

Spectrally equivalent metrics. For SPD W write $\|A\|_W = \|W^{1/2}AW^{-1/2}\|_2$ for the expansion of a linear map in the W-norm. If $aK^{-1} \preceq W \preceq bK^{-1}$ for constants $0 < a \le b$ and a frozen-branch response $A = (I + cKD)^{-1}$ is nonexpansive in the K^{-1}-norm (Proposition 4.1), then $\|Ax\|_W^2 \le b\|Ax\|_{K^{-1}}^2 \le b\|x\|_{K^{-1}}^2 \le (b/a)\|x\|_W^2$, so $\|A\|_W \le (b/a)^{1/2}$. For the V-cycle weight W_V of Appendix C.1, whose $W_V K$ has spectrum in $[0.871, 1.000]$, every frozen-branch response therefore expands by at most $(1/0.871)^{1/2} \approx 1.072$ in the W_V-norm, which is why an approximate inverse nearly retains the stability property of the exact one. This bound concerns frozen-response stability, not training or Newton-work ordering.

B.4 No fixed residual norm orders exact Newton work

Below are a certified same-instance witness and an exact single-layer construction in which the proposal with the smaller starting residual needs more exact Newton work.

Same-instance witness. For one fixed problem, a trap proposal has a smaller starting residual than a spike proposal in the ℓ_1, ℓ_2 and ℓ_∞ norms and in the K^{-p} family at $p \in \{-1, 0, 0.5, 1, 1.5, 2\}$, yet needs $8\times$ more outer steps and $30.75\times$ more matrix-vector products. A Hölder bound on the same modal data extends the residual ordering to all $p \in [0, 2]$; its endpoint values are float64, not outward-rounded. A diagonal metric concentrated on the known mislabeled ball reverses the ordering, so it does not hold for every SPD metric. The construction below makes the extra work exact.

Conditional locality. Write $F(h) = h + cK\beta(h) - g$, with $\beta(x) = \min(x, 0) + \max(x - 1, 0)$. On a frozen branch $\beta(v) = Dv + o$, an exact undamped update solves $(I + cKD)h^+ = g - cKo$. If a node and all its neighbors are on the zero-valued plateau, its new value is $h_i^+ = g_i$, which alone does not keep it on the plateau.

Lemma B.2 (Data-conditioned locality)**.** *Let V be the set of grid nodes. Suppose h^0 is strictly in $(0, 1)$ on an interior set $\Omega \subset V$ with nonempty complement, and $g_i \in (0, 1)$ at every $i \in \Omega$ with no neighbor outside Ω. Let $\delta(i) = \mathrm{dist}(i, V \setminus \Omega)$ be the graph distance along grid edges. Under exact undamped frozen-branch Newton, $h_i^k \in (0, 1)$ whenever $\delta(i) > k$.*

Proof. The claim holds at $k = 0$. If $\delta(i) > k + 1$, node i and every neighbor have distance greater than k, so their frozen constitutive values are zero and $h_i^{k+1} = g_i \in (0, 1)$. Induction completes the proof. A neighbor changing branch need not force a flip, so this gives no general upper bound, exact termination count or finite-tolerance lower bound. □

Exact single-layer family. Use the 41×41 five-point Dirichlet grid and Manhattan radius r_i from its center. For $m \in \{2, \ldots, 10\}$ define $\Omega = \{r \le m\}$ and $c = 3/10$, $\rho = 25/3$, $a = 1/5000$,

$A = 1 + 4c = 11/5$, $s_i = a\rho^{r_i - m}$ for $i \in \Omega$. Set $z_i = -s_i$ on Ω, $z_i = 1/2$ elsewhere, and $g = z + cK\beta(z)$.

Proposition B.3 (Exact termination of the specified construction)**.** *Start with any h^0 strictly in $(0, 1)$ on Ω and equal to $1/2$ elsewhere. Use exact arithmetic, exact frozen affine solves and undamped Newton. After update $k = 1, \dots, m+1$, the cold set is exactly $\{m - k + 1 \le r \le m\}$ and all other nodes are strictly on the plateau. The first exact root is reached at update $m + 2$, and equals z.*

Proof. Initially all branches are plateau, so $h^1 = g$, and on Ω, $g_i = -As_i + c\sum_{j\sim i,\, j\in\Omega} s_j$. For $r_i < m$, at least two outward neighbors have magnitude ρs_i, so $0 < (2c\rho - A)s_i \le g_i < 4ca < 1$ with $2c\rho - A = 14/5$. At $r_i = m$, all neighbors within Ω are inward, giving $g_i \le -(A - 4c/\rho)s_i < 0$, since $A - 4c/\rho = 257/125$. Outside Ω, $g_i \in [1/2, 1/2 + 4ca] \subset (0, 1)$. Thus exactly the outer shell becomes cold.

Suppose the current cold set is $C_R = \{R \le r \le m\}$, $1 \le R \le m$, and other nodes are plateau. All frozen affine offsets are zero. The cold-block matrix $B = I + cK_{C_R, C_R}$ is a strictly diagonally dominant M-matrix, so $B^{-1} \ge 0$ and $B^{-1}\mathbf{1} \le \mathbf{1}$. Write $h^+_{C_R} = z_{C_R} + e$. Comparison with g gives $Be = b$ with $b_i = c\sum_{j\sim i,\, j\in\Omega\setminus C_R} s_j \ge 0$. Only the inner boundary contributes. With $s = a\rho^{R-1-m}$, there are at most four contributing neighbors, each of magnitude s, so $0 \le e_i \le 4cs$. Every old cold node has $s_i \ge \rho s$, hence $h^+_i \le -\rho s + 4cs < 0$ because $\rho - 4c = 107/15$.

At a new shell node $r_i = R - 1$, eliminating old cold contributions yields

$$h^+_i = -As + c \sum_{j\sim i,\, j\in\Omega\setminus C_R} s_j + c \sum_{j\sim i,\, j\in C_R} e_j.$$

Unflipped neighbors are one level inward and have magnitude s/ρ (none at the center), so $h^+_i \le (-A + 4c/\rho + 16c^2)s = -\frac{77}{125}s < 0$. Deeper nodes have no cold neighbor, so stay at their data-frozen $g_i \in (0, 1)$. Exterior nodes satisfy $h^+_i = \frac{1}{2} + c\sum_{j\sim i,\, j\in C_R} e_j \in [\frac{1}{2}, \frac{1}{2} + 16c^2 s] \subset [\frac{1}{2}, 195326/390625] \subset (0, 1)$, since $s \le a/\rho$. This proves exactly one inward shell per update, without any hot node.

After $m + 1$ updates every node of Ω is cold. The next frozen system is the true solution branch and has solution z. The preceding update cannot already equal z: its old cold set omits the center node i_0, and substituting z into that frozen system leaves the nonzero vector $-z_{i_0} cK\mathbf{1}_{\{i_0\}}$, where $\mathbf{1}_{\{i_0\}}$ is the coordinate vector of i_0. Earlier iterates have wrong strict branches. The nonlinear root is unique, because for $u \ne v$, $(u - v)^T K^{-1}(F(u) - F(v)) = \|u - v\|^2_{K^{-1}} + c\sum_i (u_i - v_i)(\beta(u_i) - \beta(v_i)) > 0$. □

The parameters are fixed construction data, not fitted to a trajectory. Float64 checks for $m = 2, \dots, 10$ at residual threshold 10^{-13} reproduce the count numerically, without a roundoff certificate. The deployed damped, inexact solver takes 16 steps for depth 8, compared with the exact count $m+2$. For a full time window, the analogous interior-column condition would require $h_{0,i} + \Delta t \sum_{j\le n} s_{j,i}$ to stay strictly on the same plateau for every relevant n; we do not assert this for the multi-layer experiments. The result is an exact count for one layer. Scaling residual amplitudes below a stopping threshold shows that no universal finite-tolerance lower bound can follow from it, and it does not explain the trained savings.

B.5 Residual ordering and exact correction work

Let $P(h) = \mathrm{clip}(h, 0, 1)$, $\beta(h) = h - P(h)$ and K be the positive five-point Dirichlet matrix. For a fixed initial frame and source, write

$$F_n(H) = h_n - h_{n-1} + cK\beta(h_n) - f_n, \qquad h_n - h_{n-1} + cK(h_n - q_n) = f_n.$$

The second equation defines decoder feasibility and gives $F_n(H) = cK(q_n - P(h_n))$. The physical implicit resolvent $b \mapsto (I + cK\beta)^{-1} b$ and the exact frozen-branch Newton update $H \mapsto H - J_D^{-1} F(H)$ are different maps. The latter is constant within an affine branch and can jump across branch boundaries.

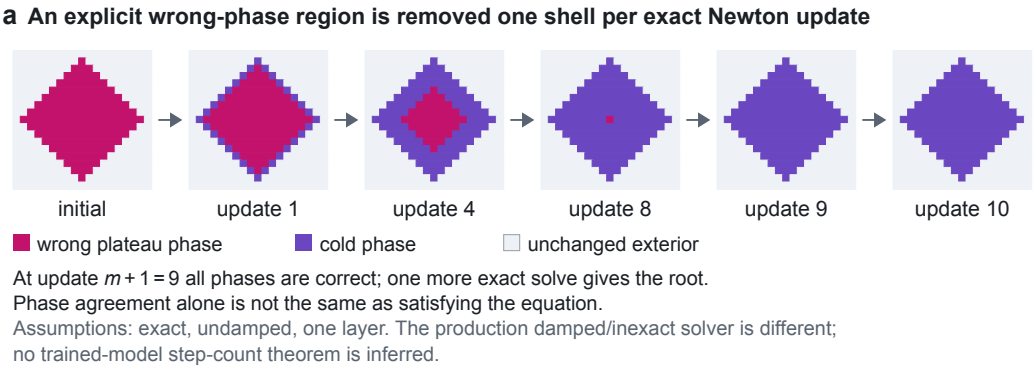

Figure B.1: **An exact branch-evolution example.** The stated one-layer construction at depth $m = 8$, shown on a central crop of its 41×41 grid. The phase classes are prescribed analytically, not learned or measured. All phases agree by update nine, but the exact root is first reached at update ten.

Statement (R1). For every integer $m \geq 2$, finite window length $S \geq 1$ and ratio $\tau > 1$, there are a two-dimensional five-point grid of side $N = 2m + 3$, $c = 3/10$, one fixed h_0, one source sequence, a unique root Z, and two proposals $H^{\mathrm{trap}}, H^{\mathrm{good}}$ with strictly feasible latents $q \in (0, 1)$, such that

$$F(H^{\mathrm{good}}) = -\tau F(H^{\mathrm{trap}}) \neq 0.$$

Consequently, every vector norm ranks the good proposal's initial residual τ times larger, and every fixed SPD quadratic residual loss ranks it τ^2 times larger. The good proposal lies strictly in the root's affine branch and reaches the root in one exact Newton update. Exact undamped full-window Newton first reaches the root from the trap proposal in $m+2$ updates. An explicit positive cumulative ℓ_1 tolerance separates the good step from the first $m + 1$ trap updates. Grid, data and tolerance may vary with m, so the statement is not an asymptotic lower bound at a fixed production tolerance.

Root and source construction. Let r_i be Manhattan distance from the grid center, $\Omega = \{r_i \leq m\}$ and $E = \Omega^c$, so each point of Ω has four grid neighbors. Set $\rho = 25/3$ and

$$s_i = a\rho^{r_i - m}\ (i \in \Omega), \quad s_i = 0\ (i \in E), \qquad z_i = -s_i\ (i \in \Omega), \quad z_i = \tfrac{1}{2}\ (i \in E).$$

Take $h_0 = \frac{1}{2}\mathbf{1}$, $f_1 = z - h_0 - cKs$ and $f_n = -cKs$ for $n \geq 2$. Then $Z = (z, \ldots, z)$ is a root. Its uniqueness follows layer by layer from

$$\begin{aligned}&\langle u - v, (u + cK\beta(u)) - (v + cK\beta(v))\rangle_{K^{-1}} \\ &\quad = \|u - v\|^2_{K^{-1}} + c\sum_i (u_i - v_i)(\beta(u_i) - \beta(v_i)) \geq \|u - v\|^2_{K^{-1}}.\end{aligned}$$

A strictly feasible plateau trap. Write $G = (I + cK)^{-1}$, which is strictly positive entrywise with $G\mathbf{1} \leq \mathbf{1}$. Put $s = a\bar{s}$ and choose

$$b = a\bar{b}, \qquad \bar{b} > \max_{i \in \Omega} \frac{\bar{s}_i}{(G\mathbf{1}_E)_i}.$$

First use $q^{\mathrm{trap}}_n = 0$ on Ω and $q^{\mathrm{trap}}_n = \frac{1}{2} - nb$ on E. Relative to $q^* = P(z)$, its decoder error satisfies

$$e_n = Ge_{n-1} + (I - G)(-nb\mathbf{1}_E) = -nb\mathbf{1}_E + b\sum_{k=1}^{n} G^k\mathbf{1}_E.$$

Thus $h^{\mathrm{trap}}_{n,\Omega} = -s + b\sum_{k=1}^{n} G^k\mathbf{1}_E > 0$ and, for $n \geq 2$, its temporal increment is $bG^n\mathbf{1}_E > 0$. Replace the zero latent on Ω by sufficiently small $\eta > 0$, proportional to a. The decoder perturbation is $\eta(\mathbf{1}_\Omega - G^n\mathbf{1}_\Omega)$, so strict positivity and the temporal increments persist; moreover $h^{\mathrm{trap}}_{n,\Omega} > \eta = q^{\mathrm{trap}}_{n,\Omega}$. After fixing $m, S, \tau, \bar{b}, \eta/a$, choose a small enough that all trap states and latents lie strictly in $(0, 1)$.

The anti-collinear good proposal. Let $D^* = \mathrm{diag}(\mathbf{1}_\Omega)$ and let J^* be the full-window block lower bidiagonal matrix with diagonal blocks $I + cKD^*$ and subdiagonal $-I$. Define $R = F(H^{\mathrm{trap}})$, $w = (J^*)^{-1}R$, $H^{\mathrm{good}} = Z - \tau w$. For $B = I + cK_{\Omega\Omega}$ write $w_{n,\Omega} = s_\Omega + v_n$. The plateau trap gives

$$Bv_1 = h^{\mathrm{trap}}_{1,\Omega} > 0, \qquad Bv_n = v_{n-1} + h^{\mathrm{trap}}_{n,\Omega} - h^{\mathrm{trap}}_{n-1,\Omega} > 0.$$

Because $B^{-1} \geq 0$ with positive diagonal, $v_n > 0$ and $h^{\mathrm{good}}_{n,\Omega} = -(1+\tau)s_\Omega - \tau v_n < 0$. On E, the good state is $\frac{1}{2} + O(a)$ and remains strictly mixed for small a. It therefore occupies the root's branch, giving the exact identity $F(H^{\mathrm{good}}) = J^*(H^{\mathrm{good}} - Z) = -\tau R$. Define $q^{\mathrm{good}}_n = P(h^{\mathrm{good}}_n) - \tau(q^{\mathrm{trap}}_n - h^{\mathrm{trap}}_n)$. On Ω, this is $\tau(h^{\mathrm{trap}} - q^{\mathrm{trap}}) > 0$ of order a; on E it is $\frac{1}{2} + O(a)$. A further reduction of a gives $q^{\mathrm{good}} < 1$, and the residual identity verifies the same linear decoder balance. Finite sigmoid logits can represent these latent values. This establishes latent feasibility, not reachability by the fixed trained network.

Synchronous peeling in every time layer. Choose also $a < 1/(8cS)$. Suppose before a full-window frozen update that every time layer has cold set $C_R = \{R \leq r \leq m\}$, with all remaining sites mixed. Write the updated cold block as $y_{n,C} = z_C + e_n$ and set $B = I + cK_{CC}$. The prescribed first source cancels the initial-state offset and yields

$$Be_1 = b_C, \qquad Be_n = e_{n-1} + b_C, \qquad b_{C,i} = c \sum_{j \sim i,\ j \in \Omega \setminus C} s_j.$$

For $s = a\rho^{R-1-m}$, $b_C \leq c(4 - \deg_C)s$ and $B\mathbf{1} = \mathbf{1} + c(4 - \deg_C)$. Comparison with $s\mathbf{1}$ and $B^{-1} \geq 0$ gives $0 \leq e_n \leq s\mathbf{1}$ for all layers. Old cold points remain negative because $y_{n,i} \leq -\rho s + s < 0$. At a new-shell point $r = R - 1$,

$$y_{1,i} = -s + \Delta_{1,i}, \qquad y_{n,i} = y_{n-1,i} + \Delta_{n,i},$$
$$\Delta_{n,i} = -4cs + c \sum_{\text{unfrozen inner } j} s_j + c \sum_{\text{old cold } j} e_{n,j} \leq 0.$$

All its neighbors lie in Ω, with inner magnitude at most s/ρ and old-cold error at most s. Hence the new shell is negative in every layer. Deeper points have no cold neighbor and satisfy $y_{n,i} = z_i - nc(Ks)_i \in (0, 1)$; their upper bound is below $4cSa < 1/2$. On E, $y_{n,i}$ lies between $1/2$ and $1/2 + 4cSs < 1$. Starting from the all-plateau trap gives $y_n = z - ncKs$, which flips precisely the outer shell. Induction yields cold set $\{m - k + 1 \leq r \leq m\}$ after updates $k = 1, \ldots, m+1$ in every layer. Update $m + 2$ solves the true affine system and first reaches Z.

Let ν_n equal the updated layer y_n on the newly switched shell and zero elsewhere. That shell was frozen as mixed ($\beta = 0$) but now has $\beta(y_n) = y_n$ there, while every other site keeps its frozen branch; hence the post-update nonlinear residual is exactly $F_n = cK\nu_n$. Each newly cold shell has one parity, so no two of its points are neighbors; each of its points has four grid neighbors; and $\nu_n < 0$ on the whole shell, because the new shell is negative in every layer. Thus $(K\nu_n)_i = 4\nu_{n,i}$ on the shell, and every other entry $-\sum_{j \sim i} \nu_{n,j}$ adds terms of one sign, so $\|K\nu_n\|_1 = 4\|\nu_n\|_1 + 4\|\nu_n\|_1 = 8\|\nu_n\|_1$. (The common sign is needed for equality; without it only $\|K\nu_n\|_1 \geq 4\|\nu_n\|_1$ holds.) Since each layer has a new point of magnitude at least $a\rho^{-m}$,

$$\sum_n \|F_n(H^k)\|_1 \geq 8cSa\rho^{-m}, \qquad 1 \leq k \leq m + 1.$$

Thus $\epsilon_W = 4cSa\rho^{-m} > 0$ separates these updates from the exact root. Initial rejection can be ensured by choosing a smaller positive tolerance; the fixed-size certificate below verifies it directly.

Convex damping and the inexact boundary. For an exact frozen candidate y^k and any $\alpha_k \in [0, 1]$, set $h^{k+1} = (1 - \alpha_k)h^k + \alpha_k y^k$. A site and its neighbors that remain mixed have first-layer $y^k_i = g_i$, where $g = h_0 + f_1 = z - cKs$; finite propagation therefore leaves the first-layer center i_0 positive through $k = m$. Write $u = h^k_1$ and $\Phi(x) = x + cK\beta(x)$, so $F_1(h^k) = \Phi(u) - \Phi(z)$. Since the root's center is $z_{i_0} = -a\rho^{-m}$, $u_{i_0} - z_{i_0} > a\rho^{-m}$. The monotonicity identity above and Cauchy–Schwarz in the K^{-1} inner product give $\|F_1(h^k)\|_{K^{-1}} \geq \|u - z\|_{K^{-1}}$. Cauchy–Schwarz

with $K_{i_0 i_0} = 4$ gives $u_{i_0} - z_{i_0} = \langle K^{1/2}\mathbf{1}_{\{i_0\}}, K^{-1/2}(u-z)\rangle \le 2\|u-z\|_{K^{-1}}$, and $\|x\|_1 \ge \|x\|_2 \ge \lambda_{\min}(K)^{1/2}\|x\|_{K^{-1}}$ with $\lambda_{\min}(K) = 8\sin^2\frac{\pi}{2(N+1)} \ge 8/(N+1)^2$. Together,

$$\|F_1(h^k)\|_1 \ge \frac{2\sqrt{2}}{N+1}\cdot\frac{u_{i_0}-z_{i_0}}{2} > \frac{\sqrt{2}a\rho^{-m}}{N+1}, \qquad 0 \le k \le m.$$

At $\epsilon_D = a\rho^{-m}/[2(N+1)]$, at least $m+1$ updates are needed for acceptance; eventual trap convergence is not asserted for arbitrary damping. A standard true-residual Armijo search that tries the unit good step accepts its exact root. With an inexact frozen solve, residual $\xi = F(h) + J_D(y-h)$ instead yields $y_i = g_i + \xi_i$ at deep mixed sites. The extra pointwise condition $|\xi_i| < \min(g_i, 1-g_i)$ is sufficient to preserve the locality argument. A global Krylov forcing condition alone does not imply this condition.

Certificate at the manuscript's grid and tolerance. The specified synthetic witness uses $N = 64$, $S = 50$, $m = 2$, $c = 3/10$, $a = 1/2000$, $b = 1/500$, $\eta = 1/20000000$ and $\tau = 2$. It has $\epsilon_W = 27/62500 = 0.000432 > 0.0004$. The certificate computes exact rational defects of binary floating-point candidates and applies analytic M-matrix enclosures; the decision uses these exact values, not floating-point displays. It verifies strictly feasible good latents, correct good branches, exact anti-collinearity, four exact-inner trap updates versus one good update, and unit-step true-residual Armijo acceptance at $\sigma = 10^{-4}$. The deployed inexact BiCGSTAB work and wall time lie outside the certificate. The main-text comparison at 64^2 (one exact-inner Newton update from the good start versus four from the trap; Section 5.1) comes from this certificate, not from Statement (R1). R1 is an existence statement: for each m it supplies some grid of side $2m+3$ (side 7 at $m = 2$), with data and tolerance that may depend on m, and it fixes no grid or tolerance. The certificate checks the stated 64^2 instance directly.

Table B.2: Outward-rounded intervals from the exact-rational certificate. The first four trap states fail 4×10^{-4} acceptance. Exact rational values of our run are in the supplementary archive; reruns agree to the printed digits.

Trap update k	Certified cumulative residual interval
0	[62.81601926, 62.81601928]
1	[14.50703988, 14.50704012]
2	[0.61709145, 0.61709175]
3	[0.018266629, 0.018266918]
4	exact root

The proof above is not machine-formalized; its constructions are checked with exact rational certificates. It separates the utility of a training objective from a universal ordering of the correction costs of individual proposals.

C Residual geometry and training controls

C.1 Metric effects and matched controls

Controls follow Appendix A.1's recipe (seeds 53/101/202 unless augmented). All tabulated arms accept 32/32 development and 64/64 held-out requests, except Euclidean at 1,000 updates (58 of 64 held-out requests per seed; capped requests keep their capped counts). The same requests are reused across seeds.

Balance defect at the evaluated checkpoints. For the twelve final checkpoints on the 128 training and 64 held-out requests, decoder and residual share the forcing, so b_n is only rounding: $\|b_n\|_2 \le 1.9 \cdot 10^{-13}$ in the float64 evaluation path and $\le 8.3 \cdot 10^{-5}$ over 4,096 nodes in the float32 training path (median $\|b_n\|/\|cKd_n\| \approx 2 \cdot 10^{-5}$). A deliberately inconsistent boundary forcing raises the relative discrepancy to 0.03–0.98; the 16,000-update models were not checked.

Seed augmentation. Seeds 53/101/202 were trained on an RTX 4090 and 303/404/505 on an RTX A6000; both arms of a seed share the device. Before the added seeds' held-out results were read, a bridge run (seed 53 retrained on the A6000, not counted as a seventh seed) fixed the pooling rule: pool if the device difference stays below 7.8 percentage points (6.8 on development). The bridge and original outer effects are -39.7% and -38.9% (differences $-0.8/-1.1$ points outer/matvec), so the seeds are pooled. Original/added/combined reductions are 38.9/34.6/37.0% outer and 41.5/38.2/39.4% matvec (Table C.1). The same test rejects pooling for the one-phase problem (Appendix D.1).

Table C.1: Headline paired-seed work. E/I: Euclidean/Inverse held-out totals; percentages are Inverse-relative-to-Euclidean changes (negative: less work). The bridge run is excluded.

	Held-out totals: E/I		Development (%)		Held-out (%)	
Seed	Outer	Active matvec	Outer	Matvec	Outer	Matvec
53	682/417	4360/2550	−34.1	−37.1	−38.9	−41.5
101	701/409	4514/2475	−38.9	−43.7	−41.7	−45.2
202	686/432	4414/2655	−32.9	−36.6	−37.0	−39.9
303	626/420	3963/2586	−29.2	−32.0	−32.9	−34.7
404	647/423	4149/2564	−32.5	−37.1	−34.6	−38.2
505	672/423	4274/2611	−34.3	−39.3	−37.1	−38.9

Retrospective resampling (10,000 draws of 16 requests per family, indices shared across seeds, arms and metrics). Conditional on the six fitted model pairs, 95% request-bootstrap intervals are $[-42.1, -31.6]\%$ (outer) and $[-45.4, -33.3]\%$ (active matvec). Also resampling model pairs within device strata gives *sensitivity ranges* $[-42.3, -31.2]\%$ and $[-45.7, -32.9]\%$, not confidence intervals over seeds or hardware (small strata, device and cohort confounded).

Scale and diagonal controls. $W_{\mathrm{tr}} = [\mathrm{tr}(K^{-1})/4096]I = .674717I$; $W_{\mathrm{loss}} = sI$ with $s = 1.506749$ matches seed 53's first 64 training-batch initial losses and was fixed before the nine new models were trained; $W_{\mathrm{diag}} = \mathrm{diag}(K^{-1})$, not $(\mathrm{diag}\, K)^{-1}$. Every route accepts all requests. Held-out outer totals (seeds 53/101/202) are 680/658/681 (W_{tr}), 693/685/689 (W_{loss}) and 668/687/654 (W_{diag}), against Euclidean and Inverse in Table C.1; inner, matvec and backtracking totals are in the supplementary details. Inverse uses less work than all three controls in every counter, seed and set (development outer reductions 32.3–35.5/32.5–36.6/32.8–37.6%). W_{tr} reduces held-out outer work by 0.3–6.1% relative to Euclidean; W_{loss} versus Euclidean and W_{diag} versus W_{tr} have mixed signs.

Reference-error matching. Evaluating each request of the separate 64-request accuracy study at eight tolerances from $4 \cdot 10^{-2}$ to $4 \cdot 10^{-6}$ and charging $c(\tau) = \min\{\mathrm{outer}(\varepsilon) : E_\infty(\varepsilon) \le \tau\}$ (selection with the help of the reference, not a stopping rule), Inverse changes accuracy-matched outer work at $\tau = 10^{-4}$ by -32.9 to -39.1% per seed on the diffusion problem (totals 412–425 against 623–676) and matvecs by -37.2 to -44.9%; on the 32^2 reaction problem it needs $+10.2$ to $+16.4\%$ more outer work. No request is censored and no selected step reaches a cap. At the residual tolerance $\varepsilon = 4 \cdot 10^{-4}$ the same six seeds change outer work by -36.32, -38.26, -35.27, -32.59, -32.34 and -35.20% (median -35.24%, the matched-residual row of Table 1). For $\tau = 10^{-3}, 10^{-4}, 10^{-5}$ the diffusion outer medians are $-36.4/-35.7/-35.1\%$; divided by this median, these six-seed accuracy-matched medians give effect ratios 1.034/1.013/.995; reaction medians are $+12.6/+13.5/+9.9\%$ (minimum $+7.1\%$). Matching field error therefore leaves both preferences in place; per-seed values are in the supplementary details.

Longer training, V-cycle metric and learning rate. Both arms trained for 16,000 updates (full cosine span, identical first 4,000 minibatches) change outer/matvec work by $-18.1/-20.7\%$ on development (held-out: Table C.2); held-out outer totals are 443/438/464 (Euclidean) and 361/355/351 (Inverse). A preregistered arm replaces K^{-1} in the Inverse loss by W_V, one symmetric three-level geometric multigrid V-cycle (64^2, 32^2, 16^2; Galerkin coarse operators, two weighted-Jacobi sweeps with $\omega = 0.8$ before and after, direct coarsest solve). W_V is SPD, the spectrum of $W_V K$ lies in $[0.871, 1.000]$, and on 100 random residuals $\|W_V r - K^{-1} r\|_2 / \|K^{-1} r\|_2$ has median 0.4% and maximum 0.7%. On held-out requests W_V changes outer work by -38.9% (seeds -38.9, -42.4,

-37.6%) and matvecs by -41.5%, the same as the exact inverse; development outer is -35.3% against -34.1%. All requests are accepted, and the preregistered criterion (at least three quarters of the exact-inverse saving, every seed favoring W_V over Euclidean) is met. At learning rates $3 \cdot 10^{-3}/10^{-3}/3 \cdot 10^{-4}$, Inverse's development medians are 194/211/240 (held-out 386/417/457), so the rule would select $3 \cdot 10^{-3}$; for Euclidean 10^{-3} is best (322 versus 544 and 626). The headline keeps the shared rate 10^{-3}.

Training-budget sweep and initial residual size. A preregistered sweep trains both arms for complete cosine schedules of 1,000 to 32,000 updates (seeds 53/101/202; Table C.2). All three seeds favor Inverse in both counters at every budget, but the reduction narrows: the 32,000/16,000 ratio of reductions is 0.65 (outer) and 0.71 (matvec), below the preregistered stabilization threshold 0.75. A second preregistered extension trains the same recipe for 64,000 and 128,000 updates under a rule fixed before training: the stabilization criterion is met if the ratio r of the 128,000- to the 64,000-update reduction is at least 0.75 in both counters and Inverse needs fewer outer iterations and matvecs at every seed at 128,000 updates. Held-out r is 0.88 (outer) and 0.78 (matvec), Inverse is cheaper at every seed in both counters (outer totals, Euclidean/Inverse, 358/323, 359/332 and 348/324 for seeds 53/101/202), and the reduction persists at 7.5% at 128,000 updates; development requests give the same reading (r 0.80 and 0.87, reduction 5.65%). From 32,000 to 64,000 updates the ratios are 0.70 and 0.66, below 0.75, so the criterion is met over the last doubling only, and the matvec ratio clears the threshold by 0.03. The 8,000-update pairs and the 64,000-update seed-101 pair were trained on the RTX A6000, the other extension runs on the RTX 4090, and all were evaluated on the RTX 4090.

Two preregistered readings of the same records compare residual size with outer work (Table C.3): the ℓ^1 residual $R_0 = \sum_n \|r_n\|_1$ and $S_p = \sum_n (r_n^\mathsf{T} K^{-p} r_n)^{1/2}$, $p = 0, 1, 2$, recomputed for all 36 checkpoints. Within a checkpoint S_1 tracks work more closely than R_0 (median Spearman correlation 0.62 against 0.53).

Retrospective ordering across checkpoints. Its written plan predates the computation but follows the 64,000-update readings, in which R_0 misordered two of the three 64,000-versus-32,000 pairs and S_1 and S_2 ordered all three. It asks which start statistic follows held-out Newton work across all 48 checkpoints of the sweep (two losses, eight budgets from 1,000 to 128,000 updates, seeds 53/101/202): outer work is each checkpoint's held-out total over the 64 requests, R_0, S_0, S_1 and S_2 are medians over the same requests, and the wrong-phase fraction of the starts, defined as in Section 5.2 against the same-seed accepted state of the 128,000-update Inverse checkpoint, is pooled over all nodes, steps and requests. With b_n at rounding level, $S_2 = \sum_n \|K^{-1} r_n\|_2 \approx c \sum_n \|d_n\|_2$ measures the constitutive defect in the Euclidean norm. Table C.4 reports (A) Spearman correlations over the 48 checkpoints; (B) among the 24 same-budget, same-seed Euclidean–Inverse pairs, those ordered, i.e. in which the checkpoint with the smaller statistic needs strictly fewer outer iterations (no ties occur); and (C) the same count over all 84 same-seed pairs of a Euclidean checkpoint and an Inverse checkpoint trained for fewer updates. The wrong-phase fraction follows work most closely, with Spearman 0.993 over the 48 checkpoints and 24 of 24 same-budget and 79 of 84 cross-budget pairs ordered, then S_2 and S_1; R_0 orders every same-budget pair but only 48 of the 84 cross-budget pairs. The plan expected S_1, S_2 and a wrong-phase fraction to follow work more closely than R_0 in all three readings, and all three do. The wrong-phase column is a retrospective corrective analysis whose definition was fixed before its computation: the written plan had defined the fraction as the share of nodes at which the network's latent liquid fraction $\widehat{q}$ and the phase $P(\widehat{h})$ of the decoded state lie on different branches, with $\widehat{q}$ counted as solid or liquid only where it equals 0 or 1 exactly; because $\widehat{q}$ is a float32 network output, that count tracked the share of exactly saturated outputs (Spearman -0.95) rather than wrong phases, and it is not reported. Using each checkpoint's own accepted state as reference, or excluding reference nodes within 10^{-3} of a kink, changes the three readings to at most 0.992, 24 and 77.

C.2 Isospectral and earlier solver controls

Sign-masked (sign-conjugate) weights $\Sigma K^{-1} \Sigma$ with diagonal $\Sigma^2 = I$ keep the spectrum and diagonal of K^{-1} and lose the saving. Against matched Inverse, M2 (six valid seeds) needs a median 63.3% more outer work and M3 (three seeds) 63.1%; the original three M2 seeds give 66.2/69.7/55.3% and change outer work against Euclidean by $+1.6/-1.0/-2.2\%$ (three seeds are not an equivalence

Table C.2: Held-out work change of Inverse relative to Euclidean by training budget (%; medians of the three per-seed changes, not ratios of the median totals; all seeds favor Inverse at every budget) and median outer totals. The 8,000-update pairs and the 64,000-update seed-101 pair were trained on the RTX A6000; all runs were evaluated on the RTX 4090. †Euclidean accepts 58 of 64 requests per seed at 1,000 updates; capped counts are retained.

	Median outer total		Paired change (%)	
Updates	Euclidean	Inverse	Outer	Matvec
1,000	1,184†	523	−55.0	−58.2
2,000	942	453	−52.5	−56.9
4,000	686	417	−38.9	−41.5
8,000	524	388	−25.6	−26.7
16,000	443	355	−18.9	−19.6
32,000	391	344	−12.3	−13.9
64,000	368	341	−8.6	−9.2
128,000	358	324	−7.5	−7.2

Table C.3: Smaller initial residuals with more outer work (held-out starts; per-seed ranges over seeds 53/101/202; same-device pairs unless marked). The last column is Euclidean's outer work relative to the Inverse checkpoint; request counts are over 192 request–seed pairs.

Pair (updates, Euc. vs. Inv.)	Measure	Euclidean	Inverse	Euc. outer
16,000 vs. 4,000	median R_0 ratio	0.55–0.60		+6.2 to +7.4%
	median ℓ^1 residual	340–348	563–624	
	median S_0	14.85	24.0–26.4	
	median S_1	13.2–13.8	15.9–17.2	
	outer totals	438–464	409–432	
	smaller residual, more work	67 (R_0), 68 (S_0), 49 (S_1) of 192		
	S_2	ranks the pair correctly		
32,000 vs. 16,000	median R_0 ratio	0.86–0.97		+6.4 to +16.5%
32,000 vs. 4,000	median R_0 ratio	0.38–0.44		−4.4 to −7.9%
8,000 vs. 4,000 (A6000)	preregistered branch	first budget with smaller R_0		+21.3 to +29.0%

Table C.4: Retrospective: which start statistic orders held-out outer work across the 48 budget-sweep checkpoints. Wrong-phase (ref.) is measured against the same-seed accepted reference; statistics and readings (A)–(C) are defined in the text.

Statistic	R_0	S_0	S_1	S_2	Wrong-phase (ref.)
(A) Spearman correlation (48 checkpoints)	0.923	0.921	0.979	0.987	0.993
(B) Same-budget pairs ordered (of 24)	24	23	24	24	24
(C) Cross-budget pairs ordered (of 84)	48	51	70	73	79

test). M1 is not reported because three of its six training runs are invalid. The masks also alter low-frequency preference, so the contrast does not attribute the gain uniquely to spatial coupling; the checkerboard weight $(8I - K)^{-1}$ shares K's eigenbasis, so not every sign mask removes it. All nine M2/M3 checkpoints and their Inverse partners accept every request; the per-seed changes are +55.3 to +69.7% for M2 and +62.5 to +68.7% for M3.

Earlier controls (not rerun; supplementary details): other Newton merit functions keep Inverse's 64^2 advantage (outer medians −17.8 to −47.8%); supervised BT fine-tuning and online refinement (six of 12 configurations pass) show no uniform improvement; a boundary-driven Stefan control changes work only slightly; and a conjectured universal one-step tail fails (612 of 715 paths, 85.6%).

D CONSTITUTIVE LAWS, GRID AND BACKBONE SCOPE

D.1 MODEL PROBLEM EQUATIONS AND COMPLETE OBJECTIVE COMPARISONS

Exact equations, constitutive laws and decoders. All model problems use K and c of Section 3, the forcing $F_n = \Delta t\, s_n + c g_\partial$, and (outside the main model problem) $N = 32$, $S = 50$, $\Delta t = 0.02$; $x_+ = \max(x, 0)$. The laws of Table D.1 enter the diffusion term. For each placement, the table below fixes the residual, the linear decoder (head output q_n) and the defect identity, with b_n the computed balance defect; moving a law between placements changes the equation and decoder, not only the metric.

Placement	Residual r_n	Linear decoder	Defect identity
Diffusion (D)	$h_n - h_{n-1} + cK\beta(h_n) - F_n$	$(I+cK)e_n = e_{n-1} + q_{n-1} - q_n + F_n$; $e_0 = \beta(h_0)$, fixed $q_0 = P(h_0)$, $P(h) = h - \beta(h)$, $\widehat{h}_n = e_n + q_n$	$r_n - b_n = cKd_n$, $d_n = q_n - P(\widehat{h}_n)$
Reaction (R)	$u_n - u_{n-1} + cKu_n + a\beta(u_n) - F_n$; $a = 4c$, the diagonal of cK	$((1+a)I + cK)e_n = e_{n-1} + q_{n-1} - q_n - cKq_n + F_n$; $e_0 = \beta(u_0)$, $\widehat{u}_n = e_n + q_n$	$r_n - b_n = ad_n$, not cKd_n
Obstacle	$y_n + cK\gamma(y_n) - \gamma(y_{n-1}) - F_n$, $y = u + q$	$(I + cK)\widehat{u}_n = \widehat{u}_{n-1} - q_n + F_n$; $\widehat{u}_0 = u_0 \geq \psi$, $q_0 = 0$	$r_n - b_n = cKd_n - d_{n-1}$, $d_n = \gamma(\widehat{u}_n + q_n) - \widehat{u}_n$, $d_0 = 0$
Mixed	constitutive term weighted by $\theta cK + 4(1-\theta)cI$, linear diffusion by $(1-\theta)cK$	matching decoder	$4cd$ at $\theta = 0$ (reaction), cKd at $\theta = 1$ (diffusion)

Table D.1: Diffusion-placement laws. Boundary temperature $\beta(h_\partial) = -0.2$, except Richards (0.5). Ranges include limiting values; q is not a bounded phase fraction for asymmetric B or Richards.

Model problem	$\beta(h)$	Slopes; kinks	Range of P; head
Two-phase	$\min(h, 0) + (h-1)_+$	1, 0, 1; 0, 1	$[0, 1]$; sigmoid
Eutectic A	$\min(h, 0) + \mathrm{clip}(h-1, 0, 0.5) + (h-2.5)_+$	1, 0, 1, 0, 1; 0, 1, 1.5, 2.5	$[0, 2]$; 2 sigmoid
Asymmetric B	$0.5\min(h, 0) + 2(h-1)_+$	0.5, 0, 2; 0, 1	$(-\infty, 1]$; 1−softplus
One-phase	$\min(h, 0)$	1, 0; 0	$[0, \infty)$; softplus
Richards-type	$5\min(h, 0) + \mathrm{clip}(h, 0, 1) + 5(h-1)_+$	5, 1, 5; 0, 1	$\mathbb{R}$; identity

The decoder does not need P monotone. Model problems with non-unit slopes use a generalized fixed-phase corrector, which never differs between Euclidean and Inverse within one problem. The Richards-type problem (total storage, piecewise-affine pressure-head law, constant conductivity, no gravity) is a regularized model, not a full soil model.

The reaction model problems place the two-phase or eutectic-A law in the reaction term ($u_\partial = -0.2$). In the parabolic obstacle, u is constrained by a supplied, time-independent obstacle ψ (a network input), with the nonpositive scaled multiplier $q = -\Delta t\lambda$ as latent variable: $(I + cK)u_n - u_{n-1} + q_n - F_n = 0$, $u_n \geq \psi$, $q_n \leq 0$, $q_n(u_n - \psi) = 0$; $\gamma(y) = \max(y, \psi)$, $P_\psi(y) = \min(y - \psi, 0)$, the head is $-\mathrm{softplus}(z_n)$ and $u_\partial = -0.2$. The lagged term distinguishes this model problem from (D); its corrector is contact-restricted.

Registered 64^2 obstacle comparison. Registered before training, this comparison predicts that Euclidean needs less work (obstacle compatibility classes for the physical-state step: Table B.1 and App. B.3). It uses 64^2 ($\Delta x = 1/65$) with all other parameters as at 32^2, new data (128 training, 32 development and 64 held-out requests, each reference passing a KKT certificate), seeds 53/101/202/303/404/505, 4,000 updates with the same recipe for both arms, and $\varepsilon = 4 \cdot 10^{-4}$, scaled per node as for the main problem. Inverse changes held-out outer work by a median $+22.9\%$ (seeds $+19.8$ to $+25.9\%$), with all six seeds in the same direction on both splits (development and matvec medians: Table D.2). Held-out outer totals, Euclidean/Inverse, are 396/484, 400/490, 396/488, 405/485, 398/491 and 397/500 (development 204/271, 202/263,

200/271, 201/270, 201/273, 202/278). The directional preference is substantially larger in the registered 64^2 comparison than in the earlier 32^2 test; the median contact fraction of the 64^2 data is about 8%, against 12.9% on the 32^2 development set.

Mixed placement. With $P(u) = \mathrm{clip}(u, 0, 1)$ and $\beta = u - P(u)$, residual and decoder both change along θ under the weighting of Proposition B.1 (equations and request settings in the supplementary details). It uses $N = 64$ and $c \in [0.4225, 1.69]$ with residual/reference tolerances $4\times10^{-4}/4\times10^{-7}$; the reaction panel uses $N = 32$, $c \in [0.1089, 0.4356]$ and tolerances $10^{-4}/10^{-7}$ on different requests.

Objectives and complete comparison. Euclidean, Inverse and BT use each model problem's own residual and defect; on reaction and obstacle model problems BT is transferred without a proof. Table D.2 gives $100(W_X - W_{\mathrm{Euc}})/W_{\mathrm{Euc}}$ per model problem and request set, computed per training seed from all-request work totals (negative: less work than Euclidean). Three-seed rows use 53/101/202; six-seed rows add 303/404/505 on the same requests. Every arm accepts every request at every seed. The "new held-out" panels evaluate the fixed models on 64 fresh requests generated after the evaluation was frozen.

Table D.2: Work relative to Euclidean (%), seed median. D: development (32), H: held-out (64); cells give the D/H changes. A dash marks a comparison not reported here (the 64^2 BT single-window comparison is in Appendix G.2, Table G.1). All seeds share the direction except where marked: two-phase 32^2 D*, five of six seeds lower in each counter; one-phase H*, two of three.

Model problem (seeds; sets)	Inverse outer	Inverse matvec	BT outer	BT matvec
Two-phase 64^2 (3; D/H)	-34.1 / -38.9	-37.1 / -41.5	– / –	– / –
Two-phase 32^2 (6; D*/H)	-2.4 / -5.5	-3.3 / -8.4	– / –	– / –
Eutectic A (3; D/H)	-16.1 / -20.65	-19.95 / -24.8	-15.75 / -19.8	-22.05 / -23.7
Asymmetric B (3; D/H)	-17.7 / -22.8	-19.6 / -23.9	-17.4 / -22.9	-19.2 / -23.9
Richards-type (3; D/H)	-50.95 / -48.8	-50.6 / -48.6	-51.6 / -51.8	-52.0 / -52.2
One-phase (3; D/H*)	-2.2 / -2.5	-2.35 / -1.6	-4.5 / -3.3	-3.4 / -3.8
Two-phase reaction (6; D/H)	$+9.8$ / $+11.9$	$+9.5$ / $+12.9$	$+123.7$ / $+98.4$	$+131.9$ / $+105.3$
Eutectic A reaction (6; D/new H)	$+7.1$ / $+7.4$	$+7.2$ / $+6.0$	$+100.7$ / –	$+115.2$ / –
Obstacle (3; D/new H)	$+2.5$ / $+3.4$	$+6.2$ / $+5.7$	$+34.5$ / –	$+30.75$ / –
Obstacle 64^2 (6; D/H)	$+34.9$ / $+22.9$	$+38.1$ / $+25.5$	– / –	– / –

Reading the table. The registered reaction comparison's 10.6% (denominator W_{Inv}) re-expresses the $+11.9\%$ cell. Backtracking can move the other way (new eutectic-A reaction panel -50.0%; new obstacle panel $+11.5\%$), so fewer backtracks alone do not mean less correction work. One-phase seeds 303/404/505, trained later on another device, favor Euclidean in development outer work ($+1.5\%$ at each seed) and are mixed on held-out outer work ($-1.2/+0.4/-1.2\%$), with held-out matvec favoring Inverse at all three; three of four device-bridge checks reject pooling, so they are reported rather than pooled. Neither the compatible metric nor operator energy therefore ranks the objectives universally.

D.2 Grid transfer and architecture sensitivity

Zero-retraining transfer and classical routes. Frozen 64^2 models are evaluated at 128^2 on 16 fresh requests; Pq prolongs the coarse latent q and decodes it on the fine grid, direct runs the network on the fine grid, and tolerance $1.6 \cdot 10^{-3}$ preserves the coarse per-node budget, not field accuracy (complete route table in the supplementary details). Two classical routes see the same requests and acceptance test: Classical + correction solves each request with per-step FISTA on the 64^2 grid, prolongs h to 128^2 with the exact h_0 and hands it to the same fine-grid Newton, and fine FISTA solves on the 128^2 grid directly. Table D.3 keeps both sides of the classical result: the coarse-grid

route needs fewer fine-grid Newton steps than Inverse/Pq, and its online time is longer; these online seconds describe costs rather than speedups. Maximum field errors over all finite outputs are at most 1.24×10^{-4} for Inverse/Pq and the supervised baselines DS-Q and DS-H (Appendix G.2), versus 1.06–1.07 for Euclidean/direct and 0.733–0.803 for Euclidean/Pq.

Table D.3: Zero-retraining transfer of frozen 64^2 models to 128^2 and two classical routes on the same 16 fresh requests (seed order 53/101/202 where three values are listed). Work is the mean per request over all attempts, capped ones included; online seconds are one pass on a host that was not quiet, coarse solve included for the classical route; field errors are given in the text. The lower block divides Inverse/Pq work by the comparator's work on the n requests both accept.

Route	Accepted of 16	Newton outer	Active matvec	Online (s)
Inverse/Pq	16/16/16	16.500/17.250/17.000	152.6/155.5/156.6	0.527/0.530/0.551
Euclidean/Pq	13/12/11	–	–	–
Euclidean/direct	9/8/9	–	–	1.036/1.002/1.011
DS-Q, DS-H	16/16/16	–	–	–
Classical + correction	16/16	10.562	93.8	1.045
Fine FISTA at 128^2	16/16	–	–	1.640
Comparator	**Jointly accepted n**	**Outer ratio**	**Matvec ratio**	
Euclidean/Pq	13/12/11	0.654/0.664/0.703	0.611/0.600/0.678	
Full NSD3	16/15/16	0.895/0.894/0.961	0.946/0.911/0.989	
NSD3W	14/15/14	0.734/0.755/0.743	0.740/0.746/0.743	

Grids trained from scratch and the diffusion-number control. The 48^2 models are evaluated at their training resolution on 64 new requests ($\varepsilon = 2.25 \cdot 10^{-4}$) against the registered criterion that both counters are lower at every seed. A lower-diffusion 64^2 control with the 32^2 diffusion-number range shows that the grid profile follows the diffusion number, not resolution alone (Table D.4).

Table D.4: Grid and backbone scope, Inverse versus Euclidean unless stated; per-seed values in seed order 53/101/202, then 303/404/505 where six are listed. Ratios are Inverse/comparator totals on jointly accepted requests (n).

Setting	Measured	Outcome
48^2, trained at 48^2	Outer ratios 0.731/0.680/0.697, matvec 0.707/0.632/0.657; all six routes 64/64; median reductions 30.3% outer, 34.3% matvec	Criterion met
Grid profile	Held-out outer reduction 4.9/30.3/38.9% at $32^2/48^2/64^2$; 10.3% at lower-diffusion 64^2	Follows diffusion number
128^2, continued from 64^2	Accepted held-out 64/64/64 vs. 36/40/39; development 32/32/32 vs. 19/21/22	Inverse accepts all
128^2 from scratch (preregistered)	Accepted development 32/32/32/32/32/31 vs. 8, held-out 63/64/64/64/64/64 vs. 16; on the 16 jointly accepted held-out requests outer 122/121/131/119/120/122 vs. 364, matvec 853/850/954/836/882/871 vs. 2,955	Euclidean collapses (six seeds)
256^2 from scratch (preregistered)	Accepted held-out 21/17/25 vs. 2/2/2, development 8/7/8 vs. 0/0/0; on the two jointly accepted held-out requests per seed, outer 10/16/17 vs. 34	Neither arm reaches 60 of 64; Euclidean collapses
256^2, diffusion number of 64^2 (preregistered)	Accepted held-out 63/64/63 vs. 62, development 32 vs. 32 at every seed; jointly accepted held-out outer 388/393/385 vs. 900 ($-56.9/-56.3/-57.2\%$), matvec 2,620/2,780/2,623 vs. 6,315, development outer 186/188/189 vs. 415	Inverse cheaper at every seed; Euclidean collapses
WindowFNO, 64 independent requests	Reductions by $p = 1$ over $p = 0$: diffusion outer 44.9/39.6/40.6%, matvec 48.0/43.8/44.2% (including one capped failure); reaction outer 0.7/7.0/2.7%, matvec $-1.5/5.2/-1.0\%$	Diffusion kept; reaction not reproduced

An earlier campaign trained Euclidean from scratch at 128^2 with learning rates 10^{-3}, $3 \cdot 10^{-4}$ and 10^{-4} (three seeds each, recipe otherwise unchanged), and every run collapsed to a near-constant proposal. A preregistered comparison then trained both arms from scratch with the headline recipe at seeds 53/101/202, and a later registration added seeds 303/404/505 with the same collapse criterion and an outcome rule. All six Euclidean runs collapse by the preregistered criterion (the same proposal across requests; gradients below 10^{-5} from update 30–41 on at the first three seeds and by update 1,000 at the others). At every new seed Inverse accepts at least 60 and Euclidean at most 32 of the 64 held-out requests, meeting the registered outcome rule (Table D.4); one Inverse held-out request at seed 53 and one development request at seed 505 reach the 40-iteration cap.

A further registration repeats the from-scratch comparison at 256^2 as a stress test of the same pipeline at higher resolution. The physical parameters stay those of the 64^2 problem, so the diffusion number is $(257/65)^2 \approx 16$ times the 64^2 range, and $\varepsilon = 6.4 \cdot 10^{-3}$ keeps the per-node budget. Both arms train at seeds 53/101/202 on the RTX 4090 for 4,000 updates with the headline recipe, on new certified data from the same generator (128 training, 32 development and 64 held-out requests). Memory forces batch 2 with four accumulation steps for both arms, fixed before any comparison was read. Four outcomes were fixed in advance: (a) Inverse accepts at least 60 and Euclidean at most 32 of the 64 held-out requests at every seed; (b) both arms accept at least 60 at every seed; (c) some Inverse seed accepts fewer than 60; (d) any other pattern. The result is (c). Its preregistered sentence reads "from scratch at 256^2 neither arm reaches acceptance within 4,000 updates", which by the registration means that neither arm reaches 60 of 64 (acceptance in Table D.4). All three Euclidean runs collapse to one and the same proposal. The sigmoid head of $\widehat{q}$ saturates (fraction 1.0), proposals differ across requests by less than 10^{-23} and gradients stay below 10^{-5} from update 1,000 on; the recorded losses at updates 1,000 to 4,000 agree across seeds in every digit. No Inverse run collapses (saturated fraction 0.00/0.21/0.17). In the training log, the median residual of the development starts stays at 42,987 for Euclidean from update 1,000 on, whereas for Inverse it falls to 31,674–34,142 by update 4,000. At every seed Inverse is accepted on more requests than Euclidean and needs less outer work on the jointly accepted ones (Table D.4). With two such requests per seed, this work comparison is descriptive only; no development request is accepted by both arms.

A last registration keeps the 256^2 grid but multiplies the diffusivity by $(65/257)^2$, so that the diffusion number returns to the 64^2 range. The data are new, recipe, batch and evaluation are as above, and both arms of a seed run on one GPU model (seed 53 on the RTX A6000, seeds 101 and 202 on the RTX 4090). Euclidean again collapses at every seed (saturated fraction 1.0), but at this diffusion number the corrector accepts its constant proposal for 62 of the 64 held-out requests. Inverse does not collapse (saturated fraction 0.25/0.26/0.26). As registered, the direction is decided on the requests accepted by both arms; it holds in all twelve signs (three seeds, two counters, held-out and development requests; Table D.4). Magnitudes are reported per seed and not pooled across GPU models; Table C.1 gives the same three seeds at 64^2, where every request is accepted.

Second backbone. The second backbone, a WindowFNO, also changes time sharing and adds a cumulative-source feature, so it does not isolate architecture. Twelve $p = 0/1$ endpoint models (three seeds) are evaluated on 64 independent requests; 767 of 768 are accepted, and the $p = 0$ failures (one development request per seed, one independent request at seed 53) are solved by $p = 1$. The registered interaction test requires every request to be accepted, so the interaction is left unclassified; the reaction preference is not reproduced (the required opposite preference is -5%).

E Prospective settings, exponent choice and mixed placement

E.1 Prospective protocol and seed augmentation

Registration index. Table E.2 indexes the registered comparisons (Section 3) behind main-text claims; unmet rules are reported like met ones. The other registered items (the zero-retraining transfer and 48^2 comparisons, the V-cycle, the exponent, mixed-placement and second-backbone tests, the corrector variants and the timing sessions) state their rule where reported. Retrospective analyses are labeled as such, and the design benchmark is marked as reused (Section 6.1).

Placement tests and the amended entry test. Before data generation, the law–placement comparisons of Table E.1 registered the compatible metric, data seeds, recipe, controls and verdict rule

Table E.1: **Original three-seed prospective tests and the inverse control.** The first five rows compare Inverse with Euclidean on 64 new requests, three seeds (53/101/202), at 32^2.

Operator / law	Placement	Predict	Outer change (%)	Outcome
K_0: Asymmetric $\frac{1}{2}$/0/2	Reaction	Euclidean	+12.0/+15.0/+10.9	Supported
K_0: Slopes 1/0/2	Diffusion	Inverse	-32.4/-31.6/-24.0	Supported
K_0: Slopes 1/0/3	Diffusion	Inverse	-49.9/-48.7/-52.1	No verdict[a]
K_a: Two-phase	Diffusion	Inverse	-6.4/-10.2/-8.2	Supported
K_a: Two-phase	Reaction	Euclidean	+15.7/+17.5/+12.1	Supported
K_a: inverse control	Diffusion	Heuristic	$-0.7/+2.0/+0.3$	Mixed[b]

[a]Euclidean accepts 63/64 at each seed and Inverse 64/64; the entries are attempted-work differences, not completed-solve savings. [b]Separate heuristic: $100(W_{K_0^{-1}} - W_{K_a^{-1}})/W_{K_a^{-1}}$; all arms accept 64/64.

Table E.2: Registration index. Provenance: PN, prospective on new data; RE, registered extension on reused requests; RT, retrospective (posed after related results were known); AM, amended protocol. The rule and the outcome are stated once, in the subsection of the last column.

Prov.	Comparison	Rule fixed before	Reported in
PN	Parabolic obstacle, 64^2, six seeds	training	§4.2; D.1
PN	128^2 from scratch, seeds 53/101/202, then 303/404/505	the runs; second registration with the same criteria	§4.3; D.2
PN	256^2 from scratch, three seeds	the runs	§4.3; D.2
PN	256^2, diffusion number of the 64^2 range	the runs	D.2
PN	Same starts under Newton and Picard	the runs	§5.3; F.1
PN	Targeted repair against displacement control	the runs	§5.2; F.1
PN	Classical starts and solver forms, third and design benchmarks	any reading; routes frozen before the third benchmark was generated	§6.2; H.3
RE	Matched solution labels in both losses	the runs	§4.1; G.2
RE	Budget sweep, 1,000–32,000 updates, three seeds	the runs	§4.1; C.1
RE	Budget extension, 64,000 and 128,000 updates	training	§4.1; C.1
RE	Residual size against outer work, 36 sweep checkpoints	the readings	§5.1; C.1, Table C.3
RT	Start statistics against outer work, 48 sweep checkpoints	the computation; plan written after the 64,000-update readings	§5.1; C.1
RT	Wrong-phase fraction against an accepted reference, 48 sweep checkpoints	the computation (its definition); a corrective analysis	C.1, Table C.4
AM	Placement tests: five laws or operators, diffusion or reaction placement, 32^2	data generation; entry test amended before any new diffusion law was trained	§4.2; E.1, Table E.1

(held-out median plus all three seed directions: support; plus two: partial support; otherwise failure), with seeds 53/101/202, 32^2, 4,000 updates, 128/32/64 requests and tolerance 10^{-4}. The protocol was amended twice before any new diffusion law was trained, replacing an uncalibrated entry test by a censoring guard that allows no unaccepted request (history in the supplementary details); slopes 1/0/3 therefore stays without a verdict (Table E.1, note a). The two decided laws have held-out matvec medians $+14.75/-32.2\%$ and development outer medians $+9.5/-44.4\%$; reaction backtracking instead favors Inverse (0 versus 1–2). The later six-seed held-out outer medians of the four registered 32^2 rows of Table 3 reuse the original requests.

Variable-coefficient operator. Face-midpoint discretization of $K_a = -\nabla \cdot (a\nabla)$ uses $a = 1 + 0.8 \sin 2\pi x \sin 2\pi y$ (face range $[0.20, 1.80]$, contrast 8.91; best scalar fit to K_0 with residual 0.381), with dense float64 eigendecomposition in place of DST inverses; all fifteen routes accept every request. The mismatched-inverse row of Table E.1 trains with K_0^{-1}: against K_a^{-1} it changes held-out outer work by a median $+0.3\%$ (two of three seeds in the predicted direction), and against Euclidean by -7.9%. Since $K_0^{-1}K_a$ has spectrum $[0.205, 1.795]$, here, as with the V-cycle (Appendix C.1), a metric spectrally equivalent to K^{-1} keeps the saving without the exact inverse.

E.2 EXPONENT CHOICE AND MIXED PLACEMENT

Finite exponent comparison. Each arm minimizes $M^{-1}\sum_{i,n} r_{i,n}^T K^{-p} r_{i,n}$ (4,000 updates, seeds 53/101/202, each problem's own corrector); every reported cell accepts all requests, and the low-diffusion control divides κ by $(65/33)^2$. Table E.3 gives the held-out outer changes; all 126 seed–split–exponent cells are in the supplementary archive. The main two-phase and eutectic-A medians favor $p = 1$, whereas Richards improves further at $p = 1.5$ and 2; the registered $p = 1$ versus 2 contrast reverses between Richards and eutectic A in both counters, splits and all seeds. Exceptions rule out a universal or per-seed optimum: two-phase development seed 202 uses 212 outer steps at $p = 1.5$ versus 216 at $p = 1$; eutectic-A seed 202 favors 0.5 over 1 on both splits; low-diffusion development $p1 - p1.5$ differences are $+1/-6/-5$; obstacle seed 53 ties $p0/p.5$ at 202 outer steps while matvec favors 0.5 (936 versus 940). The registered prediction that $p = 1$ is special therefore receives an unfavorable verdict, and the prediction contrasting compatibility with conditioning remains unresolved. At $p = 2$ the loss is best conditioned in the defect coordinates d; conditioning in the network parameters, through $J_d^\mathsf{T} K^{2-p} J_d$ with J_d the Jacobian of d, was not measured.

Reaction-placement exponents. A registered extension, posed after the $p = 0$ and $p = 1$ reaction results were known, trains the two-phase reaction problem (32^2, same recipe, seeds 53/101/202) at $p \in \{-1, -\frac{1}{2}, 0, \frac{1}{2}, 1\}$; a negative p applies $K^{|p|}$ in the sine basis. It is read on 64 held-out requests generated after registration (Table E.3, last row). Against $p = 0$, $p = -\frac{1}{2}$ changes outer work by $+0.4$, -0.8 and -1.2% at seeds 53, 101 and 202 (matvecs $+1.3$, -1.2 and $+0.3\%$), whereas $p = -1$, $\frac{1}{2}$ and 1 need more work at every seed in both counters. The cheapest exponent is $p = 0$ at seed 53 and $p = -\frac{1}{2}$ at seed 101; at seed 202 it is $-\frac{1}{2}$ for outer iterations and 0 for matvecs. Neither registered outcome holds: $p = 0$ is not cheaper than $p = -1$, $-\frac{1}{2}$ and $\frac{1}{2}$ at every seed in both counters, and no negative exponent is cheaper than $p = 0$ at every seed in both counters. The historical held-out and development requests give the same verdict, and all 45 exponent–seed–set cells accept every request. The $p = -1$ seed-101 run stalled in training (loss 8.00 after 4,000 updates, against 0.05–1.92 for the other 14 runs) yet accepts every request with about twice the outer work of $p = 0$; setting that run aside, $p = -1$ still costs more at the other two seeds.

Table E.3: Held-out outer change relative to $p = 0$ (%), seed medians of paired changes (–: untested). "Stable" in the archive means three same-sign seed differences, not statistical significance.

Model problem	$p=-1$	$p=-\frac{1}{2}$	$p=\frac{1}{2}$	$p=1$	$p=\frac{3}{2}$	$p=2$
Main two-phase, 64^2	–	–	-31.4	-38.9	-36.0	-20.4
Low-diffusion, 64^2	–	–	–	-10.3	-6.3	$+26.9$
Eutectic A, 32^2	–	–	-19.4	-20.7	-13.3	$+9.6$
Richards-type, 32^2	–	–	-42.1	-48.8	-50.4	-53.2
Two-phase reaction, 32^2	$+5.3$	-0.8	$+4.0$	$+10.9$	–	–

Mixed-placement comparison. The 64^2 interpolation (Appendix D.1) changes both residual and decoder; Figure E.1 plots the medians of Table E.4. Metrics are I, K^{-1} and the matched $B_\theta^{-1} = (\theta K + 4(1-\theta)I)^{-1}$, newly trained only at the first three interior points. Both counters and splits favor Euclidean up to $\theta = 0.935648$ and Inverse at 0.989906 and 1. At 0.968649 the outer and held-out matvec signs are mixed (held-out matvec differences, Euclidean minus Inverse, $+135/-39/+47$), so the tested points do not locate a unique or monotone crossing. The theoretical boundary $t_{64} = 0.667703274630$ is not that crossing: at 0.755527 Euclidean beats the compatible matched metric on every seed, counter and split (matched minus Euclidean outer $+3/+3/+8$ development and $+2/+5/+7$ held-out). The newly trained matched weighting beats Inverse on every seed but never both alternatives in both counters, so the registered joint-win prediction holds at none of the three interior points. At 0.989906 the endpoint-separation check fails (0.0067 against 0.01), so this point is not clearly distinct from the endpoint; its effect is still reported.

Table E.4: Mixed-placement work changes of Inverse relative to Euclidean (%). Each entry is the median of seed-paired changes. All 120 arm–seed–split cells accept every request (32/32 or 64/64) with finite paths and unchanged initial states, so these are same-acceptance comparisons.

	Development (32)		Historical held-out (64)	
θ	Outer	Matvec	Outer	Matvec
0	+8.4	+8.6	+8.3	+10.1
0.435869	+6.9	+8.1	+10.7	+13.4
0.755527	+16.8	+18.2	+10.9	+12.7
0.935648	+14.9	+8.9	+14.8	+11.8
0.968649	+0.5	−3.6	+2.6	−2.5
0.989906	−23.5	−25.3	−30.3	−31.6
1	−34.1	−37.1	−38.9	−41.5

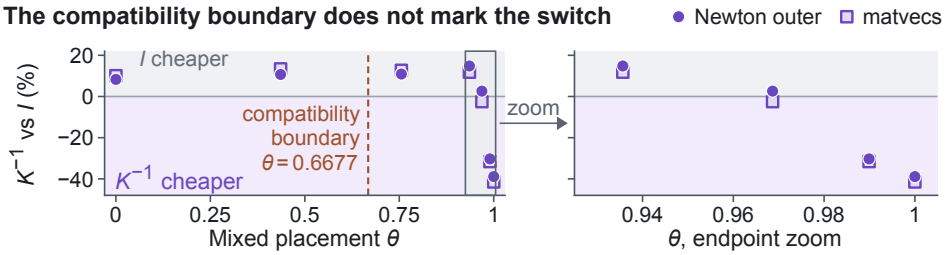

Figure E.1: **Mixed placement: the compatibility boundary does not mark the switch.** $B_\theta = \theta K + 4(1-\theta)I$ weights the constitutive term from reaction ($\theta = 0$) to diffusion ($\theta = 1$). Points are three-seed medians of the change in Newton outer iterations (circles) and active matvecs (squares) of K^{-1} against I; negative values favor K^{-1}. The dashed line is the compatibility boundary $t_{64} \approx 0.6677$ (Proposition B.1); the measured preference switches between $\theta = 0.94$ and 0.99. The left panel spans all settings and the right panel enlarges the marked interval $\theta \geq 0.925$; the seven settings are $\theta = 0, 0.44, 0.76, 0.94, 0.97, 0.99$ and 1.

F State interventions and branch-response diagnostics

F.1 Corrector dependence and phase interventions

Frozen-proposal protocol. Three fixed model pairs are evaluated on 64 new requests in eight blocks (an eight-request pilot is excluded), with identical q, h across correctors. Picard uses per-request relaxation, a cap of 600 iterations and acceptance checks every five iterations; all 768 learned and 64 cold-control evaluations accept. The relaxation is $\omega = 2/(a_{\min} + a_{\max})$ with $a = 1/(1+c\lambda)$ at the extreme eigenvalues $\lambda = 4 \mp 4\cos(\pi/(N+1))$ of K, which minimizes $\max_\lambda |1 - \omega a|$ for the linear part. It depends only on c, uses no reference, is shared by every arm on a request and also defines the L2WS fixed-point map. For $d = \log(1 + k_{\text{Inv}}) - \log(1 + k_{\text{Euc}})$, the registered request-block bootstrap of $J = d_P - d_N$ gives mean .3572 and 95% interval [.2939,.4191], above $\log(1.10)$ and conditional on the models. A registered rerun with every-iteration Picard checks moves the first-acceptance indices by 0–4, so the Picard comparison is robust to the check interval (Table F.1, last column; Section 5.3 reports the per-iteration values).

Association and robustness. Request 000, inspected earlier, is excluded, leaving 63; the aggregate effects were known before this analysis. Phase labels use the stored same-seed Euclidean/Newton terminal reference over frames 1–50, and deltas are Inverse minus Euclidean. With average-tie ranks, (i) the change in wrong-phase sites correlates with the Newton work change at 0.784/0.723/0.723 (seeds 53/101/202; log residual: $-0.153/-0.247/-0.014$), and (ii) with J at $-0.764/-0.715/-0.643$ (log residual: 0.185/0.251/0.095), meeting the registered thresholds (at least .5 for (i), at most $-.5$ for (ii)); these one-at-a-time correlations are not a causal decomposition. Excluding sites within 10^{-3} of the reference threshold (0–5,875 of 204,800) changes the correlations by less than .0018,

Table F.1: Native counts for the same frozen proposals. Raw effect compares summed counts; the shifted effect compares geometric means of $1 + k$. Euc./Inv.: Euclidean/Inverse. The combined row reuses the same 64 requests across three fixed pairs.

	Newton outer			Picard first-acceptance index			
Seed	Euc. / Inv.	Raw	Shifted	Euc. / Inv.	Raw	Shifted	Per-iter. raw
53	702/426	-39.3%	-29.4%	13,285/13,395	$+0.8\%$	$+0.6\%$	$+0.9\%$
101	708/438	-38.1%	-28.8%	13,135/13,415	$+2.1\%$	$+1.7\%$	$+2.3\%$
202	689/435	-36.9%	-28.3%	12,965/13,295	$+2.5\%$	$+3.0\%$	$+2.5\%$
Combined	2,099/1,299	-38.1%	-28.8%	39,385/40,105	$+1.8\%$	$+1.7\%$	median $+2.3\%$

leave-one-out retains the thresholds, and greedy deletion needs at least eight cases against a criterion of seven. The Picard checks fail (.792 against .75, and 89.8/87.3% against 90%).

Targeted repair and displacement control. A random subset of Euclidean's wrong-phase sites, as large as Inverse's reduction, is projected into the reference phase interval with all other sites fixed (five draws; matched-count displacement control; full repair designated a ceiling before evaluation). $R = (C_{\mathrm{Euc}} - C_{\mathrm{repair}})/(C_{\mathrm{Euc}} - C_{\mathrm{Inv}})$ uses work totals on the same 63 requests within each seed, with medians over draws within seed and then across seeds; it describes the reference-assisted operation of Section 5.2, not a mediated fraction of the training effect. Figure 3a is a protocol-consistent visualization replay of one request; the work values in Figure 3b,c come from the original runs. Outer median .976 versus .386 meets the registered half-control criterion; matvec .787–.800 versus

Table F.2: Outer iterations summed over 63 requests; five draws per arm.

Seed	W_{Euc}	W_{Inv}	Repair W	R	Control W	R_{ctrl}
53	690	417	432–442	0.923	573–584	0.392
101	697	430	429–439	0.985	590–609	0.386
202	677	427	429–437	0.976	577–600	0.380

.407–.457 fails it at every seed and for the median. The control raises phase errors and residual yet recovers 31–43%; full repair recovers 1.63–1.70, leaving a 2–21-iteration gap. All arms accept 63/63 under both correctors; repair beats the control under both, but only Newton improves substantially over the original Euclidean start (Table F.3).

Table F.3: Same-state intervention cross-read: percent change relative to unmodified Euclidean, seeds 53/101/202.

Initial state	Newton outer work (%)	Picard native work (%)
Inverse	$-39.6/-38.3/-36.9$	$+0.9/+2.3/+2.5$
Targeted repair	$-36.5/-37.7/-36.0$	$+0.4/+0.5/+0.5$
Displacement control	$-15.5/-14.8/-14.0$	$+12.6/+14.5/+15.8$

Corrector variants on the same starts. Preregistered after the main analyses, this evaluation reuses the starts of the twelve final static checkpoints on the 64 headline held-out requests and changes only the corrector (both weightings on the same host); the main configuration reproduces the headline totals exactly. GMRES is unrestarted with the 80-iteration inner cap; exact inner solves factor the per-step diagonal blocks with sparse LU; fixed forcing sets $\eta = 10^{-2}$; the FISTA variant starts step n from the learned $\widehat{h}_n$, one request at a time. All 768 evaluations per corrector are accepted (Table F.4). Outside the registration, FISTA from the clamped previous state needs 6.4–8.0% fewer gradient iterations than from either learned start.

Table F.4: Inverse-relative-to-Euclidean work change (%) by corrector and seed, in outer iterations (gradient iterations for FISTA); negative favors Inverse.

Corrector	53	101	202	303	404	505	Lower
Newton, BiCGSTAB (main)	-38.9	-41.7	-37.0	-32.9	-34.6	-37.1	6/6
Newton, GMRES	-39.1	-41.4	-37.1	-33.7	-34.4	-37.2	6/6
Newton, exact LU	-40.0	-43.0	-38.7	-34.3	-35.4	-36.7	6/6
Newton, fixed forcing	-39.8	-42.8	-38.4	-34.2	-34.7	-36.7	6/6
FISTA from learned start	$+0.1$	$+0.8$	$+0.4$	$+1.2$	-0.0	-0.7	2/6

F.2 Fixed-branch response analysis

Table F.5: Fixed-branch response analysis by training seed, without a mediation fraction or population confidence interval. Negative ΔS: smaller fixed-branch response for Inverse.

Quantity	53	101	202
Mean / median ΔS	-0.0215 / -0.0189	-0.0235 / -0.0215	-0.0223 / -0.0234
Partial rank correlation	0.5631	0.5414	0.4070
Unconditional Spearman	0.8171	0.7556	0.6995
Control design rank; unresolved signs	5; 1	5; 1	5; 1

For a proposal a, fix its affine constitutive branch $\beta(h) = D_a h + o_a$ over the window and let $\Phi(a)$ solve the resulting full-window linear system $((I + cKD_{a,n})\Phi_n - \Phi_{n-1} = f_n - cKo_{a,n}, \Phi_0 = h_0)$. With the same-seed accepted reference z, the fixed-branch response error is $S(a) = \mathrm{RMS}(\Phi(a) - \Phi(z))$ and $\Delta S = S(a_{\mathrm{Inv}}) - S(a_{\mathrm{Euc}})$. The same 63 requests serve all three seeds, and all 567 solves qualify by their affine-system residual. One paired sign per seed is unresolved by a conservative error bound and counted neither as zero nor as signed. Both correlations are between paired response-error and work differences; the partial rank correlation controls for five prespecified variables (wrong-phase count, state error, Inverse residual energy, weighted defect energy and affine-branch energy) by residualizing both rank vectors; the design has rank five at every seed. Exact definitions, the error bound and the regression are in the supplementary details. These are retrospective associations for fixed full-window affine responses, not production Newton-Jacobian conditioning.

G Learning-baseline identities and complete static comparisons

G.1 Positioning against the closest method families

The baselines adapted in Appendix G.2 are the closest learned method families: L2WS and SIP, whose distinct loss-gradient semantics Appendix G.2 specifies (Sambharya et al., 2024; Holl et al., 2022), together with our own Newton-step displacement, Jacobian-smoothed and supervised variants. Active-set finite termination (Brugnano & Sestini, 2011), slow Newton-min paths in positive-definite complementarity (Dussault et al., 2019), Hermitian Jordan-product bounds (de Bruijn, 1980) and absolute-value-equation residual error bounds (Zamani & Hladík, 2023) are the classical precedents for the finite-grid, decoder-feasible exact paths and production-solver limits of Section 5.1. A comparison matrix is in the supplementary archive.

G.2 Solver-aware training and external baselines

Objectives and information. The methods below also change components other than the residual metric; we adapt them to our task rather than reproduce the source papers' experiments, and compare proposals under one common Newton corrector or each method's native iteration. *Composite operator-energy training (Energy/BT)* is our own objective, not a published method; Energy/BT is our shorthand for it. It minimizes $\mathcal{L}_{\mathrm{BT}} = B^{-1}\sum_{i,n}(\|b_{i,n}\|_1 + 2c_i\sqrt{d_{i,n}^{\mathsf{T}} K d_{i,n}})$ for the balance and constitutive defects of Section 3. At exact diffusion-decoder balance the second term is $2\|r_{i,n}\|_{K^{-1}}$,

Table G.1: **All adapted single-window baseline configurations.** Median totals over 64 requests and three seeds. Eq: 4,000 updates; Cal: calibration-derived updates. These counts are not times.

Method	Labels	U_{Cal}	Min acc.[a]	Outer		Matvec	
			Eq/Cal	Eq	Cal	Eq	Cal
Euclidean	No	4,000	64/64	702	702	4815	4815
Inverse	No	4,000	64/64	435	435	2746	2746
Energy/BT	No	4,000	64/64	425	425	2711	2711
NSD3	No	1,749	64/64	446	507	2769	3146
JSN2	No	455	64/60	507	829	3560	5713
L2WS-FP15	No	3,466	64/64	430	440	2682	2734
L2WS-FP60	No	2,444	64/64	622	633	3820	3958
L2WS-Reg1	Yes	4,096	64/64	432	425	2675	2650
L2WS-Reg60	Yes	2,469	61/60	826	892	5053	5498
SIP-Newton	No	2,604	64/64	438	477	2746	2926

[a]Minimum accepted /64 across seeds for Eq/Cal; failed work retained. JSN2-Cal: 64/60/62; Reg60: 61/64/62 (Eq), 60/64/61 (Cal). Selection, native deployment and per-seed counts: Appendix G.2.

whose square Inverse uses, so BT also changes homogeneity and request weighting. *Newton-step displacement (NSD3)* advances the decoded proposal by up to three damped Newton steps and penalizes its displacement; the implemented gradient includes $2(1-\rho_i)$, $\rho_i = \prod_j (1-\alpha_{ij})$ with accepted damping lengths α_{ij}, unlike regression to a fully stopped teacher. Full NSD3 uses 4,000 updates and NSD3W 1,749, each with its own full cosine schedule. *Jacobian-smoothed Newton training (JSN2)* instead penalizes the displacement of two smoothed Newton steps. *L2WS* (Sambharya et al., 2024) forms $z_k = \mathcal{T}^k(z_0)$ and minimizes the unsquared norm $\|\mathcal{T}(z_k) - z_k\|_2$ (FP) or $\|z_k - z^*\|_2$ (Reg), where for decoder $\mathcal{D}$ and relaxation ω, $\mathcal{T}(z) = (1-\omega)z + \omega P(\mathcal{D}(z))$ and $\mathcal{T}(z) - z = -(\omega/c)K^{-1}r$ at exact balance. Unrolling, unsquared homogeneity and request-dependent scaling distinguish FP from static $p = 2$. *Supervised baselines (DS-Q, DS-H)* are trained on reference solutions: DS-Q predicts q, decodes it with (2) and minimizes the squared error of the decoded h; DS-H predicts h directly, without the decoder, with the same loss. *Physics-inversion training (SIP-Newton)* (Holl et al., 2022) regresses the decoded proposal to a Newton-teacher state without differentiating that target; teacher computations and unrolled steps are charged to training. *Budgets and selection.* Calibrated counts (Table G.1; Appendix A.1) match update investment, not realized wall time. The primary common-Newton configurations, fixed before evaluation, are FP15, Reg1 and SIP; native FP/Reg use 60 steps; all other configurations and failures are reported.

External baselines on single windows. The confirmation uses 64 new single-window requests with the two budgets and the deployment-specific configurations above, all fixed before evaluation, and acceptance at $4 \cdot 10^{-4}$ (Section 3). Table G.1 includes all ten configurations and their completion. Inverse and SIP-Newton share the median matvec total 2,746 by coincidence: their per-seed totals are 2,746/2,724/2,813 and 2,750/2,746/2,688 (seeds 53/101/202). In native deployment all learned FP/Reg arms accept 64/64 with 2,610–2,910 whole-window Picard applications (shared batches), and k60 improves over k15/k1; hold-initial-state accepts all with 2,365, its unrelaxed version 62/64 with 4,045, and FISTA 64/64 (work not recorded). Native and Newton units differ. Two timing attempts of this campaign failed the timing protocol, so no times are reported from them.

Matched solution labels. Both weightings receive the same solution labels (preregistered): each arm minimizes $\mathcal{L}_p/\bar{\mathcal{L}}_p + \mathcal{L}_{\mathrm{DSQ}}/\bar{\mathcal{L}}_{\mathrm{DSQ}}$, where $\mathcal{L}_{\mathrm{DSQ}}$ is the DS-Q squared error of the decoded state against the reference solution and each $\bar{\mathcal{L}}$ is fixed at its first-batch value at initialization; everything else follows the headline recipe. At zero label weight the headline trainers are reproduced exactly; all arms accept every request (Table G.2). Inverse with labels needs 3.9, 6.8 and 9.2% less outer work than Euclidean with labels (median -6.8%, smaller in magnitude than the preregistered -10% mark), and Euclidean with labels needs 3.6, 7.1 and 3.7% more than headline Inverse without labels. Adding labels changes Inverse by a median -0.5% and Euclidean by -36.7%.

Table G.2: Held-out totals with and without matched solution labels, seeds 53/101/202.

Arm	Outer	Active matvec
Euclidean (headline)	682/701/686	4,360/4,514/4,414
Inverse (headline)	417/409/432	2,550/2,475/2,655
Euclidean with labels	432/438/448	2,595/2,627/2,672
Inverse with labels	415/408/407	2,549/2,485/2,431

H Repeated deployment and qualified timing reads

H.1 Fine-tuning on inputs from accepted states

Inputs and checkpoints. In the appendix tables, A, B and D label the accepted-state, original-input and fresh-input continuations of Inverse (Section 6.1), and C and F the Euclidean counterparts of A and B (C with its own pool from Euclidean chains). A's training chains start from the first-window requests of the original training set (window-2–8 sources from a separate stream, parent κ kept); each seed's fixed update-4,000 model and corrector generate the accepted entries, detached for continuation. D draws 896 fresh first-window requests and takes the first n_s to match A's pool size (896 entries at every seed except 892 at seed 202), not request-wise source/κ. All arms use 4,000 matched-optimizer continuation updates and the final update-8,000 checkpoint; development diagnostics do not select. Batches contain four original first-window inputs plus four pool inputs for A/D and eight original inputs for B.

The design benchmark's 64 eight-window parents are new draws whose generator reads neither the training requests nor learned models, but prior diagnostics on it informed the adaptation design and it is reused for evaluation; the fresh benchmark (Appendix H.3) removes this reuse. Each route propagates its own accepted states. Resampling eight eight-trajectory blocks gives conditional 95% intervals for A/B (Table H.1) of $[-34.5, -31.8]\%$ and $[-33.8, -31.2]\%$ (original/added seeds, design benchmark) and -32.8% $[-33.3, -32.4]$ and -31.9% $[-32.4, -31.5]$ (fresh benchmark); they describe request composition for fixed models, not new seeds or timing sessions.

Table H.1: Windows-2–8 Newton outer iterations summed over 64 trajectories, and changes (%). D has only the original three seeds.

Seed	53	101	202	303	404	505
Design benchmark						
A / B	4,109/6,057	4,195/6,278	4,143/6,366	4,182/6,079	4,063/6,239	4,112/6,093
D	6,182	6,372	6,523			
A vs. B	−32.2	−33.2	−34.9	−31.2	−34.9	−32.5
A vs. D; D vs. B	−33.5; +2.1	−34.2; +1.5	−36.5; +2.5			
Fresh preregistered benchmark						
A / B	4,063/5,973	4,192/6,197	4,111/6,233	4,175/5,952	4,057/6,073	4,057/6,028
D	6,092	6,316	6,448			
A vs. B	−32.0	−32.4	−34.0	−29.9	−33.2	−32.7
A vs. D	−33.3	−33.6	−36.2			

Design-benchmark medians: A vs. D -34.2%, D vs. B $+2.1\%$, A vs. B -32.8% over six seeds.

H.2 Deployment routes and their algorithms

This subsection defines every route that the service tables of Appendix H.3 time; all routes solve the same eight-window trajectories under the acceptance test of Section 3. A learned route proposes a whole window and corrects it with the whole-window Newton corrector of Appendix A.1; a rejected window restarts from the same incoming state with charged q-form FISTA (fallback rules in Appendix H.4). Continuations A, B, C, D and F are defined in Appendix H.1. *Energy/BT continued* follows A's recipe with the Energy/BT loss in place of the Inverse loss; the continuation loss

reproduces the original loss and gradients exactly. *DS-Q continued* follows A's recipe on reference labels for the 2,684 pool windows (1,502 CPU core-seconds of labeling). The static configurations of Table H.2 are the update-4,000 models of Appendix G.2 without continuation (NSD3W: 1,749 updates; Table H.2). *Copy initial state* uses the initial frame of the window as the proposal for every step and enters the same corrector and fallback.

Classical routes. *Default q-form FISTA* is the per-step solver of Appendix A.1: a fixed-momentum accelerated projected-gradient method in the phase fraction q, started from $P(h_{n-1})$, with residual checks at iteration zero, every five iterations and at the 2,000 cap. *Tuned q-form FISTA (route 1)* starts each step from the extrapolation $2h_{n-1} - h_{n-2}$; of five variants read three times each on the development replay before timing (residual checks every two iterations or every iteration, gradient adaptive restart, the extrapolated start, and the first three combined), this one was fastest (0.703 s per parent against 0.906–1.280 s). *Coarse-grid classical route (route 2)* restricts each request to 32^2, solves it with the same FISTA ($\varepsilon = 10^{-4}$, whole-window acceptance required), prolongs h bilinearly with the exact fine h_0 and corrects with A's whole-window Newton; it rebuilds its 32^2 spectral tables for every batch. *Per-step semismooth Newton* solves each of a window's 50 steps as a one-step window with the same corrector and per-step tolerance $\varepsilon/50$, started from the previous accepted state or from the extrapolated start, with FISTA fallback.

FISTA in temperature form (route 3). With b and K as in the q-form step of Appendix A.1, the same implicit step $h + cK\beta(h) = b$ is solved for the temperature $T = \beta(h)$ instead of the phase fraction q. Since $h = q + T$ with $q \in [0, 1]$, $q = 0$ where $T < 0$ and $q = 1$ where $T > 0$, the step is the optimality condition of the strictly convex problem $\min_T \frac{1}{2}T^\mathsf{T}(I + cK)T - b^\mathsf{T}T + \sum_i \max(T_i, 0)$, whose subdifferential condition $(I + cK)T - b + q = 0$ with $q_i \in \partial \max(T_i, 0)$ is the step itself. The smooth part has gradient $T + cKT - b$, which applies K by the five-point stencil, so no inverse of $I + cK$ is formed, and the proximal map of the nonsmooth part is explicit: $\mathrm{prox}_{\tau \max(\cdot,0)}(z)_i = z_i - \min(\max(z_i, 0), \tau)$. The route runs the q-form's accelerated proximal-gradient iteration, $T^+ = \mathrm{prox}_{g/L}\big(y - (y + cKy - b)/L\big)$, $y^+ = T^+ + \theta(T^+ - T)$, with $g = \sum_i \max(T_i, 0)$, $L = \max(1 + c\lambda)$ and $\mu = \min(1 + c\lambda)$ over the DST-I eigenvalues λ of K, and the fixed momentum $\theta = (\sqrt{L} - \sqrt{\mu})/(\sqrt{L} + \sqrt{\mu})$; its condition number L/μ equals that of the q-form. The state is recovered from the step equation, $h = b - cKT$, and each step stops on the same true step residual of h as the q-form, with the same check schedule. The timed configuration starts step n from the temperature $\beta(2h_{n-1} - h_{n-2})$ of the extrapolated state, with residual checks every five iterations and no restart; it was chosen on the development replay (read on a shared host) by the selection rule of route (1) among four variants, and a fifth, with adaptive restart, failed a device-port count check and was excluded before timing.

H.3 Service times by session

Design benchmark: static configurations and continuations. Tables H.2–H.4 give the service seconds of every timed configuration by session (routes in Appendix H.2; clocks and read qualification in Appendix H.4). Table H.2 lists the static configurations; the continuations of the same session are in Table 4, and its A row is summarized in Table H.3. Every static learned configuration is slower than FISTA at every seed, and A is faster than FISTA at every seed.

Fresh benchmark. After the configurations were fixed, a second benchmark was registered before generation: 64 eight-window trajectories from a fresh seed with the same generator and path-residual contract, used only here. After reproducing the design benchmark's counts and signatures for A (seed 53) and FISTA, 35 frozen configurations were timed under the same clocks and guard; all 175 reads are valid and complete 64/64. Both preregistered questions hold at every seed: A needs less windows-2–8 outer work than B (six seeds) and D (three; Table H.1), and A is faster than FISTA (six seeds). A, B, D and FISTA use no fallback; C uses one fallback window at seeds 53 and 202, F two to five, and the copy-initial-state start completes through fallback in 302 windows. FISTA is 3.7% slower here than on the design benchmark, so B, D (53/101), L2WS-Reg1, DS-Q (53/202) and L2WS-FP15 (101) also finish below it, by 0.1–8.6%.

Later sessions. Three later preregistered sessions reuse the harness (Table H.3). Per-step semismooth Newton completes 64/64 without fallback using 4.12 outer iterations and 12.40 matvecs per

Table H.2: **Static configurations on the design benchmark.** Service seconds per complete eight-window trajectory (lower is better), median of five timing repeats; all repeats complete 64/64.

Configuration	s=53	101	202	Configuration	s=53	101	202
Common Newton: residual/defect				*Common Newton: supervised*			
Euclidean	1.212	1.284	1.218	L2WS-Reg1	0.752	0.744	0.739
Inverse	0.774	0.786	0.845	DS-Q (h label)	<u>0.729</u>	<u>0.739</u>	<u>0.726</u>
Energy/BT	0.741	0.762	0.778	*Native Picard*			
Common Newton: solver-assisted				L2WS-FP60	0.766	0.768	0.761
NSD3	0.775	0.770	0.780	L2WS-Reg60	0.762	0.757	0.753
NSD3W	0.917	0.854	0.927				
L2WS-FP15	0.775	0.768	0.842	*Classical*			
SIP-Newton	0.769	0.765	0.767	FISTA	**0.712** (one configuration)		

4,000 updates except NSD3W (1,749) and FISTA (no training). NSD3W is NSD3 trained for 1,749 updates on its own cosine schedule, the count whose total steady-state price equals that of 4,000 Energy/BT updates. DS-Q predicts q, decodes it with (2) and minimizes the squared error of the decoded h against the reference solution. **Bold**/<u>underline</u>: fastest/second per seed. Fallback: Appendix H.4.

step, and the order A, FISTA, per-step Newton holds at every seed. The extrapolated start makes FISTA 6.7% (design) and 6.1% (fresh) faster than the default; A is faster than tuned q-form FISTA at every seed, and continued DS-Q is faster than A at every seed, on both benchmarks. On the separate 32^2 eutectic-A problem the 64-parent workload takes 9.152–9.501 s (A), 11.829–12.909 s (B), 11.486–12.507 s (D) and 30.458 s (FISTA), that is 0.143–0.148 s per trajectory for A against 0.476 s for FISTA.

Table H.3: Service seconds per eight-window trajectory by timing session and route (median of five reads; every read completes 64/64). Percentages compare readings of the same session and are computed from the unrounded medians, so they can differ from ratios of the printed three-decimal values by up to 0.2 points.

(a) Fresh benchmark, preregistered session	s=53	101	202	303	404	505
A	**0.509**	**0.534**	**0.506**	**0.523**	**0.506**	**0.497**
A faster than FISTA (%)	31.1	27.7	31.5	29.1	31.5	32.7
B	0.675	0.705	0.716	0.680	0.701	0.675
D	0.696	0.725	0.743			
C (Euclidean, returned)	0.889	0.828	0.854			
F (Euclidean, original)	0.992	1.009	0.997			
DS-Q	0.732	0.743	0.707			
Energy/BT	0.745	0.761	0.773			
L2WS-FP15	0.743	0.737	0.799			
L2WS-Reg1	0.735	0.734	0.727			
FISTA (default) 0.738 and copy initial state 1.917, one configuration each						

(b) Other sessions, ranges over seeds 53/101/202 (six seeds where stated)

Design benchmark, first service timing (Table 4): FISTA (default) 0.712; A 0.501–0.537 (six seeds), 24.6–29.6% faster than FISTA.
Design benchmark, Energy/BT continuation session: FISTA (default) 0.747; A 0.519–0.539; Energy/BT continued 0.510–0.534, 28.5–31.8% faster than FISTA and faster than A at seed 101 only.
Tuned-FISTA and supervised-continuation sessions, design / fresh benchmark: FISTA 0.747 / 0.725 default and 0.697 / 0.681 tuned; A 0.522–0.533 / 0.508–0.533, 23.5–25.2% / 21.7–25.4% faster than tuned FISTA; DS-Q continued 0.490–0.492 / 0.494–0.498, 5.7–8.1% / 2.8–7.2% faster than A; static DS-Q 0.733 / 0.727 (seed 53).
Fresh benchmark, per-step Newton session: FISTA (default) 0.780; A 0.505–0.543 (six seeds); per-step Newton 2.488 from the previous accepted state and 2.162 from the extrapolated start.

Registered comparison with classical starts and solver forms. This session times A (seeds 53/101/202) against routes (1)–(3) of Appendix H.2 on a quiet host under the workload of Ap-

Table H.4: Registered classical comparison: service seconds per eight-window trajectory (median of five reads; every read completes 64/64 without fallback) and fine-grid Newton outer iterations over all windows of the third benchmark; seed order 53/101/202. The percentages in the text are computed from the unrounded medians.

Route	Third benchmark (s)	Outer	Design benchmark (s)
A	0.542/0.536/0.535	4,517/4,642/4,552	0.535/0.546/0.538
(1) Tuned q-form FISTA	0.707	–	0.739
(2) Coarse-grid classical route	1.024	7,077	1.048
(3) Temperature form	0.424	–	0.418

pendix H.4. It was registered before any reading; every route was frozen and hashed on the development replay before a third benchmark of 64 eight-window trajectories was generated from a new stream seed, and the design benchmark is timed again as reused, secondary evidence. The registered rule calls a route faster than A only if it is faster at every seed on the third benchmark, with the design benchmark agreeing (Table H.4).

Route (3) is faster than A at every seed on both benchmarks, by $21.8/20.9/20.8\%$ ($21.8/23.4/22.4\%$ on the design benchmark). A is faster than route (1), by $23.3/24.2/24.3\%$ ($27.7/26.2/27.2\%$), and than route (2), by $47.1/47.7/47.8\%$ ($49.0/47.9/48.6\%$). Route (2) also needs more fine-grid Newton work than A: 7,077 outer iterations and 49,263 matvecs against 4,517–4,642 and 32,218–33,163 (design benchmark: 7,279 against 4,485–4,575 outer iterations). Route (2) spends 57,205 coarse FISTA iterations and 71,846 coarse inverse applications on the third benchmark.

Five-read ranges on the third and the design benchmark are 0.519–0.554 and 0.526–0.554 s for A, 0.685–0.734 and 0.716–0.757 s for route (1), 0.970–1.034 and 1.013–1.053 s for route (2), and 0.365–0.480 and 0.369–0.451 s for route (3); by registration only medians decide.

H.4 Timing protocol, read qualification and sensitivity

Workload, clocks and qualification. The service timing runs the 64 eight-window parents of a benchmark (batch eight) one route process at a time on one RTX 4090, with acceptance $\ell_1 \le 4 \times 10^{-4}$; common routes share Newton and native FP60/Reg60 use relaxed Picard. Charged FISTA fallback restarts from the same incoming state; under the strict-acceptance contract of Appendix A.1, a rejected fallback ends its parent. After one warm-up batch, synchronized host-wall service time covers transfers, prediction, setup, correction, acceptance, record handling and fallback, divided by the 64 parents; loading and warm-up fall in a separate cold clock. The five reads repeat the same checkpoints; their median is descriptive and was fixed in advance for the fresh benchmark. Every read must pass a quiet-host guard and a numerical-signature check of its whole record (guard mechanics in the supplementary details), and every unit starts a cold process for each read. JSN2 was not timed.

Qualified reads and fallback. All 275 reads of the first session qualify and complete 64/64 parents; 85 reads invoke fallback for at least one window (a count of reads, not of windows or parents), whereas A/B (all six seeds), D (53/101/202) and FISTA never fall back. In the classical session, one read of route (3) on the design benchmark was voided by the host-load guard and repeated; every read of that session completes 64/64 without fallback, work counts are identical across reads, and A's and FISTA's counts on the design benchmark reproduce the earlier tuned session. Twenty of the 45 design-benchmark reads of the tuned-FISTA session overlapped CPU-only jobs on the same host; the load guard passed every read, and using only non-overlapping reads leaves the preregistered outcome unchanged.

Sensitivity. Across the five reads of the first session, 119 of 1,485 route pairs change service-time ordering. With identical work counts, FISTA's median shifts by up to 5.7% and A's by -0.2 to $+4.5\%$ across sessions, which is why comparisons use readings of one session only. Reductions $100(1-t_s/t_{\mathrm{FISTA}})$ use five-read medians; they describe complete configurations on the stated bench-

mark, not a confidence interval or the isolated effect of residual weighting, and exclude training and pool preparation.

Typed correction work. Newton outer steps and active matvecs are summed over request–windows, whereas batch matvec, preconditioner and fixed-phase calls, native Picard maps and FISTA gradient iterations, inverse calls and residual checks are shared-batch counts; they are neither added across coordinates nor treated as equal costs, and a dash marks a count that does not apply, not zero work. Over all eight windows, A/B/D use 4,439–4,575/6,434–6,753/6,563–6,905 outer steps (identical across the five reads, zero fallback) and FISTA 127,735 gradient iterations, 156,482 inverse calls and 28,747 residual checks; the per-seed records are in the supplementary details.

I TRAINING INFORMATION, RESOURCE ACCOUNTING, EVIDENCE ACCESS AND SCOPE

All learned routes differentiate through the linear decoder, a 50-step recursion; Euclidean/Inverse/BT losses use no future-solution labels or nonlinear updates, and DS-Q supervises the decoded future h (no static row in Table 2). Solver-aware methods unroll the steps listed in Appendix G.2 (Newton inner cap 200; JSN2 uses a local two-step smoothed-Newton adjoint). Continuation losses contain no nonlinear solve, and pool acquisition is charged.

Account **I** charges new continuation/pool work given the parent checkpoint and original data; **N** deduplicates all required recorded training, data and pool costs, including static training; **R** keeps consumer allocations symbolic. Units are run-wall seconds, not GPU-card time or money. Selection, inherited, setup, search and failure costs are unknown and not set to zero. Amortization equates one recorded run-wall second with one deployment-wall second: with service saving $\Delta t > 0$ and known upfront difference ΔU, $N = \lceil \Delta U/\Delta t \rceil$ ($N = 0$ if the upfront cost is also lower) and $N_{64} = 64\lceil N/64 \rceil$. Unknown cost differences can have either sign; these counts are neither economic payback nor full break-even bounds. Across seeds 53/101/202/303/404/505 (per-seed accounts in the supplementary details), A's upfront seconds are 1151.00–1277.41 (I) and 1950.27–2926.07 (N); against default FISTA, A repays its continuation and pool after $N = 5{,}472$–$7{,}063$ trajectories ($N_{64} = 5{,}504$–$7{,}104$) and all preparation after $N = 10{,}190$–$16{,}708$ ($N_{64} = 10{,}240$–$16{,}768$); against B the counts are 0–3,092 (N_{64} 0–3,136; 0 at seeds 303/404).

Application scope. Predictive twins link observations, models and decisions (Kapteyn et al., 2021); we study simulation on synthetic model problems with supplied physics and no experimental data, not equipment validation, assimilation or control (Shao et al., 2023), which would need calibrated laws, process tolerances, complete costs and independent measurements.