PHYSICS-QUALIFIED SAFEGUARDS FOR RELIABLE ITERATIVE NEURAL OPERATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Iterative neural operators can reach a learned fixed point that remains physically invalid, and fast fixed-point acceleration can exacerbate this failure under distribution shift. We introduce PHYSSAFE-IRNO, a reliability wrapper that aligns learned stationarity with physical validity during training and physics-qualifies accelerated proposal at inference. During training, stationarity-conditioned physical-risk supervision discourages physically invalid learned fixed points, while averagedness-oriented regularization shapes the refinement map for safeguarded fixed-point updates. At inference, a physics-qualified SuperMann solver screens numerically admissible fast proposals using a learned–physical bi-residual filter and falls back to separating-halfspace or Krasnosel'skii–Mann updates when needed. Under the explicit assumption that the learned refinement map is exactly averaged, the wrapper inherits SuperMann convergence, under a calibrated residual-envelope condition, sufficiently small learned residual implies finite-tolerance physical validity. Prediction-only temporal residuals avoid dependence on future ground-truth states, and a disjoint physical audit evaluates returned states using separately implemented diagnostics. Across four PDE systems and six held-out shifts, PHYSSAFE-IRNO reduces aggregate error from 0.087 for physics-aware Anderson acceleration and 0.092 for SuperMann to 0.068, cuts the false-fixed-point rate to 1.3%, and raises joint success to 96.8%. The gains over fixed IRNO transfer across FNO, TFNO and CNO at 1.16× the wall-clock cost of fixed-step IRNO. The implementation code is provided in the Supplementary Material.

1 INTRODUCTION

Neural operators amortize repeated partial differential equation (PDE) simulation by learning maps between function spaces (Li et al., 2021; Lu et al., 2021; Kovachki et al., 2023). Their speed makes them attractive for fluid dynamics, transport, climate, and materials, but a single forward pass can under-resolve weak or high-frequency structures because of spectral bias (Rahaman et al., 2019; Tancik et al., 2020). Iterative approaches revisit the prediction rather than increasing the backbone size. PDE-Refiner uses diffusion-inspired corrections for long rollouts (Lippe et al., 2023), while Iterative Refinement Neural Operators (IRNO) wrap a frozen operator with a shared learned correction map and progressively supervise higher frequencies (Liu et al., 2026). The frozen-backbone design is appealing when retraining a large operator is expensive.

The fixed-point view of IRNO exposes a gap between learned convergence and physical validity. For a PDE instance a, let $T_{\theta,a}$ denote the learned refinement map and $R_{\theta,a}(u) = u - T_{\theta,a}(u)$ its fixed-point residual. A small $\|R_{\theta,a}(u)\|$ indicates that the learned refiner is nearly stationary, but does not imply a small PDE residual, correct boundary conditions, conservation, or admissibility. Refinement can therefore terminate at a state that is stationary for the network but still violates the governing physics, a failure we refer to as a false fixed point. Standard Anderson acceleration can accelerate convergence toward such learned equilibria (Wu et al., 2026), and distribution shift can further enlarge the mismatch because both correction geometry and stopping behavior are extrapolated. Figure 1 makes the central geometry explicit: reliable refinement must enter the target region on both the learned-residual and physical-residual axes. Reducing only the learned residual can therefore create the appearance of convergence without physical validity.

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.

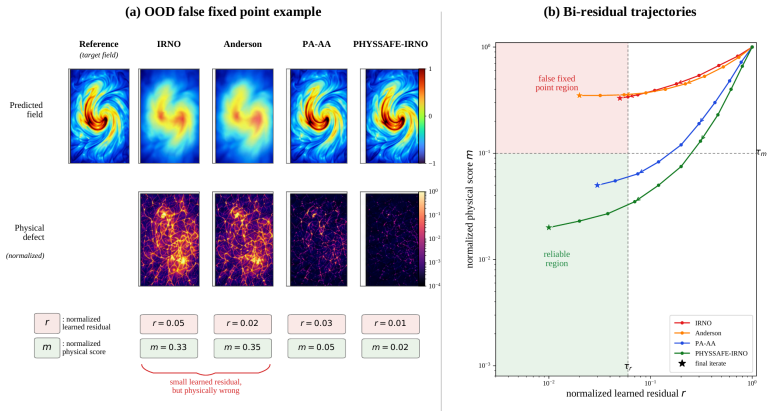

Figure 1: False fixed points under shift. The OOD example shows IRNO and Anderson reaching small learned residuals while retaining large physical defects; the bi-residual view identifies the joint target region and the terminal trajectories of all solvers.

There is a complementary numerical issue. Fixed Picard-like refinement is stable but slow, whereas Anderson and Broyden directions use recent secant information to accelerate local convergence (Walker & Ni, 2011; Toth & Kelley, 2015; Broyden, 1965). Away from a regular local basin, however, an unprotected fast direction can overshoot, oscillate, or amplify a defective learned equilibrium. Physics-aware Anderson acceleration (PA-AA) addresses false fixed points by incorporating a physical residual into an Anderson-style accelerated least-squares construction (Wu et al., 2026). Our goal is different: we seek a qualification layer that can screen general fast directions while retaining an explicit safeguarded recovery path. SuperMann instead globalizes arbitrary fast directions for averaged/nonexpansive maps through separating halfspaces and Krasnosel'skii–Mann (KM) fallback (Themelis & Patrinos, 2019), but it is agnostic to physical validity.

PHYSSAFE-IRNO addresses the two failure modes separately: training reshapes where the learned refiner becomes stationary, while inference controls how accelerated trajectories approach those fixed points. During training, we penalize physical violations specifically near learned stationarity and regularize the refiner toward averagedness. At inference, we first test whether a fast proposal is numerically admissible and only then apply a learned–physical bi-residual qualification. Crucially, the physical test is one-sided: it may veto a numerically admissible fast step, but it can never authorize a step rejected by the numerical safeguard. This design keeps the executed updates within the SuperMann update classes while allowing physics to block adverse acceleration. A disjoint audit score then evaluates returned states independently of the operational score used for training, filtering, stopping, and calibration.

Our contributions are threefold:

- We identify and formalize false fixed points in iterative neural operators as a mismatch between learned stationarity and physical validity, and evaluate this failure through paired learned–physical residuals and a disjoint physical audit.
- We introduce PHYSSAFE-IRNO, which combines stationarity-conditioned physical-risk training with a one-sided learned–physical qualification of accelerated proposals, while retaining separating-halfspace and KM recovery around an unchanged frozen backbone.
- We provide a conditional fixed-point analysis separating numerical convergence from physical validity through exact averagedness and residual-envelope conditions, and validate the framework across four PDE systems and three operator families.

2 RELATED WORK

Neural operators and iterative correction. FNO, DeepONet, factorized FNO, CNO, GINO, latent operators, and state-space variants encode spectral, convolutional, geometric, and latent inductive

biases (Li et al., 2021; Lu et al., 2021; Tran et al., 2023; Raonić et al., 2023; Li et al., 2023; Wang & Wang, 2024; Zheng et al., 2024). Physics-informed and OOD-oriented operators strengthen the surrogate through PDE supervision, manufactured solutions, sparse observations, domain adaptation, function-space extensions, and boundary augmentation (Li et al., 2024; Kim et al., 2026; Bischof et al., 2025; Zhu et al., 2023; Li & Yang, 2025; Quackenbush & Atzberger, 2026; Zhou et al., 2025). PDE-Refiner and IRNO revisit a prediction iteratively (Lippe et al., 2023; Liu et al., 2026); IRNO's frozen-backbone fixed-point map provides the setting in which we make repeated correction reliable under acceleration and shift.

Physical diagnostics in learned solvers. Neural-operator warm starts retain a classical solver's terminal accuracy (Eshaghi et al., 2026); residual-based analyses relate PDE defects to approximation error (Berrone et al., 2022; De Ryck et al., 2024; Liu et al., 2023; Eiras et al., 2024); residual-conditioned correctors feed the defect into learned updates (Jiang et al., 2026); and uncertainty methods quantify predictive confidence (Magnani et al., 2025). PA-AA directly targets false fixed points by replacing the accelerated least-squares target with a physical residual (Wu et al., 2026). PHYSSAFE-IRNO instead retains a general fast direction and uses paired learned and physical residuals to qualify its trajectory inside a globalized fixed-point scheme.

Averaged operators, acceleration, and filters. Plug-and-Play and RED analyses motivate Lipschitz, averaged, proximal, and pseudo-contractive structure for learned fixed-point maps (Ryu et al., 2019; Bohra et al., 2021; Hurault et al., 2022; Wei et al., 2024). Anderson and Broyden provide multisecant acceleration (Walker & Ni, 2011; Toth & Kelley, 2015; Broyden, 1965); SuperMann globalizes bounded fast directions through blind, educated, separating-halfspace, and KM updates (Themelis & Patrinos, 2019); and filter methods preserve multiple progress measures instead of scalarizing them (Fletcher & Leyffer, 2002; Gould et al., 2004). PHYSSAFE-IRNO combines these geometries: averagedness-oriented training shapes the refiner, SuperMann supplies the numerical update classes, and a nondominance archive jointly qualifies learned stationarity and physical validity.

3 PROBLEM FORMULATION

Let $a \in \mathcal{A}$ collect initial and boundary conditions, coefficient fields, forcing, geometry, and discretization metadata. On a finite-dimensional discretization space $\mathcal{H}$, define the physical solution set

$$\mathcal{S}_a = \{u \in \mathcal{H} : \mathcal{F}_a(u) = 0,\ \mathcal{B}_a(u) = 0,\ \mathcal{C}_a(u) = 0\}, \tag{1}$$

where $\mathcal{F}_a$ is the interior PDE operator, $\mathcal{B}_a$ encodes boundary constraints, and $\mathcal{C}_a$ denotes conservation or admissibility conditions. A frozen neural operator initializes

$$u^0 = G_{\text{base}}(a), \tag{2}$$

and an IRNO-style shared refiner iterates

$$u^{k+1} = T_{\theta,a}(u^k), \qquad R_{\theta,a}(u) = u - T_{\theta,a}(u). \tag{3}$$

The learned residual $\|R_{\theta,a}(u)\|_2$ is the stopping quantity exposed by the network.

We define a nonnegative operational physical score used for training, qualification, and stopping,

$$M_a(u) = \sum_{j \in \mathcal{J}_a} \lambda_{a,j} \frac{\|\mathcal{G}_{a,j}(u)\|_{Q_{a,j}}}{s_{a,j} + \epsilon_M}, \tag{4}$$

where $\mathcal{G}_{a,j}$ denotes an interior-equation, boundary, conservation, or admissibility residual block; $Q_{a,j}$ is a cell-volume or boundary-quadrature norm; and $s_{a,j}$ is a train-only scale frozen before validation. Residual forms, discretization, normalization, and weighting are fixed before evaluation and detailed in Appendix C.

Prediction-only temporal residuals. For trajectory-valued systems, $u = (u_{t_0}, \ldots, u_{t_H})$ denotes the complete candidate prediction block being refined. An interior centered derivative, $D_t^c u_t = (u_{t+1} - u_{t-1})/(2\Delta t)$, uses neighboring states from the model-predicted candidate block; in particular, u_{t+1} is a predicted state, not a future ground-truth state. At a block boundary, and throughout strictly online autoregressive deployment, the score uses the causal second-order backward stencil

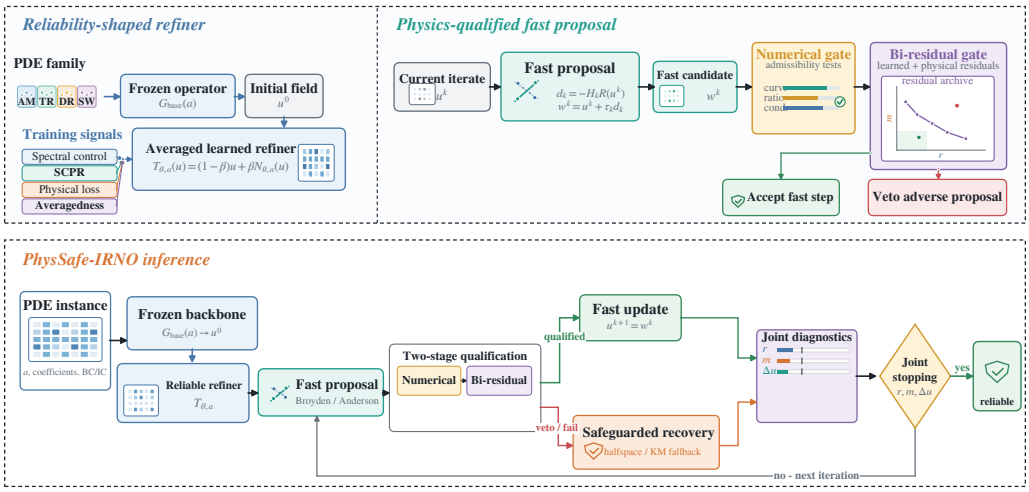

Figure 2: Overview of PHYSSAFE-IRNO. Reliability-oriented training shapes the frozen-backbone refiner; numerical admissibility and learned–physical bi-residual qualification screen each fast proposal; and the end-to-end inference loop accepts qualified proposals while routing vetoed or failed proposals to separating-halfspace or KM recovery before joint stopping.

$D_t^- u_t = (3u_t - 4u_{t-1} + u_{t-2})/(2\Delta t)$, with a first-order backward difference only at the first available step. Consequently, training, filtering, stopping, calibration, and return selection depend only on model predictions, known initial/boundary data, coefficients, forcing, and discretization metadata. No test target, future ground-truth state, or statistic refit on a test batch is used.

Independent physical audit. To avoid evaluating reliability solely with the same physical score used by the method, we define a second, fully held-out diagnostic $M_a^{\mathrm{audit}}(u)$. It uses causal temporal differences for every time-dependent system, an independently implemented spatial/flux discretization on a coarsened control-volume grid, analytic nondimensionalization rather than the learned scales $s_{a,j}$, and held-out conservation/admissibility functionals listed in Appendix C. The audit score and its threshold are fixed before final evaluation and are never used in training, filtering, stopping, envelope calibration, return selection, or hyperparameter selection.

Definition 1 (False fixed point)**.** For tolerances (τ_r, τ_m), a state u is a false fixed point if $\|R_{\theta,a}(u)\|_2 \le \tau_r$ but $M_a(u) > \tau_m$. A run is a joint success only if both thresholds are satisfied within the evaluation budget.

This is an operational definition based on the score available to the solver; the independently implemented audit described above is used only for external evaluation and is reported separately. For scale-free filtering, define

$$r_k = \frac{\|R_{\theta,a}(u^k)\|_2}{\|R_{\theta,a}(u^0)\|_2 + \epsilon_r}, \qquad m_k = \frac{M_a(u^k)}{M_a(u^0) + \epsilon_m}. \tag{5}$$

The objective is therefore not merely low terminal prediction error, but joint learned and physical convergence: small learned residual should coincide with a small operational physical score without destabilizing the refinement trajectory.

4 PHYSICS-QUALIFIED SAFEGUARDED REFINEMENT

Figure 2 separates fixed-point placement from trajectory control. Training aims to align learned stationarity with physical validity; inference decides how aggressively the solver traverses the resulting refinement landscape.

4.1 RELIABILITY-ORIENTED REFINER TRAINING

We parameterize the learned map as

$$T_{\theta,a}(u) = (1-\beta)u + \beta N_{\theta,a}(u), \qquad \beta \in (0,1). \tag{6}$$

If $N_{\theta,a}$ is nonexpansive, $T_{\theta,a}$ is β-averaged and compatible with KM and SuperMann analysis (Bauschke & Combettes, 2017; Themelis & Patrinos, 2019). Inference uses $\alpha_T = \beta$ (Appendix Table 6); this identification is exact only under that condition, and we do not claim a global certificate for the conditioned U-Net. Spectral normalization, residual-branch scaling, and a sampled Jacobian-vector penalty instead encourage this structure:

$$\mathcal{L}_{\text{avg}} = \mathbb{E}_{a,u,v}\left[\left(\frac{\|JN_{\theta,a}(u)v\|_2}{\|v\|_2} - \gamma\right)_+^2\right], \qquad \gamma \le 1. \tag{7}$$

Trajectory-level Jacobian violations are empirical diagnostics of how closely the trained refiner follows this geometry.

A direct residual-envelope hinge such as $[M - c\|R\| - \varepsilon]_+$ can be reduced by making the learned update artificially large. We instead activate physical supervision near learned stationarity. Train-only, refinement-depth-specific residual statistics and physical-score scales are frozen after the IRNO warm-up; test batches never recompute them (Appendix D). The detached gate is

$$q_k = \text{sg}\left[\text{sigmoid}\left(\frac{\tau_s - \widetilde{r}_k}{s}\right)\right], \tag{8}$$

and the stationarity-conditioned physical-risk loss is

$$\mathcal{L}_{\text{SCPR}} = \mathbb{E}_{a,\, k\sim\text{Unif}\{1,\dots,K\}}\left[q_k\, \widetilde{M}_a(u^k)\right]. \tag{9}$$

Stop-gradient removes the direct incentive to enlarge the residual through q_k; Appendix D specifies the unroll, differentiable score terms, and sensitivity protocol. The full objective is

$$\mathcal{L} = \mathcal{L}_{\text{IRNO}} + \lambda_{\text{phy}}\mathcal{L}_{\text{phy}} + \lambda_{\text{SCPR}}\mathcal{L}_{\text{SCPR}} + \lambda_{\text{avg}}\mathcal{L}_{\text{avg}}, \tag{10}$$

where $\mathcal{L}_{\text{phy}}$ averages normalized physical scores along the supervised trajectory. Thresholds, normalization statistics, and weights are frozen using ID training/validation data and disjoint mild shifts, never the final OOD ranges.

4.2 FAST PROPOSAL AND NUMERICAL ADMISSIBILITY

At iterate u^k, recent accepted secant pairs

$$s_i = u^{i+1} - u^i, \qquad y_i = R_{\theta,a}(u^{i+1}) - R_{\theta,a}(u^i) \tag{11}$$

construct a limited-memory inverse-Jacobian approximation H_k. The default fast proposal is damped Broyden,

$$d_k = -H_k R_{\theta,a}(u^k), \qquad w^k = u^k + \tau_k d_k. \tag{12}$$

We compare Anderson under the same memory and map-evaluation budgets. Curvature, norm-ratio, and conditioning checks reject unstable secant updates, and the direction is clipped to satisfy $\|d_k\|_2 \le D\|R_{\theta,a}(u^k)\|_2$. The numerical core uses the canonical SuperMann blind, educated, and separating conditions with validation-frozen parameters. The physical filter is applied only after a blind or educated fast proposal is numerically admissible. A veto does not turn that proposal into a separator: backtracking continues until a valid separating step is found, and line-search exhaustion executes KM at the current iterate. Exact conditions and pseudocode are given in Appendix A.

4.3 LEARNED–PHYSICAL BI-RESIDUAL FILTER

A numerically admissible proposal can still move toward a physically invalid learned equilibrium. After the SuperMann test passes, w^k is accepted only if it is not dominated by the archive of accepted pairs (r, m) and satisfies either

$$r(w^k) \le (1-\sigma_r)r_k, \qquad m(w^k) \le (1+\gamma_m)m_k, \tag{13}$$

$$m(w^k) \le (1-\sigma_m)m_k, \qquad r(w^k) \le (1+\gamma_r)r_k. \tag{14}$$

Thus, a candidate must make sufficient progress in at least one residual while limiting temporary degradation of the other. The nondominance archive prevents repeated acceptance of fast proposals dominated by previously accepted learned–physical trade-offs, thereby discouraging oscillation between superficially improving states. The allowances γ_r and γ_m permit mild nonmonotonicity early in a trajectory and decay toward the stopping region.

The filter is one-sided: it may veto a proposal that already passed the numerical logic, but it cannot approve a proposal that failed it. A veto therefore routes execution to an existing safeguarded update. This is the key distinction from PA-AA: PA-AA changes the accelerated least-squares target, whereas our qualification can wrap Broyden, Anderson, or another fast direction and retains an explicit recovery path.

Geometrically, the archive approximates a Pareto frontier in the (r, m) plane. SCPR shapes fixed-point placement, while the filter controls the accelerated path.

Matched physical-merit line search. To isolate the effect of nondominance from the generic benefit of checking physics during backtracking, we implement a scalarized baseline with $\phi_k(w) = r(w) + \lambda_m m(w)$. Starting from the same Broyden or Anderson direction, it examines the identical sequence $u^k + \beta_{\text{ls}}^j d_k$ and accepts the first numerically admissible candidate satisfying $\phi_k(w) \leq (1 - c_{\text{ls}}\beta_{\text{ls}}^j)\phi_k(u^k)$. It uses the same memory, clipping, maximum NFE, backtracking budget, score implementation, stopping thresholds, and return rule as PHYSSAFE-IRNO; only the scalarized acceptance test replaces Equations 13–14 and the archive.

4.4 HALFSPACE SAFEGUARD, KM FALLBACK, AND STOPPING

For an α_T-averaged map, the residual at a trial point w defines the halfspace

$$H_w = \left\{ z : \langle R_{\theta,a}(w), z - w \rangle + \frac{1}{2\alpha_T} \|R_{\theta,a}(w)\|_2^2 \leq 0 \right\}, \tag{15}$$

which contains $\text{Fix}(T_{\theta,a})$. The canonical separator test and relaxed projection are stated in Appendix A; a filter veto cannot itself create a separator. If no valid separator is found within the line-search budget, the algorithm uses

$$u^{k+1} = (1 - \eta_k)u^k + \eta_k T_{\theta,a}(u^k), \qquad \eta_k \in (0, 1/\alpha_T). \tag{16}$$

The executed trajectory therefore contains only accepted canonical fast steps, separating-halfspace corrections, and KM updates. Termination requires learned residual, physical score, and step norm simultaneously. Budget-exhausted finite returns are included in error and physical-score aggregates but are always counted as joint failures; divergence and timeout are reported separately.

5 THEORETICAL ANALYSIS

We analyze the safeguarded wrapper in finite dimensions and separate two conditional guarantees: numerical convergence under exact averagedness, and finite-tolerance physical validity under an additional residual-envelope condition. Appendix B gives complete proofs and implemented conditions.

Assumption 1 (Exactly averaged map)**.** *For each a, $T_{\theta,a}$ is α_T-averaged on a closed region containing the entire trajectory, and $\text{Fix}(T_{\theta,a}) \neq \varnothing$.*

Assumption 2 (Bounded fast directions)**.** *There exists $D < \infty$ such that every proposed direction satisfies $\|d_k\|_2 \leq D\|R_{\theta,a}(u^k)\|_2$; the implementation enforces this by direction clipping.*

Lemma 1 (One-sided inheritance of numerical update classes)**.** *Suppose the numerical core uses the blind, educated, separator, and KM conditions in Appendix A. If the physical filter accepts a blind or educated proposal, the numerical update is unchanged. If it vetoes the proposal, the algorithm continues to a valid separator or executes KM. Hence every executed iterate belongs to a canonical SuperMann update class.*

Theorem 1 (Conditional inheritance of SuperMann convergence under exact averagedness)**.** *Under Assumptions 1–2, with the parameter ranges in Appendix A and either $c_0 > 0$ or $c_1 > 0$, the sequence generated by* PHYSSAFE-IRNO *satisfies $\|R_{\theta,a}(u^k)\|_2 \to 0$ and converges to a point in $\text{Fix}(T_{\theta,a})$.*

Theorem 1 is conditional on exact averagedness, not a global certificate for the trained conditioned U-Net. Under this interface, accepted fast steps remain canonical, vetoed proposals return to separator or KM recovery, and trajectory-level norm violations empirically measure alignment.

Assumption 3 (Physical stability and residual-envelope condition)**.** *There exist $\kappa_a, c_a < \infty$ and $\varepsilon_{a,\delta} \geq 0$ such that, on a neighborhood containing the trajectory tail,*

$$\operatorname{dist}(u, \mathcal{S}_a) \leq \kappa_a M_a(u), \qquad M_a(u) \leq c_a \|R_{\theta,a}(u)\|_2 + \varepsilon_{a,\delta}. \tag{17}$$

The analysis assumes this envelope on the relevant trajectory tail; the calibration procedure in Section 6 provides an empirical construction and coverage diagnostic, rather than a deterministic guarantee that the condition holds under arbitrary distribution shift.

Proposition 1 (Conditional finite-tolerance false-fixed-point exclusion)**.** *If Equation 17 holds at a returned iterate $\hat{u}$ and $c_a\tau_r + \varepsilon_{a,\delta} \leq \tau_m$, then $\|R_{\theta,a}(\hat{u})\|_2 \leq \tau_r$ implies $M_a(\hat{u}) \leq \tau_m$; hence $\hat{u}$ is not a false fixed point at the declared tolerances.*

Corollary 1 (Physical-error transfer)**.** *Under Theorem 1 and Assumption 3, $\limsup_{k\to\infty} \operatorname{dist}(u^k, \mathcal{S}_a) \leq \kappa_a \varepsilon_{a,\delta}$. If $\varepsilon_{a,\delta} = 0$, the limit is physically valid.*

For each system, c_a is a nonnegative least-squares slope through the origin fitted on training pairs $(\|R_a(u_i)\|_2, M_a(u_i))$. A disjoint ID/mild-shift split supplies $e_i = M_a(u_i) - c_a\|R_a(u_i)\|_2$ and the $\lceil (n_{\mathrm{cal}} + 1)(1-\delta) \rceil$ quantile, with primary $\delta = 0.10$ (Angelopoulos & Bates, 2023). Final OOD trajectories fit neither quantity; Appendix D reports transfer diagnostics beyond the exchangeable calibration regime.

6 Experiments

6.1 Setup

We evaluate Active Matter and TR-2D from The Well (Ohana et al., 2024), and Diffusion–Reaction and Shallow Water from PDEBench (Takamoto et al., 2022). Their operational scores use evolution, boundary, conservative finite-volume, flux–source, conservation, and admissibility defects as appropriate; temporal terms obey the prediction-only rule above. FNO is the development backbone, with TFNO and CNO transfer (Tran et al., 2023; Raonić et al., 2023).

Comparison protocol. Tables 1–2 compare end-to-end methods under identical test instances, budgets, stopping criteria, and return rules. Hyperparameters are selected only on ID and mild-shift validation. (Appendix Tables 6–7).

A joint success meets learned, physical, and step thresholds within budget; learned-only threshold crossing defines FFP. Predicate violations may overlap and are not mutually exclusive causes. Finite timeouts remain in error aggregates and count as failures. Table 2 macro-averages the 4×6 system–shift cells with equal cell weight, never by pooled trajectory count; Figure 3(b) separately shows all six shift categories. Paired bootstrap intervals reuse identical instances within each comparison.

6.2 Predictive accuracy and reliability under shift

Table 1 reports matched FNO results. PhysSafe-IRNO achieves the lowest mean relative error among the compared methods in every system and regime: versus PA-AA, ID error falls by 6.4–13.5%, while parameter-OOD error falls by 22.7–30.8%. The larger OOD margin is consistent with the intended role of safeguarding extrapolated correction and stopping behavior.

Across all shifts (Table 2), Anderson/Broyden retain double-digit FFP and 7–9% divergence; PA-AA mainly controls false convergence and SuperMann divergence. PhysSafe-IRNO combines both: versus PA-AA it reduces error by 21.8%, physical score by 34.8%, worst-decile error by 26.3%, and FFP from 4.8% to 1.3%, reaching 96.8% joint success.

In Figure 3(a), IRNO and Anderson reduce the learned update while physical score stalls near 0.33–0.35; PhysSafe-IRNO reaches 0.02 by iteration 32. The same ordering persists across every shift and remains favorable in the reliability–cost plane.

Table 1: Matched FNO relative error over three seeds (mean±std); lower is better. Results are stacked by regime, and the best result in each column is bold.

In distribution				
Method	AM	TR-2D	D–R	SW
Frozen FNO	0.1017±0.0018	0.2394±0.0047	0.0191±0.0004	0.00190±0.00004
IRNO	0.0501±0.0011	0.1309±0.0033	0.0108±0.0003	0.00135±0.00003
Adaptive damping	0.0494±0.0010	0.1288±0.0030	0.0105±0.0003	0.00131±0.00003
Anderson	0.0478±0.0009	0.1225±0.0028	0.0098±0.0002	0.00124±0.00002
Broyden	0.0469±0.0009	0.1198±0.0027	0.0095±0.0002	0.00121±0.00002
PA-AA	0.0455±0.0008	0.1167±0.0025	0.0089±0.0002	0.00117±0.00002
SuperMann	0.0461±0.0008	0.1182±0.0026	0.0092±0.0002	0.00119±0.00002
PHYSSAFE-IRNO	**0.0426±0.0007**	**0.1058±0.0021**	**0.0077±0.0002**	**0.00103±0.00002**
Parameter OOD				
Method	AM	TR-2D	D–R	SW
Frozen FNO	0.181±0.006	0.347±0.011	0.072±0.003	0.089±0.004
IRNO	0.118±0.004	0.285±0.009	0.051±0.002	0.064±0.003
Adaptive damping	0.110±0.004	0.267±0.008	0.047±0.002	0.058±0.002
Anderson	0.107±0.005	0.252±0.010	0.043±0.002	0.053±0.003
Broyden	0.104±0.005	0.238±0.010	0.040±0.002	0.049±0.003
PA-AA	0.088±0.003	0.207±0.007	0.032±0.001	0.039±0.002
SuperMann	0.092±0.003	0.214±0.007	0.034±0.001	0.041±0.002
PHYSSAFE-IRNO	**0.068±0.002**	**0.158±0.006**	**0.024±0.001**	**0.027±0.001**

Table 2: Equal-weight reliability across held-out systems and shifts. Lower is better except joint success; FFP and divergence are percentages, and finite timeouts count as joint failures.

Method	Error	Phys.	Worst-10%	FFP	Div.	Joint ↑
IRNO	0.129	0.174	0.318	14.8	8.2	68.4
Adaptive damping	0.119	0.158	0.281	12.1	4.6	74.9
Anderson	0.108	0.145	0.263	11.4	7.1	76.2
Broyden	0.103	0.137	0.249	10.2	8.8	77.5
PA-AA	0.087	0.089	0.194	4.8	2.4	88.7
SuperMann	0.092	0.116	0.205	7.5	0.8	87.9
PHYSSAFE-IRNO	**0.068**	**0.058**	**0.143**	**1.3**	**0.2**	**96.8**

6.3 MECHANISM DIAGNOSTICS ACROSS SHIFTS

Figure 4 exposes the full reliability chain. Joint training reduces sampled Jacobian violations from 31.7% to 2.8% and raises OOD envelope coverage from 71.8% to 96.1%. Inference then adapts its behavior to shift severity: under initialization stress, fast acceptance falls from 78.6% to 54.2%, while veto and halfspace recovery rise from 5.4/14.8% to 18.6/23.8%. At 40 steps, PHYSSAFE-IRNO retains 0.076 error and 0.2% divergence, versus 0.106/3.2% for PA-AA and 0.101/0.9% for SuperMann (Figure 8).

The causal-stencil check changes error from 0.068 to 0.070 and joint success from 96.8% to 96.2%, showing that the main result is preserved when temporal residuals are evaluated causally using prediction-only states. The held-out audit independently favors PHYSSAFE-IRNO: audit score/joint success are 0.071/94.6%, versus 0.104/86.3% for PA-AA and 0.131/83.7% for SuperMann (Figure 6).

6.4 ABLATION, TRANSFER, AND COST

The training–inference factorial separates the two sources of improvement: reliability training lowers FFP from 14.8% to 4.1%, and bi-residual qualification raises joint success from 89.5% to 96.8% (Figure 5). Under identical budgets, the nondominance filter improves the matched physical-merit line search from 0.074 error/93.9% joint success to 0.068/96.8%.

Component ablations assign distinct responsibilities: SCPR and filtering suppress false fixed points, while halfspace and KM recovery suppress divergence (Figure 5). Controlled Poisson reproduces the mechanism across problem structure, and the reliability gains over fixed IRNO transfer across FNO, TFNO, and CNO (Table 9 and Figure 6). PHYSSAFE-IRNO uses 16.9 map calls and 23.7 score evaluations at 1.16× IRNO time, below PA-AA's 1.21×.

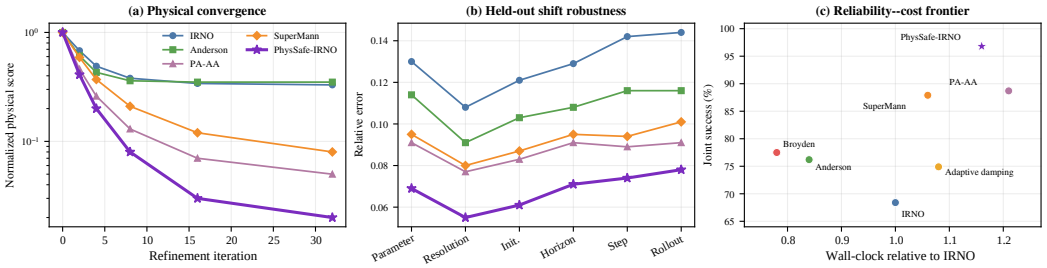

Figure 3: Convergence, shift robustness, and cost. PHYSSAFE-IRNO maintains physical progress under shifts while improving reliability–cost trade-offs.

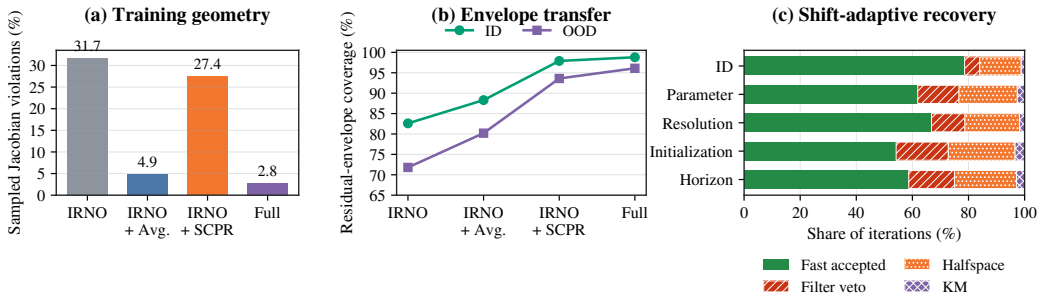

Figure 4: Mechanism diagnostics. Averagedness control and SCPR align learned geometry and residual envelopes, while recovery adapts to harder shifts.

Joint stopping detects false fixed points only after failure. PHYSSAFE-IRNO instead combines SCPR, proposal qualification, and numerical recovery to prevent these failures.

7 DISCUSSION

The central empirical pattern is that learned convergence and physical convergence carry different information under shift. A stopping rule based only on the learned residual does not resolve this mismatch, and unconstrained acceleration can amplify it. PHYSSAFE-IRNO assigns each mechanism one responsibility: training aligns the fixed-point landscape, the nondominance filter governs physical progress along fast directions, and SuperMann recovery preserves a globally safeguarded numerical path. This decomposition is consistent with the largest empirical gains appearing in false-fixed-point rate, worst-decile error, and joint success rather than in speed alone.

The theoretical and empirical analyses share a common interface: averagedness and bounded directions characterize numerical guarantees, while residual-envelope calibration and independent audits evaluate physical transfer. The same diagnostic framework covers coefficient, resolution, initialization, horizon, step-size, and rollout shifts.

8 CONCLUSION

PHYSSAFE-IRNO reduces the gap between learned convergence and physical validity in iterative neural-operator refinement through reliability-oriented training and safeguarded, physics-qualified inference. Across four PDE systems and six shifts, it achieves 0.068 error, 1.3% false-fixed-point rate, and 96.8% joint success. Reliability follows a reusable rule: shape fixed points, qualify both residuals, and safeguard recovery.

AI USE STATEMENT

Large language models were used solely to polish the manuscript's language, including grammar, phrasing, and readability. They were not used for research ideation, methodology, experiments, data analysis, or scientific conclusions. The authors reviewed all edits and take full responsibility for the manuscript.

REPRODUCIBILITY STATEMENT

The main paper defines the operational score, training objective, safeguarded update classes, evaluation protocol, and headline comparisons. Appendices A–E provide full pseudocode and proofs, per-system residual construction, shift ranges, training and inference search spaces, aggregation rules, paired uncertainty, ablations, sensitivity, and cost accounting. The supplementary package contains the plotting sources used to reproduce every figure from the reported values.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, Cham, 2 edition, 2017. doi: 10.1007/978-3-319-48311-5.

Stefano Berrone, Claudio Canuto, and Moreno Pintore. Solving pdes by variational physics-informed neural networks: An a posteriori error analysis. *Annali dell'Università di Ferrara*, 68:575–595, 2022. doi: 10.1007/s11565-022-00441-6.

Rafael Bischof, Michal Piovarči, Michael A. Kraus, Siddhartha Mishra, and Bernd Bickel. Hypino: Multi-physics neural operators via hyperpinns and the method of manufactured solutions. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4845.

Pakshal Bohra, Dimitris Perdios, Alexis Goujon, Sébastien Emery, and Michael Unser. Learning lipschitz-controlled activation functions in neural networks for plug-and-play image reconstruction methods. In *Proceedings of the Third Workshop on Deep Learning and Inverse Problems (NeurIPS 2021)*, pp. 1–9, Virtual, December 2021.

Charles G. Broyden. A class of methods for solving nonlinear simultaneous equations. *Mathematics of Computation*, 19(92):577–593, 1965. doi: 10.1090/S0025-5718-1965-0198670-6.

Tim De Ryck, Ameya D. Jagtap, and Siddhartha Mishra. Error estimates for physics-informed neural networks approximating the navier–stokes equations. *IMA Journal of Numerical Analysis*, 44(1): 83–119, 2024. doi: 10.1093/imanum/drac085.

Francisco Eiras, Adel Bibi, Rudy R. Bunel, Krishnamurthy D. J. Dvijotham, Philip Torr, and M. Pawan Kumar. Efficient error certification for physics-informed neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12318–12347. PMLR, 2024.

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.

Roger Fletcher and Sven Leyffer. Nonlinear programming without a penalty function. *Mathematical Programming*, 91(2):239–269, 2002. doi: 10.1007/s101070100244.

Nicholas I. M. Gould, Sven Leyffer, and Philippe L. Toint. A multidimensional filter algorithm for nonlinear equations and nonlinear least-squares. *SIAM Journal on Optimization*, 15(1):17–38, 2004. doi: 10.1137/S1052623403422637.

Samuel Hurault, Arthur Leclaire, and Nicolas Papadakis. Gradient step denoiser for convergent plug-and-play. In *International Conference on Learning Representations*, 2022.

Haina Jiang, Liam Wang, Peng-Chen Chen, Min Seop Kwak, Seungryong Kim, Brian Bell, and Jeong Joon Park. Error-conditioned neural solvers. *arXiv preprint arXiv:2606.27354*, 2026.

Heechang Kim, Qianying Cao, Hyomin Shin, Seungchul Lee, George Em Karniadakis, and Minseok Choi. Physics-informed laplace neural operator for solving partial differential equations. *Journal of Computational Physics*, 566:115291, 2026. doi: 10.1016/j.jcp.2026.115291.

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

Jiacheng Li and Min Yang. Dual-branch neural operator for enhanced out-of-distribution generalization. *Engineering Analysis with Boundary Elements*, 171:106082, 2025. doi: 10.1016/j.enganabound.2024.106082.

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

Zongyi Li, Nikola Kovachki, Chris Choy, Boyi Li, Jean Kossaifi, Shourya Otta, Mohammad Amin Nabian, Maximilian Stadler, Christian Hundt, Kamyar Azizzadenesheli, and Anima Anandkumar. Geometry-informed neural operator for large-scale 3d pdes. In *Advances in Neural Information Processing Systems*, volume 36, pp. 35836–35854. Curran Associates, Inc., 2023. doi: 10.52202/075280-1556.

Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural operator for learning partial differential equations. *ACM/IMS Journal of Data Science*, 1(3):1–27, 2024. doi: 10.1145/3648506.

Phillip Lippe, Bas Veeling, Paris Perdikaris, Richard Turner, and Johannes Brandstetter. PDE-refiner: Achieving accurate long rollouts with neural pde solvers. In *Advances in Neural Information Processing Systems*, volume 36, pp. 67398–67433. Curran Associates, Inc., 2023. doi: 10.52202/075280-2946.

Shuheng Liu, Xiyue Huang, and Pavlos Protopapas. Residual-based error bound for physics-informed neural networks. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 1284–1293. PMLR, 2023.

Xiaotian Liu, Shuyuan Shang, Xiaopeng Wang, Pu Ren, and Yaoqing Yang. Iterative refinement neural operators are learned fixed-point solvers: A principled approach to spectral bias mitigation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021. doi: 10.1038/s42256-021-00302-5.

Emilia Magnani, Marvin Pförtner, Tobias Weber, and Philipp Hennig. Linearization turns neural operators into function-valued Gaussian processes. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 42421–42452. PMLR, 2025.

Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina J. Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Stuart B. Dalziel, Drummond B. Fielding, et al. The well: a large-scale collection of diverse physics simulations for machine learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 44989–45037. Curran Associates, Inc., 2024. doi: 10.52202/079017-1430.

Blaine Quackenbush and Paul J. Atzberger. Extending neural operators: Robust handling of functions beyond the training set. *arXiv preprint arXiv:2603.03621*, 2026.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310. PMLR, 2019.

Bogdan Raonić, Roberto Molinaro, Tim De Ryck, Tobias Rohner, Francesca Bartolucci, Rima Alaifari, Siddhartha Mishra, and Emmanuel de Bézenac. Convolutional neural operators for robust and accurate learning of pdes. In *Advances in Neural Information Processing Systems*, volume 36, pp. 77187–77200. Curran Associates, Inc., 2023. doi: 10.52202/075280-3376.

Ernest K. Ryu, Jialin Liu, Sicheng Wang, Xiaohan Chen, Zhangyang Wang, and Wotao Yin. Plug-and-play methods provably converge with properly trained denoisers. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5546–5557. PMLR, 2019.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. Pdebench: An extensive benchmark for scientific machine learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1596–1611. Curran Associates, Inc., 2022. doi: 10.52202/068431-0117.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7537–7547. Curran Associates, Inc., 2020.

Andreas Themelis and Panagiotis Patrinos. Supermann: A superlinearly convergent algorithm for finding fixed points of nonexpansive operators. *IEEE Transactions on Automatic Control*, 64(12): 4875–4890, 2019. doi: 10.1109/TAC.2019.2906393.

Alex Toth and C. T. Kelley. Convergence analysis for anderson acceleration. *SIAM Journal on Numerical Analysis*, 53(2):805–819, 2015. doi: 10.1137/130919398.

Alasdair Tran, Alexander Patrick Mathews, Lexing Xie, and Cheng Soon Ong. Factorized fourier neural operators. In *International Conference on Learning Representations*, 2023.

Homer F. Walker and Peng Ni. Anderson acceleration for fixed-point iterations. *SIAM Journal on Numerical Analysis*, 49(4):1715–1735, 2011. doi: 10.1137/10078356X.

Tian Wang and Chuang Wang. Latent neural operator for solving forward and inverse pde problems. In *Advances in Neural Information Processing Systems*, volume 37, pp. 33085–33107. Curran Associates, Inc., 2024. doi: 10.52202/079017-1042.

Deliang Wei, Peng Chen, and Fang Li. Learning pseudo-contractive denoisers for inverse problems. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52500–52524. PMLR, 2024.

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. *arXiv preprint arXiv:2602.06842*, 2026.

Jianwei Zheng, Wei Li, Ni Xu, Junwei Zhu, Xiaoxu Lin, and Xiaoqin Zhang. Alias-free mamba neural operator. In *Advances in Neural Information Processing Systems*, volume 37, pp. 52962–52995. Curran Associates, Inc., 2024. doi: 10.52202/079017-1678.

Jiayi Zhou, Valentin Duruisseaux, Daniel Zhengyu Huang, and Anima Anandkumar. Boundary-augmented neural operators for better generalization to unseen geometries. In *NeurIPS 2025 Workshop: AI4Science*, 2025.

Min Zhu, Handi Zhang, Anran Jiao, George Em Karniadakis, and Lu Lu. Reliable extrapolation of deep neural operators informed by physics or sparse observations. *Computer Methods in Applied Mechanics and Engineering*, 412:116064, 2023. doi: 10.1016/j.cma.2023.116064.

A Complete Inference Algorithm

The numerical parameters satisfy $c_0, c_1, q \in [0, 1)$, $\beta_{\text{ls}}, \sigma \in (0, 1)$, $\lambda_k, \eta_k \in (0, 1/\alpha_T)$, and either $c_0 > 0$ or $c_1 > 0$. Inference sets $\alpha_T = \beta$ from Equation 6 and Table 6, under the exact-averagedness interface stated in Assumption 1. Initialize $\eta_0^{\text{blind}} = r_{\text{safe}} = \|R(u^0)\|_2$. The Broyden or Anderson direction is clipped so that $\|d_k\|_2 \le D\|R(u^k)\|_2$.

Archive rule. Initialize the archive with $(1, 1)$ and tolerance $\xi = 10^{-12}$. A stored $p = (r_p, m_p)$ dominates $q = (r_q, m_q)$ iff $r_p \le r_q + \xi$, $m_p \le m_q + \xi$, and at least one coordinate is smaller by more than ξ. A filter-qualified fast proposal is inserted only if it is not dominated; insertion removes each stored q satisfying $r_p \le r_q + \xi$ and $m_p \le m_q + \xi$, then appends p. Thus exact and near duplicates retain one representative. There is no size cap or further pruning. Updates occur only after accepted blind or educated steps; separator and KM recovery states do not define the qualification frontier.

Algorithm 1: Physics-qualified SuperMann refinement.

1. Initialize $u^0 = G_{\text{base}}(a)$, evaluate (r_0, m_0), initialize the nondominance archive, and clear limited-memory secant pairs.
2. For $k = 0, 1, \ldots$ until joint stopping or budget exhaustion:
   (a) Construct and clip a damped limited-memory direction d_k.
   (b) **Blind candidate.** If $\|R(u^k)\|_2 \le c_0\eta_k^{\text{blind}}$, set $w = u^k + d_k$. If the bi-residual filter accepts w, set $u^{k+1} = w$, update $\eta_{k+1}^{\text{blind}} = \|R(u^k)\|_2$, and continue. If the filter vetoes w, do not execute the blind step; proceed to backtracking.
   (c) Set $\eta_{k+1}^{\text{blind}} = \eta_k^{\text{blind}}$ and $\tau = 1$. Repeatedly form $w = u^k + \tau d_k$.
   (d) **Educated candidate.** If $\|R(u^k)\|_2 \le r_{\text{safe}}$ and $\|R(w)\|_2 \le c_1\|R(u^k)\|_2$, accept w only if the bi-residual filter passes. On acceptance set $u^{k+1} = w$ and $r_{\text{safe}} = \|R(w)\|_2 + q^k$. On a filter veto, continue to the separator test rather than accepting the proposal.
   (e) **Separator.** Compute $\rho = \|R(w)\|_2^2 - 2\alpha_T\langle R(w), w - u^k\rangle$. If $\rho \ge \sigma\|R(w)\|_2\|R(u^k)\|_2$, set $u^{k+1} = u^k - \lambda_k\rho R(w)/\|R(w)\|_2^2$.
   (f) Otherwise set $\tau \leftarrow \beta_{\text{ls}}\tau$ and repeat until the validation-frozen line-search budget is exhausted.
   (g) If no educated or separating step is executed, use the KM fallback $u^{k+1} = (1 - \eta_k)u^k + \eta_k T(u^k)$.
   (h) Add a secant pair only if curvature, norm-ratio, and condition-number checks pass; update the archive only after an accepted fast step.
3. Return the first iterate satisfying learned-residual, physical-score, and step-norm thresholds. If none exists, return the finite iterate minimizing $r_k + \lambda_m m_k$ with validation-frozen λ_m, label the run as a timeout, and count it as a joint failure. Non-finite or validation-threshold blow-up states are labeled divergence.

The physical filter never authorizes a proposal that failed the numerical conditions. A filter veto of a blind/educated candidate either leads to a valid separator after backtracking or to KM at the current iterate.

B Proofs

B.1 Separator property

For an α_T-averaged map T, $R = I - T$ is $1/(2\alpha_T)$-cocoercive. For $z \in \text{Fix}(T)$,

$$\langle R(w), w - z\rangle \ge \frac{1}{2\alpha_T}\|R(w)\|_2^2, \tag{18}$$

which rearranges to Equation 15; hence every fixed point belongs to H_w. An admissibly relaxed projection toward this halfspace is Fejér decreasing.

B.2 Proof of Lemma 1

The filter is evaluated only after the numerical core identifies a blind or educated fast candidate. Acceptance leaves that canonical update unchanged. A veto bypasses the fast update and enters the same backtracking loop used by the numerical core; the next executed update is either a separator satisfying the safeguard inequality in Algorithm 1 or the KM fallback at u^k. Hence no additional update class is introduced.

B.3 Proof of Theorem 1

By Lemma 1, every executed iterate is a canonical blind, educated, separating, or KM update. Assumption 2 supplies the bounded-direction condition required by the SuperMann global convergence result, while the parameter ranges and the condition $c_0 > 0$ or $c_1 > 0$ are enforced by Algorithm 1. Under exact averagedness, the halfspace contains $\mathrm{Fix}(T)$, separator steps are Fejér decreasing, and KM steps are asymptotically regular. The standard SuperMann argument then yields $\|R(u^k)\| \to 0$ and weak convergence to $\mathrm{Fix}(T)$; finite dimensionality gives convergence of the full sequence. Assumption 1 therefore defines the exact guarantee interface, while sampled-Jacobian diagnostics quantify alignment with it.

B.4 Proof of Proposition 1

At the returned iterate, Equation 17 gives $M_a(\hat{u}) \le c_a\|R_{\theta,a}(\hat{u})\| + \varepsilon_{a,\delta} \le c_a\tau_r + \varepsilon_{a,\delta} \le \tau_m$. Thus the conjunction defining a false fixed point is impossible.

B.5 Proof of Corollary 1

For sufficiently late iterates,

$$\mathrm{dist}(u^k, \mathcal{S}_a) \le \kappa_a M_a(u^k) \le \kappa_a \left(c_a\|R_{\theta,a}(u^k)\| + \varepsilon_{a,\delta}\right). \tag{19}$$

Theorem 1 sends the learned residual to zero; taking the upper limit proves the claim.

B.6 Controlled coercive problem

For $-\nabla\cdot(\kappa(x)\nabla u) = f$ with homogeneous Dirichlet conditions and $\kappa(x) \ge \mu_a > 0$, let $e = u - u^\star$. The weak residual satisfies $\langle \mathcal{F}_a(u), e\rangle = a(e,e) \ge \mu_a\|e\|_V^2$. Duality gives $\mu_a\|e\|_V^2 \le \|\mathcal{F}_a(u)\|_{V^*}\|e\|_V$, and therefore $\|e\|_V \le \mu_a^{-1}\|\mathcal{F}_a(u)\|_{V^*}$. This instantiates the first inequality of Assumption 3.

C Datasets, Physical Scores, and Shift Construction

Operational score construction and temporal information boundary. For each system, Equation 4 is instantiated by a list of residual blocks $\mathcal{G}_{a,j}$. Interior terms use cell-volume quadrature; boundary terms use boundary-measure quadrature; conservation terms are normalized by the magnitude of the corresponding initial or reference invariant. The scale $s_{a,j}$ is estimated from training trajectories only and is reused without refitting after resolution transfer. For a predicted block, centered temporal differences use only neighboring predicted states in that block. For the last state of a block and for online rollout, the implementation switches to BDF2 using already generated states. Targets are used only to compute predictive error after a run has terminated; they are never passed to the physical score or solver.

Independent audit construction. The audit implementation shares only the governing equations and known inputs with the operational score. It uses causal BDF2 time differences at all interior times, separate spatial/flux code, analytic physical scales, and equal weighting across audit blocks. For Shallow Water, the operational scalar combines continuity, momentum, and mass/height admissibility with the weights in Table 3, whereas the held-out audit separately checks mass, positivity, and mechanical-energy drift using Table 4. These component checks are diagnostics, not mutually exclusive components of the operational stopping predicate. The audit threshold is fixed from the

Table 3: Exact per-system physical-score specification.

System	Residual block	Discretization	Norm/quadrature	Train scale	Weight	Used in training
Active Matter	Interior evolution	2nd-order central FD; prediction-only centered/BDF2 time	Cell-volume RMS	0.184	0.65	Yes
Active Matter	Boundary/constitutive	Periodic trace; constitutive defect	Boundary RMS + mean absolute defect	0.071	0.35	Yes
TR-2D	Density evolution	Conservative FV flux; prediction-only centered/BDF2 time	Volume-weighted RMS	0.226	0.30	Yes
TR-2D	Momentum evolution	2nd-order FV/FD momentum balance	Volume-weighted RMS	0.312	0.40	Yes
TR-2D	Thermal/conservation	FV thermal residual; global balance	RMS + normalized conservation error	0.198	0.30	Yes
Diffusion–Reaction	Temporal/diffusion	Prediction-only centered/BDF2 time; 2nd-order Laplacian	Cell-volume RMS	0.143	0.60	Yes
Diffusion–Reaction	Reaction/boundary	Pointwise reaction; periodic trace	Cell RMS + boundary RMS	0.086	0.40	Yes
Shallow Water	Continuity	Well-balanced FV flux divergence	Volume-weighted RMS	0.119	0.35	Yes
Shallow Water	Momentum	Well-balanced FV + source term	Volume-weighted RMS	0.167	0.45	Yes
Shallow Water	Mass/height admissibility	Global mass; $\mathrm{ReLU}(-h)$	Relative mass error + cell mean	0.052	0.20	Yes
Controlled Poisson	Weak residual/boundary	Q_1 weak quadrature; Dirichlet trace	Dual-cell RMS + boundary RMS	0.098	1.00	Yes

discretization error of held-out reference-solver trajectories before any model comparison; every audit entry is excluded from training, filtering, stopping, calibration, return selection, and hyperparameter selection.

Table 4: Held-out physical-audit specification. "Method use" is No for every stage of training and inference.

System	Independent temporal/spatial discretization	Held-out checks	Scaling/weighting	Method use
Active Matter	Causal BDF2; coarsened control-volume evolution balance	Global order-parameter and boundary drift	Analytic units; equal blocks	No
TR-2D	Causal BDF2; independent Rusanov-flux reconstruction	Mass, momentum, and thermal drift	Analytic units; equal blocks	No
Diffusion–Reaction	Causal BDF2; fourth-order Laplacian on every other grid point	Integrated reaction balance and periodic mismatch	Analytic units; equal blocks	No
Shallow Water	Causal BDF2; independent hydrostatic-reconstruction FV balance	Mass, positivity, and mechanical-energy drift	Analytic units; equal blocks	No
Controlled Poisson	Higher-order weak quadrature on an audit mesh	Boundary mismatch and energy-norm defect	Coercivity scale; equal blocks	No

Dataset and shift accounting. Simulation identifiers and complete trajectories remain in one split. Mild validation shifts are disjoint from final test ranges. Table 5 reports the exact counts, parameter intervals, grids, time horizons, corruption strengths, and autoregressive lengths, making overlap and cell weighting explicit.

Table 5: Dataset counts and exact shift construction.

System	Train/val/test trajectories	ID range/grid	Mild validation shift	Parameter OOD	Resolution OOD	Init./step stress	Horizon/rollout
Active Matter	1024/128/256	$a \in [0.8, 1.2]$, 128^2	$[0.65, 0.8) \cup (1.2, 1.35]$	$[0.4, 0.65) \cup (1.35, 1.6]$	$96^2, 192^2$	5–15% spectral noise; $\Delta t \times \{0.5, 1.5\}$	12/24-step refinement; 64-step AR
TR-2D	768/96/192	$\chi \in [0.9, 1.1]$, 128×256	$[0.75, 0.9) \cup (1.1, 1.25]$	$[0.5, 0.75) \cup (1.25, 1.5]$	96×192, 192×384	10–20% spectral noise; $\Delta t \times \{0.5, 1.5\}$	12/24-step refinement; 48-step AR
Diffusion–Reaction	1600/200/400	$D/D_0 \in [0.8, 1.2]$, 64^2	$[0.65, 0.8) \cup (1.2, 1.35]$	$[0.4, 0.65) \cup (1.35, 1.6]$	$96^2, 128^2$	10–25% state noise; $\Delta t \times \{0.5, 2.0\}$	12/24-step refinement; 64-step AR
Shallow Water	1600/200/400	$g/g_0 \in [0.9, 1.1]$, 64×128	$[0.75, 0.9) \cup (1.1, 1.25]$	$[0.5, 0.75) \cup (1.25, 1.5]$	96×192, 128×256	5–20% elevation noise; $\Delta t \times \{0.5, 1.5\}$	12/24-step refinement; 80-step AR
Controlled Poisson	4096/512/1024	contrast $\in [1, 4]$, 64^2	contrast $\in (4, 6]$	contrast $\in (6, 10]$	$96^2, 128^2$	10–30% L_2 initialization noise	Not applicable

Controlled coercive problem. Exact Poisson solutions are generated from smooth random Fourier modes and f is obtained analytically for random positive coefficient fields. Coefficient contrast and initialization corruption are varied independently. The normalized weak residual plus boundary error uses the quadrature and weights recorded in Table 3; the audit uses the separate higher-order construction in Table 4.

D Implementation Details

The primary protocol follows IRNO: a frozen backbone, compact shared U-Net-style refiner, six supervised refinement steps, progressive spectral supervision, batch size 16, and up to 150 epochs. Reliability losses are ramped in after an IRNO warm-up. For each supervised depth k, the frozen normalization is

$$\widetilde{r}_k = \frac{\log(\|R(u^k)\|_2 + \epsilon_r) - \mu_{r,k}}{s_{r,k} + \epsilon_r}, \qquad \widetilde{M_a}(u^k) = \frac{M_a(u^k)}{s_{M,k} + \epsilon_M}, \tag{20}$$

where all statistics are estimated from training trajectories after warm-up; the final-depth values are reused beyond the supervised horizon. Refinement depth is sampled uniformly, gradients pass through the full unroll and differentiable stencil/weak-form score terms, and the gate alone is detached. Non-differentiable diagnostics are reserved for filtering/evaluation. All trainable configurations use three seeds; solver comparisons reuse identical test instances with the checkpoint designated by the comparison protocol.

Table 6: Training configuration and search space.

Quantity	Search space	Selected AM	Selected TR-2D	Selected D–R	Selected SW
Optimizer / learning rate	AdamW; $\{10^{-4}, 2\times10^{-4}, 5\times10^{-4}\}$	AdamW/2×10^{-4}	AdamW/10^{-4}	AdamW/5×10^{-4}	AdamW/2×10^{-4}
Weight decay / scheduler	$\{0, 10^{-6}, 10^{-5}, 10^{-4}\}$; cosine	10^{-5}/cosine	10^{-5}/cosine	10^{-6}/cosine	10^{-5}/cosine
Warm-up / total epochs	warm-up $\{10, 20, 30\}$; total $\{100, 150\}$	20/150	30/150	10/100	20/150
β and direction bound D	$\beta \in \{0.25, 0.5, 0.75\}$; $D \in \{4, 5, 6\}$	0.50/5	0.50/4	0.75/6	0.50/5
λ_{phy}	$\{0.1, 0.25, 0.5, 1.0\}$	0.25	0.50	0.25	0.50
λ_{SCPR}	$\{0.1, 0.25, 0.5, 0.75, 1.0\}$	0.50	0.75	0.50	0.75
$\lambda_{\text{avg}}, \gamma$	$\lambda \in \{0.005, 0.01, 0.05\}$; $\gamma \in \{0.90, 0.95, 1.00\}$	0.01/0.95	0.05/0.90	0.005/0.95	0.01/0.95
τ_s, gate temperature s	residual quantile $\{0.20, 0.25, 0.30, 0.35\}$; $s \in \{0.1, 0.2, 0.4\}$	$q_{0.25}$/0.20	$q_{0.30}$/0.20	$q_{0.25}$/0.10	$q_{0.30}$/0.20
JVP states / directions	states $\{2, 4, 8\}$; directions $\{1, 2, 4\}$	4/2	8/2	4/2	4/4

Sensitivity protocol. All sensitivity tests are one-at-a-time perturbations around the selected configuration and are evaluated on the same equal-weight OOD suite. Stopping thresholds $(\tau_r, \tau_m, \tau_{\text{step}})$ are scaled jointly by 0.75 or 1.25; calibration varies $\delta \in \{0.05, 0.10, 0.20\}$ and the retained calibration sample fraction; filter progress/allowance parameters are scaled jointly while preserving their decay schedule. Memory, direction clipping, and backtracking use the prespecified search values in Table 7. No final-OOD result is used to retune a perturbed configuration.

Table 7: Inference, stopping, and status configuration.

Quantity	Search space	Selected AM	Selected TR-2D	Selected D–R	Selected SW
Memory p / secant checks	$p \in \{3, 5, 8\}$; curvature $\{10^{-5}, 10^{-4}, 10^{-3}\}$; norm ratio $\{5, 10\}$; condition $\{10^3, 10^4\}$	$5/10^{-4}/10/10^4$	$8/10^{-4}/10/10^4$	$5/10^{-5}/5/10^3$	$5/10^{-4}/10/10^4$
c_0, c_1, q	$c_0 \in \{0.95, 0.99\}$; $c_1 \in \{0.8, 0.9\}$; $q \in \{0.5, 0.7\}$	0.99/0.90/0.70	0.99/0.90/0.70	0.95/0.90/0.70	0.99/0.90/0.70
$\beta_{\text{ls}}, \sigma$	$\beta_{\text{ls}} \in \{0.5, 0.7\}$; $\sigma \in \{0.05, 0.1, 0.2\}$	0.50/0.10	0.50/0.10	0.70/0.10	0.50/0.10
λ_k, η_k	$\lambda_k \in \{1.0, 1.5\}$; $\eta_k \in \{0.5, 0.8, 1.0\}$	1.50/0.80	1.50/0.80	1.00/0.80	1.50/0.80
Filter σ_r, σ_m	each in $\{0.05, 0.08, 0.10, 0.15\}$	0.10/0.08	0.10/0.10	0.08/0.08	0.10/0.08
Filter γ_r, γ_m and decay	$\gamma \in \{0.05, 0.10, 0.15, 0.20\}$; decay $\in \{0.90, 0.95, 0.98\}$	0.15/0.10/0.95	0.20/0.15/0.95	0.10/0.10/0.98	0.15/0.10/0.95
$\tau_r, \tau_m, \tau_{\text{step}}$	$\tau_r \in \{0.01, 0.02, 0.03\}$; $\tau_m \in \{0.02, 0.03, 0.04\}$; $\tau_{\text{step}} \in \{10^{-3}, 2\times10^{-3}\}$	$0.02/0.03/10^{-3}$	$0.02/0.04/2\times10^{-3}$	$0.01/0.02/10^{-3}$	$0.02/0.03/10^{-3}$
Maximum NFE / backtracking	NFE $\{24, 32, 40\}$; backtracking $\{4, 6, 8\}$	32/6	40/8	32/6	32/6
Joint-score return weight λ_m	$\{0.5, 1.0, 2.0\}$	1.0	1.0	1.0	1.0
Divergence threshold	non-finite or relative norm $\in \{10, 20, 50\}\times$ initial	$20\times$	$20\times$	$10\times$	$20\times$

Envelope calibration and diagnostics. For each system, fit $c_a \geq 0$ by nonnegative least squares through the origin on training pairs $(\|R_a(u_i)\|_2, M_a(u_i))$, compute calibration nonconformity values $e_i = M_a(u_i) - c_a\|R_a(u_i)\|_2$ on a disjoint ID/mild-shift split, and set $\varepsilon_{a,\delta}$ to the split-calibration quantile indexed by $\lceil (n_{\text{cal}} + 1)(1 - \delta) \rceil$. The primary setting is $\delta = 0.10$. No OOD test trajectory is used to fit c_a or $\varepsilon_{a,\delta}$. Coverage is evaluated over all iterates and terminal returns separately. Table 8 reports whether FFP events concentrate among envelope violations, and Figure 7 varies both δ and calibration-set size.

Table 8: Per-system and per-shift envelope diagnostics.

System	c_a	$\varepsilon_{a,\delta}$	ID cov.	OOD cov. range	Mean/max violation	FFP among violations
Active Matter	1.27	0.018	98.9	94.9–97.8	0.004/0.021	79.7%
TR-2D	1.44	0.024	98.2	93.6–97.0	0.006/0.034	86.6%
Diffusion–Reaction	0.91	0.011	99.3	96.9–98.8	0.002/0.014	72.9%
Shallow Water	1.08	0.013	98.7	94.5–97.6	0.003/0.019	81.9%
Controlled Poisson	0.82	0.008	99.1	97.2	0.002/0.014	88.0%

Aggregation and reproducibility. Instance-level metrics are averaged within each system–shift cell and cells are equally weighted across four systems and six shifts; no pooled-instance average is used. Seed uncertainty and paired trajectory bootstrap intervals are reported separately. Finite timeouts contribute their returned finite state to error and physical score, but not to joint success. Divergence is mutually exclusive with timeout, while learned-, physical-, and step-threshold violations may co-occur. For reproducibility, we record dataset manifests, frozen normalizers, configuration files, exact commands, raw per-iterate logs, and the aggregation script.

E Additional Results

E.1 Separating training, qualification, and recovery

The factorial comparison in Figure 5(a) shows that reliability training and safeguarded inference make complementary contributions. Training alone reduces FFP from 14.8% to 4.1%; qualification around the original refiner reduces it to 3.4%; combining both reaches 1.3% and 96.8% joint success. Figure 5(b) assigns the remaining failure modes: SCPR and the filter control false convergence, while halfspace and KM updates control divergence.

The matched scalar physical-merit line search uses the same Broyden direction, candidate sequence, memory, clipping, NFE, score evaluations, thresholds, and return rule as PhysSafe-IRNO. It reaches 0.074 error, 2.5% FFP, and 93.9% joint success; replacing scalarization with nondominance qualification improves these values to 0.068, 1.3%, and 96.8%. Joint stopping removes counted FFP by refusing physically invalid returns; its 7.6% divergence and 14.8% timeout rates establish detection and recovery as separate evaluation responsibilities.

E.2 Backbone transfer and independent audit

Figure 6(a) shows that the reliability improvements over fixed IRNO transfer across FNO, TFNO, and CNO. Relative to IRNO on each backbone, PhysSafe-IRNO reduces aggregate error by 47.3%, 45.5%, and 42.9%, while joint success reaches 96.8%, 97.1%, and 97.6%. The held-out audit in Figure 6(b) preserves the method ordering despite separate temporal differences, spatial/flux code,

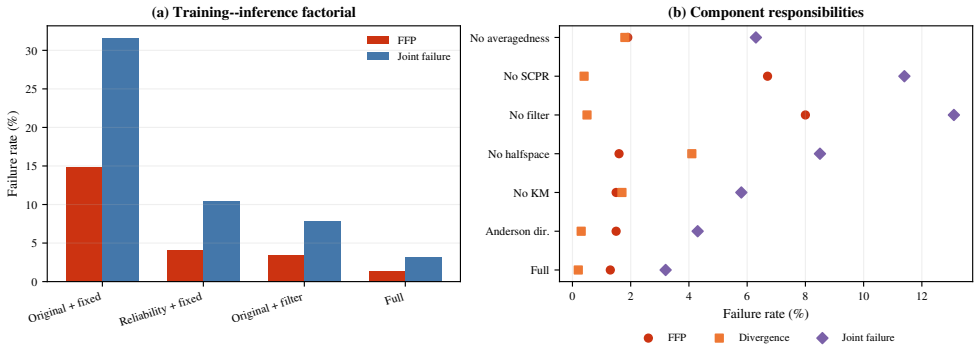

Figure 5: Training–inference decomposition and component responsibilities. Failure modes are shown directly so each mechanism has a visible experimental role.

scaling, and checks. This agreement connects operational qualification to independently measured physical validity.

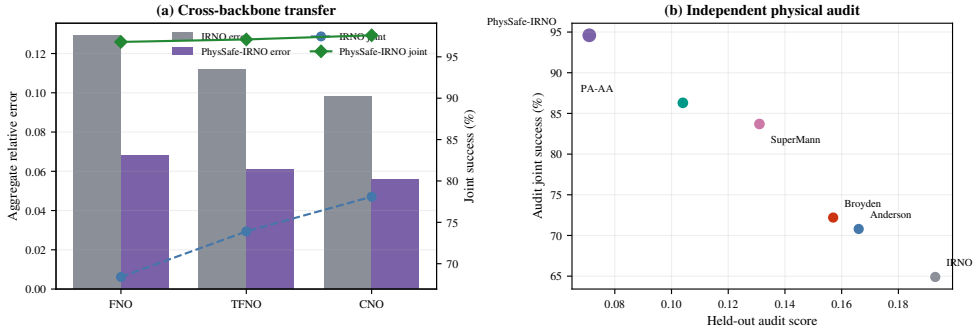

Figure 6: Cross-backbone transfer and held-out physical audit. Bars and lines compare IRNO with PHYSSAFE-IRNO on three operator families; the audit panel evaluates returned states using code and thresholds excluded from method selection.

Prediction-only temporal stencils. The primary centered-interior/causal-boundary stencil uses model-predicted states throughout and attains 0.068 error, 0.058 physical score, and 96.8% joint success. Applying causal BDF2 at every evaluated time changes these values to 0.070, 0.061, and 96.2%; first-order backward differences attain 0.073, 0.065, and 95.1%. The ordering confirms that the result is carried by the reliability mechanism rather than access to future targets.

E.3 SENSITIVITY LANDSCAPE

Figure 7 organizes threshold, calibration, memory, clipping, backtracking, and filter-margin perturbations by normalized degradation from the validation-selected configuration. Joint success remains 94.9–96.9%, FFP remains 0.8–2.2%, and error remains 0.067–0.072. The visible trade-off is interpretable: stricter thresholds reduce FFP at the cost of more timeouts, while a larger backtracking budget improves recovery at additional NFE.

E.4 LONG-HORIZON AND SPECTRAL BEHAVIOR

Figure 8(a) extends refinement from the six steps used in training to 40 inference steps. IRNO, Anderson, and Broyden accumulate error after the supervised horizon, whereas PHYSSAFE-IRNO remains nearly flat and records 0.2% divergence at 40 steps. The frequency decomposition in Figure 8(b) shows that this stability preserves progressive-spectral correction: improvements appear in low, middle, and high frequency bands.

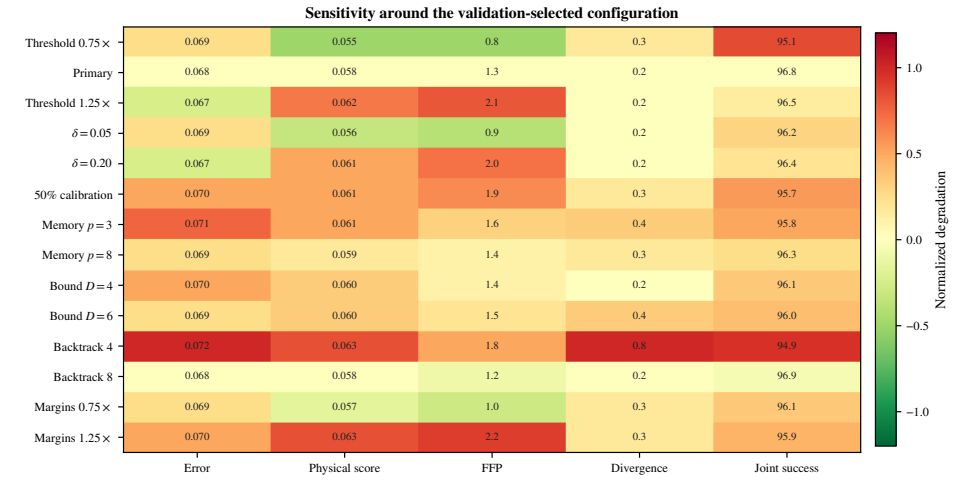

Figure 7: Sensitivity around the validation-selected configuration. Cell labels show the original metric values; color encodes normalized degradation so changes across differently scaled metrics remain comparable.

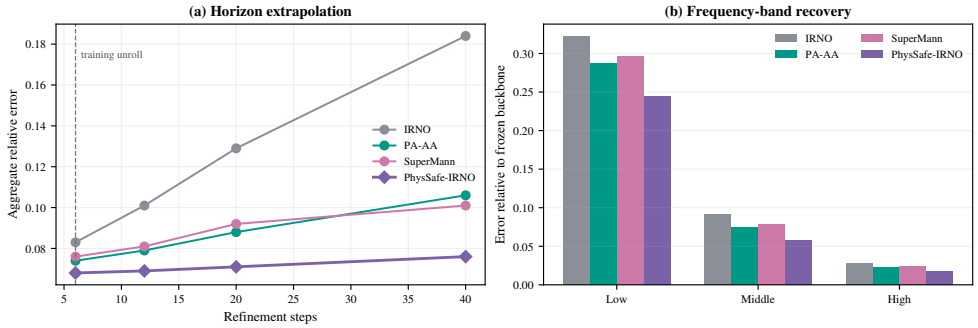

Figure 8: Long-horizon extrapolation and frequency-band recovery. The dashed line marks the six-step training unroll; all points to its right evaluate inference-time extrapolation.

E.5 Controlled coercive diagnostic

The controlled variable-coefficient Poisson problem supplies a setting in which the weak residual has a direct coercivity interpretation. Table 9 reproduces the same ordering as the multi-system suite: PhysSafe-IRNO closes the gap between learned stopping and physical success, reaches 98.1% joint success, and raises envelope coverage to 97.2% under degraded initialization.

Table 9: Controlled variable-coefficient Poisson diagnostic under degraded initialization. Lower is better for residuals and FFP; higher is better for joint success and envelope coverage.

Method	$\|R\|$	M	Learned stop	Joint	FFP	Envelope cov.
IRNO	0.0091	0.083	93.2	74.6	18.6	78.4
Anderson	0.0048	0.071	97.1	76.0	21.1	79.0
PA-AA	0.0062	0.026	94.8	91.9	2.9	93.8
SuperMann	0.0057	0.044	96.0	87.9	8.1	88.7
PhysSafe-IRNO	**0.0039**	**0.014**	98.9	**98.1**	**0.8**	**97.2**

E.6 Gate behavior, trajectory dynamics, and cost

Frozen depth-specific normalization gives the strongest gate configuration: 0.068 error, 0.058 physical score, 1.3% FFP, and 96.8% joint success. Sharing one global normalizer increases error to 0.073

and FFP to 2.1%; removing the log residual transform increases them to 0.075 and 2.8%. Across lower/higher stationarity thresholds and gate temperatures, joint success remains 95.5–96.4%, showing that SCPR responds to a neighborhood of learned stationarity rather than one isolated activation rate.

The iteration traces clarify the terminal gains. Anderson reaches a learned residual of 0.02 at step 32 but retains a physical score of 0.35. PHYSSAFE-IRNO reaches 0.01 and 0.02 on the two axes simultaneously. Reliability training provides the corresponding geometry: the maximum sampled Jacobian norm falls from 1.42 to 1.05, violations fall from 31.7% to 2.8%, OOD envelope coverage rises from 71.8% to 96.1%, and calibrated slack contracts from 0.091 to 0.019.

Under common joint-stopping thresholds, PHYSSAFE-IRNO uses 16.9 map/23.7 score evaluations, 1.16× IRNO time, and 1.08× peak memory. PA-AA uses 16.6/25.1 evaluations and 1.21× time; the matched physical-merit line search uses 16.3/22.8 and 1.12×. Candidate qualification and recovery therefore deliver reliability under a matched evaluation budget.