Neither Precision Nor Architecture Alone: Controlled Tests of Failure Remedies for Physics-Informed Neural Networks

Jinyuan Zhang
202621116012480@stu.hubu.edu.cn
Peng He*
penghe@hubu.edu.cn
He Hu
202521120012751@stu.hubu.edu.cn
Yin Yuan
202521120012766@stu.hubu.edu.cn
ShengShuo Jiao
202621120012764@stu.hubu.edu.cn

Hubei University, Wuhan, China

*Corresponding author: penghe@hubu.edu.cn

Abstract

Physics-Informed Neural Networks (PINNs) frequently fail on stiff or advection-dominated PDEs, and two recent accounts offer competing remedies: switching from FP32 to FP64 to repair an L-BFGS stopping artifact, or replacing the MLP with a state-space-model (SSM) backbone plus sub-sequence alignment to counter architectural simplicity bias. We test both under matched, seed-paired controls in a pre-registered 144-run study spanning convection, reaction, and wave, plus an independent 85-run convection/wave study; success is relative ℓ_2 error below 0.05. The two remedies act on disjoint regime-and-seed slices: neither substitutes for the other. On hard convection (β=50), alignment recovers 2/5 seeds in FP32 and 3/5 in FP64, where the unaligned SSM succeeds on 0/5 seeds at either precision and the vanilla MLP moves only from 0/5 to 1/5 across the precision switch—the recoveries trace to the alignment objective, not the backbone. On reaction the backbone alone already succeeds on 3/5–4/5 seeds, so each remedy covers a regime the other does not. Responses are also seed-specific: the same precision switch flips individual seeds in opposite directions and, on wave, lowers median error with no statistically significant success gain. Tightening the inner L-BFGS tolerance in an independent repeated-step runner likewise lowers median error at a large runtime cost, with success counts unchanged. Precision, stopping, backbone, and alignment must therefore be evaluated jointly and reported per seed.

1 Introduction

Physics-Informed Neural Networks (PINNs) Raissi et al. [2019], Karniadakis et al. [2021] embed PDE residuals into the training loss, enabling mesh-free solutions to forward and inverse problems, and have matured into a broad scientific-machine-learning toolkit supported by shared libraries Lu et al. [2021]. Despite their elegance, PINNs break down on stiff or advection-dominated PDEs—a failure pattern documented across convection, reaction, and wave equations at moderate-to-high parameter regimes Krishnapriyan et al. [2021].

Two concurrent 2025 papers offer different diagnoses and cures:

- **PINNMamba** Xu et al. [2025a] attributes failures to *continuous-discrete mismatch* and *simplicity bias* of MLPs, and fixes them with a State Space Model (SSM) backbone plus sub-sequence contrastive alignment.
- **"FP64 is All You Need"** Xu et al. [2025b] attributes failures to an *optimizer artifact*—L-BFGS's `tolerance_change` sitting near the FP32 machine epsilon (∼1.19e-7)—and fixes them by switching to FP64.

These explanations motivate different remedies, but their sufficiency and transferability have not been tested together. If FP64 alone transfers across regimes, it should rescue failures without modifying the model; if the architecture/alignment remedy is sufficient, precision should add little once that protocol is fixed. No existing study tests both interventions within the same controlled slices. Figure 1 separates the two hypotheses and specifies the matched comparison required to test them jointly.

The tests come in two complementary studies, each changing one factor at a time. The local controlled study pools 144 runs across preliminary experiments, a pre-registered evaluation, tolerance ablations, protocol comparisons, and matched follow-ups; the independent validation study adds 85 convection and wave runs for the alignment-weight, tolerance, and seed analyses (a wave configuration whose coefficients depart from the public protocol is set aside). Every comparison is matched—a

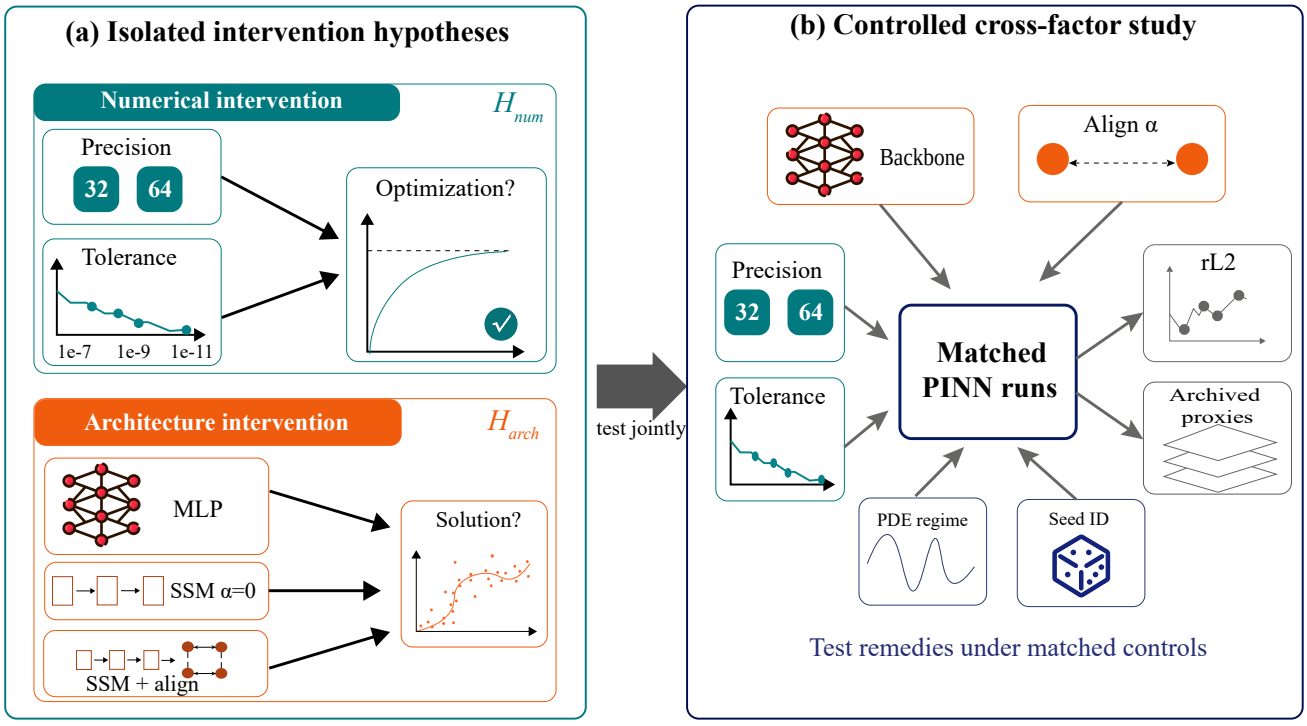

Figure 1: Why the two remedies require a joint test. Prior studies isolate numerical precision/stopping from architecture/alignment, whereas our matched comparisons vary precision, inner L-BFGS tolerance, backbone, and alignment weight while controlling the PDE regime and seed. rL2 and archived diagnostic proxies appear as descriptive summaries.

PDE–seed cell shares its collocation grid, budget, and optimizer across variants—and reported per seed, since the remedy that rescues one initialization may leave another failing.

Contributions.

1. A controlled comparison of precision, inner L-BFGS tolerance, backbone, and alignment within matched PDE and seed slices, with the effective sample size reported for each comparison.

2. Evidence that numerical and architectural interventions act on disjoint regimes and seeds: the alignment-weight sweep exhibits a threshold response, whereas tighter tolerance alters error and runtime but not success counts.

3. A quantification of seed and threshold sensitivity, together with the scope each conclusion supports.

2 Related Work

Diagnosing PINN failure modes. PINNs fail systematically on stiff or advection-dominated PDEs, and the literature offers several diagnoses that need not exclude each other. Krishnapriyan et al. [2021] documented failures on convection, reaction, and reaction-diffusion equations and attributed them to complex loss landscapes; Wang et al. [2021] localized the pathology in the gradient imbalance between residual and boundary terms (cf. Eq. 2); Wang et al. [2022b] traced it to the eigenspectrum of the Neural Tangent Kernel; and Wang et al. [2022a] to training schedules that violate temporal causality. Benchmarks and shared libraries have since standardized how such failures are produced and measured Hao et al. [2024], Lu et al. [2021]. What these accounts share is a design, not only a conclusion: each varies one factor against an otherwise fixed protocol, so which mechanism binds in which regime—and whether a remedy for one transfers to another—has not been tested under matched controls. That joint test is the one our study runs.

Two remedies proposed in 2025. Two concurrent works propose cures from opposite ends of the stack. On the architecture side, sequential inductive bias has been added through temporal Transformer attention Zhao et al. [2024], Vaswani et al. [2017] and, more recently, through selective state-space (Mamba) blocks Xu et al. [2025a], Gu and Dao [2023], the latter paired with a subsequence contrastive alignment loss and state-of-the-art results on convection-β=50 and reaction-ρ=5; a sibling line instead swaps the MLP's activation functions to

counter spectral bias Sitzmann et al. [2020]. On the optimizer side, Xu et al. [2025b] attribute the same benchmark failures to L-BFGS's `tolerance_change` sitting near the FP32 machine epsilon and report that FP64 arithmetic resolves them. The two papers evaluate overlapping failure cases but under different protocols, report one seed per configuration, and leave their compound remedies unablated—PINNMamba never separates the alignment loss from the SSM backbone, and the FP64 study never tightens the tolerance within FP32. We therefore place both remedies inside shared PDE–seed cells and vary one factor at a time.

Loss weighting and training schedules. A third family changes neither architecture nor arithmetic but reshapes the objective: gradient-based adaptive balancing Wang et al. [2021], NTK-based weights Wang et al. [2022b], co-trained self-adaptive masks McClenny and Braga-Neto [2023], and causal temporal weighting Wang et al. [2022a], alongside curriculum schedules that march the training window forward in time Krishnapriyan et al. [2021]. Read this way, sub-sequence alignment belongs to the same family: it adds an agreement term to the objective rather than a new optimizer or a new number format. Our protocol holds the rest of this family fixed and sweeps the alignment term by weight, so its contribution separates from the backbone that carries it.

The optimizer and precision view. L-BFGS Liu and Nocedal [1989] is the de facto second-order optimizer in PINN training, and its convergence tests cannot resolve progress below the roundoff of the working arithmetic Higham [2002]. Loss-landscape analysis pushes toward the same conclusion from the other side: PINN objectives are ill-conditioned enough to defeat first-order methods, which is why second-order optimizers dominate practice Rathore et al. [2024]. The FP64 remedy rests on exactly this foundation: if the stopping floor is set by arithmetic, raising precision lowers the floor. What the foundation does not predict is whether the lowered floor turns failures into successes, in which regimes, and for which seeds; that part is empirical, and we answer it per seed.

Study design. Prospective registration is standard in clinical trials but rare in machine learning, and reproducibility initiatives Pineau et al. [2021] motivate our frozen evaluation rules. Concurrent work maps optimizer effectiveness across regimes spanning PINNs, neural operators, and neural ODEs Wang et al. [2026]. Our contribution is narrower and complementary: precision, stopping, backbone, and alignment held under matched PINN controls, compared seed by seed, with the evaluation rules registered before the study expanded.

3 Method

Figure 2 summarizes how a matched run is constructed, trained, and analyzed.

3.1 Problem Setup

We study three one-dimensional PDEs that are standard PINN stress tests Krishnapriyan et al. [2021]: the convection equation $u_t + \beta u_x = 0$ at β=50, whose transported profile develops steep, high-frequency gradients; the stiff reaction equation $u_t = \rho u(1-u)$ at ρ=5, with a rapidly saturating solution; and the wave equation $u_{tt} - 4u_{xx} = 0$ under a two-frequency initial displacement, whose multi-scale oscillations stress temporal coherence. All three are posed on a periodic spatial domain and trained against analytic reference solutions on a fine grid; the parameters are deliberately hard—each regime fails under a standard MLP trained with L-BFGS, the joint baseline both remedies are meant to rescue.

Both remedies target the same objective, so it is worth stating explicitly. Writing each PDE in operator form $M[u] = 0$ on $\Omega \times [0, T]$ with boundary operator B and initial operator I, the network u_θ minimizes the physics-informed objective

$$\mathcal{L}(\theta) = w_r \mathcal{L}_r(\theta) + w_b \mathcal{L}_b(\theta) + w_i \mathcal{L}_i(\theta), \qquad (1)$$
$$\mathcal{L}_r(\theta) = \tfrac{1}{|\chi_r|} \textstyle\sum_{(x,t)\in\chi_r} \|M[u_\theta](x,t)\|^2,$$

with $\mathcal{L}_b, \mathcal{L}_i$ the mean-squared boundary and initial residuals over their collocation sets. The three regimes are precisely where this objective is hard to minimize: the residual gradient dominates the data gradients in the sense of the pathology ratio Wang et al. [2021],

$$\frac{\|\nabla_\theta \mathcal{L}_r\|}{\|\nabla_\theta \mathcal{L}_b\| + \|\nabla_\theta \mathcal{L}_i\|} \gg 1, \qquad (2)$$

equivalently the Neural Tangent Kernel is ill-conditioned, $\lambda_{\max}(K_r)/\lambda_{\min}(K_{b,i}) \gg 1$ Wang et al. [2022b]. Each remedy attempts to restore minimizability of (1), but along a different axis.

3.2 Remedy Mechanisms

The two propositions below make precise why these interventions are not interchangeable.

Precision acts on the optimizer's stopping test. L-BFGS terminates when the iterate change drops below a `tolerance_change` threshold τ Liu and Nocedal [1989], and that threshold cannot be resolved below the roundoff at which the objective is evaluated.

Proposition 1 (Precision-bounded stopping floor)**.** *Let $\mathcal{L}$ be evaluated in precision p with unit roundoff u_p (FP32: $u \approx 5.96\times10^{-8}$, machine $\varepsilon \approx 1.19\times10^{-7}$; FP64: $u \approx 1.11\times10^{-16}$). The iterate difference $|\Delta\mathcal{L}_k|$ driving the stopping test satisfies the roundoff floor $|\Delta\mathcal{L}_k| \gtrsim u_p|\mathcal{L}_k|$, so the test resolves genuine progress only down to* $\max(\tau, u_p|\mathcal{L}_k|)$.

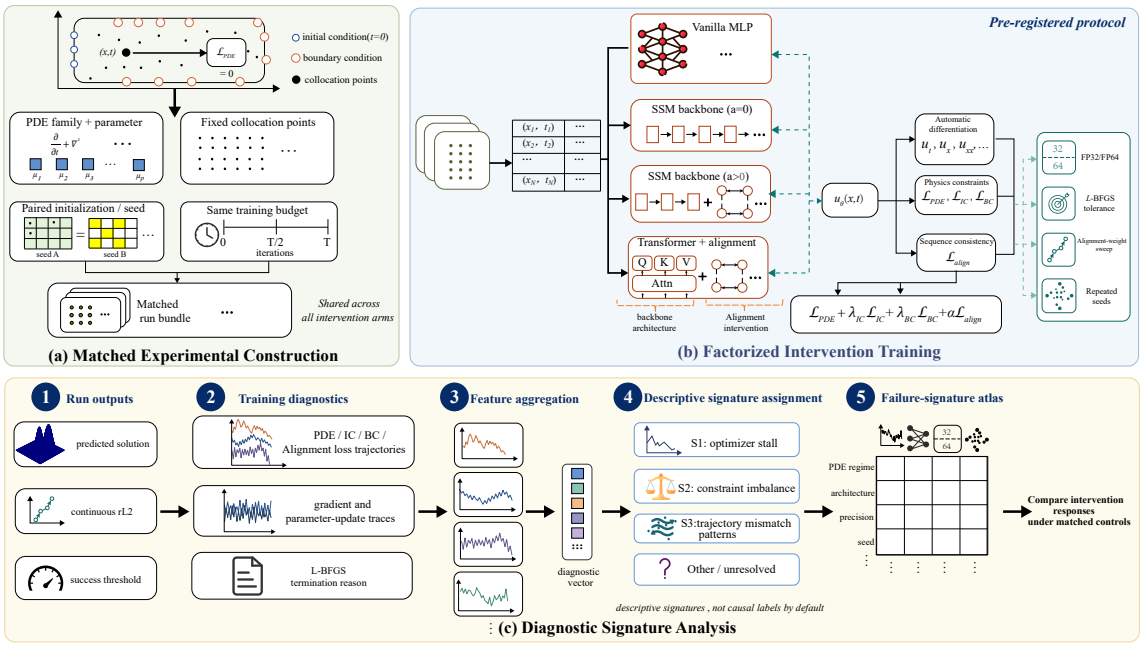

Figure 2: Overview of the controlled study workflow.

Intuition. The optimizer cannot resolve progress below the loss rounding error: in FP32 this floor coincides with the default τ, while in FP64 it lies nine orders of magnitude below it. The full proof (with the IEEE-754 error expansion) is in Appendix A.1 of the supplement.

Alignment acts on the hypothesis class. The second remedy adds the sub-sequence contrastive term of Xu et al. [2025a],

$$\mathcal{L}_{\text{align}}(\theta) = \frac{1}{(k-1)|\chi|}\sum_{(x,t)\in\chi}\sum_{j=1}^{k-1}\Delta_j(x,t;\theta), \qquad (3)$$
$$\Delta_j = \left\|u_\theta^{(s_0)}(x,t+j\Delta t) - u_\theta^{(s_j)}(x,t+j\Delta t)\right\|^2,$$

where $u_\theta^{(s)}$ is the prediction issued by sub-sequence s, so the trained objective is $\mathcal{L} + \alpha\mathcal{L}_{\text{align}}$. It is one instance of the temporal-conditioning family. Causal residual weighting Wang et al. [2022a] replaces $\mathcal{L}_r$ by a causally weighted form,

$$\mathcal{L}_c = \frac{1}{|\chi|}\sum_{(x,t)\in\chi} w_c(x,t)\|M[u_\theta](x,t)\|^2, \qquad (4)$$
$$w_c(x,t) \propto \exp\big(-\varepsilon\textstyle\sum_{t'\le t}\|M[u_\theta](x,t')\|^2\big),$$

which suppresses residual error at time t until earlier times are fit; self-adaptive masks McClenny and Braga-Neto [2023] instead use $\mathcal{L} = \sum_i s_i\|M[u_\theta](x_i,t_i)\|^2$ with co-trained weights $s_i \ge 0$. All three modify the hypothesis class, not the stopping test.

Proposition 2 (Alignment as over-determination)**.** *Let $\mathcal{F}_{\chi^*} = \{u_\theta : M[u_\theta]|_{\chi^*} = 0\}$ be the collocation-feasible set, which by Theorem 4.1 of Xu et al. [2025a] contains infinitely many spurious solutions. The agreement term (3) imposes $m = (k-1)|\chi|$ equality constraints $u_\theta^{(s_0)} = u_\theta^{(s_j)}$ at shared collocation points; whenever these are independent on the hypothesis class, the feasible set shrinks to $\mathcal{F}_{\chi^*} \cap \{\text{agreements}\}$ of dimension $\dim \mathcal{F}_{\chi^*} - m$, removing the bump-type spurious solutions.*

Intuition. Each cross-subsequence agreement is an independent level-set that lowers the feasible manifold's dimension by one; a bump-type spurious solution violates at least one such agreement and is removed (Theorem 4.1 of Xu et al. [2025a]). The full proof (bump-function construction) is in Appendix A.2 of the supplement.
Propositions 1–2 formalize our central claim: precision acts on the optimizer's stopping axis and alignment on the hypothesis-class axis, so the two remedies lie on orthogonal axes of the loss landscape Rathore et al. [2024] and neither subsumes the other.

3.3 Experimental Protocol

The local controlled study pools 144 runs from preliminary experiments, the registered evaluation, tolerance ablations, protocol comparisons, and matched follow-ups. The independent validation study contributes 85 runs and 55.7 run-hours. Because neither study forms a complete factorial block, Table 1 reports the effective sample size for each research question (RQ) instead of merging overlapping slices into a single total.

Within each comparison we control PDE (convection β=50, reaction ρ=5, wave with a two-frequency initial displacement), architecture (vanilla MLP; SSM without alignment, α=0; SSM with alignment; PINNsFormer in a smaller slice), precision (FP32, FP64), seeds (five per main cell; three or five-to-ten in smaller/independent slices), and stopping/alignment (`tolerance_change`$\in \{10^{-7}, 10^{-9}, 10^{-11}\}$ on the convection MLP sweep; $\alpha \in \{0, 100, 300, 1000, 3000\}$ on the SSM sweep). Runs in a

Table 1: Evidence coverage. Local controlled-study slices overlap; independent-study slices partition the 85 runs, with one non-comparable slice excluded from validation claims. Asterisked proxy counts have unequal observability.

Slice	RQ	n
Local study: five-seed conv/react slice	1–2,4	60
Local study: instrumented conv/react subset	5	87*
Local study: model-protocol comparison	2	36
Independent study: convection α sweep	2	25
Independent study: convection tolerance sweep	3	30
Independent study: wave MLP precision slice	4	20
Independent study: excluded wave-PINNMamba slice	–	10

*S1 is observable for 46/87 records (39/63 failures); S3b plateau data for 18/87 (13/63). The registered S2 peak-amplitude condition is absent from the archive.

comparison share the same PDE instance, collocation grid, training budget, and L-BFGS strong-Wolfe configuration. The registered evaluation uses 5,000 Adam steps then up to 1,000 outer L-BFGS steps on a 101×101 collocation grid with a 501×501 evaluation grid and disabled TF32; the independent study uses a repeated-step runner (1,000 L-BFGS calls, `max_iter`$= 20$) on an NVIDIA RTX 4090D. Accuracy is the relative ℓ_2 error $\mathrm{rL2} = \|u_\theta - u_{\mathrm{ref}}\|_2 / \|u_{\mathrm{ref}}\|_2$ over the evaluation grid, with success at unrounded $\mathrm{rL2} < 0.05$; each cell reuses the same seeds across its variants so differences reflect the intervention.

3.4 Pre-registration and Provenance

To separate confirmatory from exploratory evidence, we registered the analysis plan on May 28, 2026 after a set of preliminary experiments and before expanding the study. The registered plan fixed the success threshold ($\mathrm{rL2} < 0.05$), the diagnostic signature definitions (S1 optimizer stall; S2 constraint-mismatch collapse; S3a–c trajectory conditions), and the rule that demotes the wave PDE when too few cells admit a clean signature. The supplementary archive contains the registered document, its timestamp, and a provenance table marking each run as preliminary, registered, follow-up, or independent; the independent validation study and the local follow-ups were run after registration.

3.5 Archived Diagnostic Proxies

The registered classifier tags each failure as optimizer stall (S1), constraint mismatch (S2), or a trajectory condition (S3a–c). The archived implementation covers these rules in part (two registered tests absent; 41/87 records without L-BFGS logs), so the proxies enter as descriptive summaries, with unavailable evidence read as "not fired"; the complete rules and coverage are deferred to the supplement.

4 Experiments

Five research questions organize the experiments:

- **RQ1:** Are precision and architecture/alignment interchangeable fixes under matched controls?
- **RQ2:** Which part of the architecture-based remedy matters: alignment weight or the recorded model protocol?
- **RQ3:** How do precision and inner L-BFGS tolerance affect the repeated-step runner?
- **RQ4:** How stable are intervention responses across seeds and evaluation choices?
- **RQ5:** Which archived diagnostic proxies appear in the instrumented convection/reaction subset?

4.1 Cross-Factor Comparison (RQ1)

Precision and alignment act on different PDE and seed slices, and neither substitutes for the other. Figure 3 shows the local controlled-study response patterns, and Table 2 reports per-seed rL2 for the main convection/reaction slice at the registered outcome threshold. On hard convection, FP64 lifts vanilla-MLP success from 0/5 to 1/5 and leaves the unaligned SSM at 0/5, whereas alignment recovers 2/5 seeds in FP32 and 3/5 in FP64. On reaction, the SSM backbone already reaches 3/5–4/5 and alignment reaches 5/5 at both precisions, so the two remedies divide the regimes between them. The wave, transformer, tolerance, and alignment-weight follow-ups answer the remaining RQs.

4.2 Alignment Contribution and Weight Response (RQ2)

The hard-convection SSM recoveries require alignment, and the response depends on both the alignment weight and the PDE. Table 3 isolates the backbone's contribution on convection-β=50:

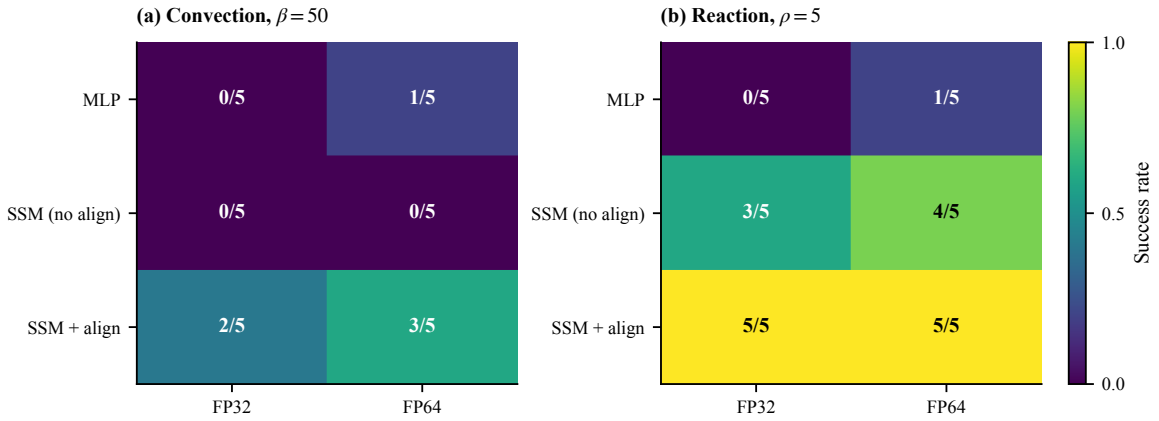

Figure 3: Local controlled-study five-seed slice (success defined as rL2 < 0.05), pooling registered and matched follow-up runs; FP64 no-alignment seeds 3–4 are drawn from the follow-ups. Precision and architecture/alignment produce regime- and seed-specific gains, with no uniform monotone improvement.

Table 2: Per-seed rL2 in the local controlled-study slice (five seeds per cell; bold is rL2 < 0.05 on unrounded values).

PDE	Regime	s=0	s=1	s=2	s=3	s=4	Rate
Conv-β50	mlp FP32	1.005	0.896	0.806	0.914	0.800	0/5
	mlp FP64	0.080	0.738	0.077	**0.045**	0.052	1/5
	ssm_noalign FP32	2.027	1.361	1.188	1.327	1.372	0/5
	ssm_noalign FP64	1.312	1.400	1.298	1.238	1.326	0/5
	ssm_align FP32	**0.011**	**0.0498**	0.549	0.386	0.904	2/5
	ssm_align FP64	**0.006**	1.287	**0.005**	0.168	**0.026**	3/5
React-ρ5	mlp FP32	0.979	0.979	0.079	0.980	0.974	0/5
	mlp FP64	0.055	0.065	0.059	0.064	**0.042**	1/5
	ssm_noalign FP32	**0.026**	0.106	**0.030**	**0.031**	0.682	3/5
	ssm_noalign FP64	**0.034**	0.643	**0.035**	**0.029**	**0.023**	4/5
	ssm_align FP32	**0.019**	**0.028**	**0.010**	**0.022**	**0.031**	5/5
	ssm_align FP64	**0.028**	**0.031**	**0.026**	**0.022**	**0.023**	5/5

	FP32	FP64
ssm_noalign	0/5 (all >1.0)	0/5 (all >1.0)
ssm_align	2/5	3/5

Table 3: Hard-convection success by alignment and precision (five seeds per cell).

Without alignment, the SSM backbone yields rL2 above 1.0 in every reported cell, with no improvement over the vanilla MLP. Under this protocol, the SSM-based recoveries are therefore attributable to the alignment objective rather than to the backbone alone. On reaction-ρ=5, the backbone alone achieves 3/5 (FP32) and 4/5 (FP64), while alignment lifts both precisions to 5/5; alignment thus has a larger effect on convection than on reaction.

Weight response. The independent sweep reuses the same five seeds at $\alpha \in \{0, 100, 300, 1000, 3000\}$ on convection PINNMamba FP32. Success counts are 0/5, 0/5, 0/5, 2/5, and 3/5, and median rL2 falls from 0.972 at α=300 to 0.801 at α=1000 and 0.041 at α=3000. Figure 4 shows a threshold-like response rather than a sharp phase transition; at five seeds on a coarse grid, the trend is clear but the threshold is not precisely calibrated.

Recorded model protocols. The local controlled study also compares PINNMamba with PINNsFormer. Table 4 reports the convection-β=50 success counts.

Convection-β=50	FP32	FP64
PINNMamba	2/5	3/5
PINNsFormer	0/5	0/3

Table 4: PINNMamba vs. PINNsFormer success on convection-β=50 (unequal seeds; protocol-level contrast).

PINNsFormer records no successful convection seed (all rL2 > 0.97); on reaction, both protocols solve every cell. The PINNsFormer manifest leaves the alignment field blank and the methods differ beyond back-

bone class, so the contrast here is protocol-level; representative trajectories are in the supplement.

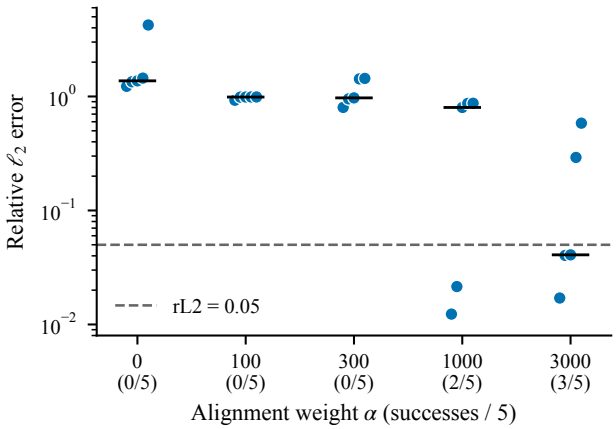

Figure 4: Independent convection PINNMamba FP32 sweep (the same five seeds per α). Points are seeds, black segments are medians, and tick parentheses give successes/5. The dashed line is rL2 = 0.05; points below it are successes.

4.3 Tolerance Response in the Repeated-Step Runner (RQ3)

Tighter inner tolerance lowers error and raises cost without changing success counts in the repeated-step implementation. The independent validation study issues 1,000 optimizer calls, each capped at 20 inner L-BFGS iterations, and sweeps `tolerance_change` across three orders of magnitude on convection MLP. Because this protocol differs from a single-call L-BFGS run, it probes the tolerance response of a repeated-step solver rather than reproducing the single-call premature-termination mechanism. Table 5 reports success count and median rL2 per cell.

Precision	tol=1e-7	tol=1e-9	tol=1e-11
FP32	1/5; 1.006	1/5; 0.720	1/5; 0.720
FP64	1/5; 1.005	1/5; 0.314	1/5; 0.314

Table 5: Repeated-step convection MLP sweep: success count and median rL2 by precision and `tolerance_change`.

Tightening tolerance leaves the success count at 1/5 for both precisions while lowering the median error in each, more markedly in FP64. Median runtime rises from 120.8 to 2,573.3–2,569.6 seconds in FP64 and from 17.0 to 278.1–275.7 seconds in FP32, so the error gain—which concentrates at the step from 10^{-7} to 10^{-9} and then saturates—comes at a sharply negative marginal return per second. Figure 5 shows the seed distributions. Inner-iteration counts and termination reasons are absent from the logs, so these numbers describe a precision-dependent tolerance response of the repeated-step runner, not the optimizer's internal interaction or mechanism.

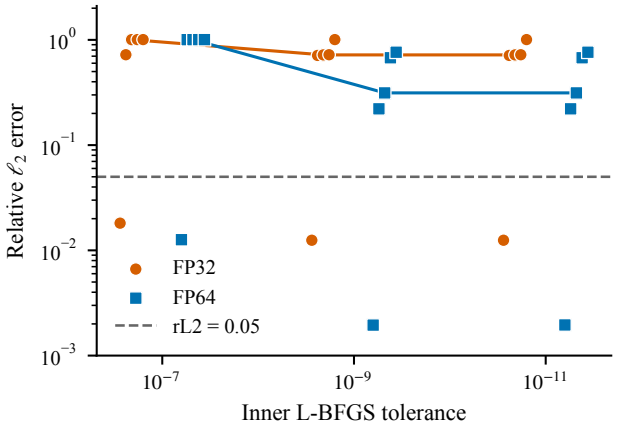

Figure 5: Independent repeated-step convection MLP sweep (the same five seeds per precision/tolerance cell). Circles are FP32, squares are FP64, and connected markers are medians. The dashed line is rL2 = 0.05; points below it are successes.

4.4 Seed and Evaluation Robustness (RQ4)

Intervention responses vary sharply across seeds. In the local controlled-study convection SSM+alignment slice, FP64 moves seed 1 from 0.0498 to 1.287 but seed 2 from 0.549 to 0.005—opposite directions under the same intervention. Reaction SSM+alignment succeeds in all ten precision–seed cells, whereas the no-alignment reaction arm still fails one FP64 cell. The independent ten-seed wave-MLP slice yields 4/10 successes in FP32 (median rL2 0.125) and 6/10 in FP64 (median 0.045). Among the paired seeds, three succeed only in FP32, five only in FP64, one in both, and one in neither; exact McNemar p=0.727, and a paired Wilcoxon test on log-rL2 gives p=0.375. FP64 therefore lowers the median error, though the success-rate gain is not statistically significant. The original three-seed wave-architecture result is exploratory, and the non-comparable independent wave-PINNMamba runs sit outside this analysis.

Table 6 sweeps the success threshold ±40% (0.03–0.07) on the 60-run local convection/reaction slice.

Threshold	0.03	0.04	**0.05**	0.06	0.07
Success rate	25%	35%	**40%**	45%	48%

Table 6: Success-rate sensitivity to the rL2 threshold on the 60-run local convection/reaction slice.

The aggregate regime ordering is stable across this range, although individual near-threshold seeds change status; the smooth rise in the success rate as the cutoff loosens confirms that the rankings are not artifacts of the 0.05 boundary.

4.5 Diagnostic Failure Signatures (RQ5)

Among the 63 failures in the 59-convection/28-reaction subset, the dominant proxy class is trajectory-related, consistent with the late-time and multi-scale distortions seen on convection and wave; the unresolved share reflects the missing L-BFGS logs noted in Archived Diagnostic Proxies. The exclusive counts (stall / constraint / trajectory / mixed / unresolved) and the bar chart are in the supplement.

5 Discussion

Revisiting the two diagnoses. Both original diagnoses survive matched controls, each within a narrower scope than first claimed. The precision diagnosis is right that arithmetic bounds the L-BFGS stopping test—FP64 lowers error in every controlled cell—yet precision alone rarely flips a hard-convection failure into a success (1/5), which makes it a complement to, rather than a substitute for, the architectural remedy. The architecture diagnosis is right that the SSM with alignment recovers hard convection, but our ablation credits the alignment objective rather than the Mamba backbone, since the unaligned SSM matches or underperforms the vanilla MLP. Each remedy holds where it was reported, and each covers a different slice of the regimes and seeds we test.

Why seed-level reporting matters. On hard convection with alignment, FP64 turns seed 2 from failure into success but seed 1 from success into failure, so the cell-level rate (3/5) hides opposite per-seed movements; reversals like these are characterized through their distribution, not confirmed by another seed—the case for per-seed reporting.

Implications. Neither "just use FP64" nor "just change the backbone" qualifies as general advice: practitioners should compare precision and stopping jointly, tune alignment as an intervention, and report per-seed error distributions (for lower error, FP64 at moderate tolerance is the efficient point, since tightening past 10^{-9} multiplies runtime without further gain). Single-configuration conclusions shift when seed, precision, stopping, and alignment vary jointly; causal interpretation of the proxies awaits counterfactual continuations.

Limitations. The corpus covers three one-dimensional PDE families at selected parameter values, not a broad scientific-computing benchmark. The PINNMamba–PINNsFormer comparison is a method-protocol comparison with unequal seeds, a blank PINNsFormer alignment field, no parameter/tuning match, and no identical initialization tensors across architectures. The independent validation uses a repeated-step runner, and its wave-PINNMamba coefficient differs from the public wave protocol. The archived proxy classifier omits two registered conditions, 41/87 records lack L-BFGS logs, S3 uses final rL2, and the causal taxonomy still awaits validation by rescue experiments.

6 Reproducibility

All random seeds are reported, and rL2 < 0.05 is applied to unrounded values. Artifacts include the run manifest, registered analysis file, archived proxy JSON, independent-validation CSV and implementation, and the scripts that regenerate Figures 3–5; provenance separates preliminary, registered, follow-up, and independent evidence, and the pinned upstream code and environment accompany the submission.

7 Conclusion

Under matched controls, precision and tolerance act on different PDE and seed slices than architecture and alignment, so the two classes of remedy are not interchangeable. The alignment recoveries, tolerance costs, and seed reversals we observe make a practical case: evaluate precision, stopping, backbone, and alignment jointly, and report every outcome per seed.

References

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Zhongkai Hao, Jiachen Yao, Chang Su, Hang Su, Ziao Wang, Fanzhi Lu, Zeyu Xia, Yichi Zhang, Songming Liu, Lu Lu, and Jun Zhu. PINNacle: A comprehensive benchmark of physics-informed neural networks for solving PDEs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/8c63299fb2820ef41cb05e2ff11836f5-Abstract-Datasets_and_Benchmarks_Track.html.

Nicholas J Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics (SIAM), 2 edition, 2002.

George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

Aditi Krishnapriyan, Amir Gholami, Shandian Zhe, Robert Kirby, and Michael W Mahoney. Characterizing possible failure modes in physics-informed neural networks. In *Advances in Neural Information Processing Systems*, 2021.

Dong C Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1):503–528, 1989.

Lu Lu, Xuhui Meng, Zhiping Mao, and George Em Karniadakis. DeepXDE: A deep learning library for solving differential equations. *SIAM Review*, 63(1):208–228, 2021.

Levi D McClenny and Ulisses M Braga-Neto. Self-adaptive physics-informed neural networks. *Journal of Computational Physics*, 474:111722, 2023.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research: a report from the NeurIPS 2019 reproducibility program. *Journal of Machine Learning Research*, 22(164): 1–20, 2021.

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

Pratik Rathore, Weimu Lei, Zachary Frangella, Lu Lu, and Madeleine Udell. Challenges in training PINNs: A loss landscape perspective. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 42159–42191. PMLR, 2024. URL https://proceedings.mlr.press/v235/rathore24a.html.

Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Sifan Wang, Yujun Teng, and Paris Perdikaris. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. *SIAM Journal on Scientific Computing*, 43(5):A3055–A3081, 2021.

Sifan Wang, Shyam Sankaran, and Paris Perdikaris. Respecting causality for training physics-informed neural networks. *arXiv preprint arXiv:2203.07404*, 2022a.

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

Yuxin Wang, Yuanzhe Hu, Xiaokun Zhong, Xiaopeng Wang, Haiquan Lu, Tianyu Pang, Michael W. Mahoney, Yujun Yan, Pu Ren, and Yaoqing Yang. Unveiling multi-regime patterns in SciML: Distinct failure modes and regime-specific optimization. *arXiv preprint arXiv:2605.29153*, 2026. Accepted at the International Conference on Machine Learning.

Chenhui Xu, Dancheng Liu, Yuting Hu, Jiajie Li, Ruiyang Qin, Qingxiao Zheng, and Jinjun Xiong. Sub-sequential physics-informed learning with state space model. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 69507–69525. PMLR, 13–19 Jul 2025a. URL https://proceedings.mlr.press/v267/xu25t.html.

Chenhui Xu, Dancheng Liu, Amir Nassereldine, and Jinjun Xiong. FP64 is all you need: Rethinking failure modes in physics-informed neural networks. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://papers.nips.cc/paper_files/paper/2025/hash/d274ea8b3c7f526f79ac9ce75ee3c8df-Abstract-Conference.html.

Zhiyuan Zhao, Xueying Ding, and B. Aditya Prakash. PINNsFormer: A transformer-based framework for physics-informed neural networks. In *International Conference on Learning Representations*, 2024.

This supplement provides the full proofs of the two propositions stated in Section 3 (Remedy Mechanisms) of the main paper. We write u_p for the unit roundoff of precision p (FP32: $u \approx 5.96 \times 10^{-8}$, machine $\varepsilon \approx 1.19 \times 10^{-7}$; FP64: $u \approx 1.11 \times 10^{-16}$), $\mathrm{fl}(\cdot)$ for floating-point evaluation, τ for the L-BFGS `tolerance_change` threshold, and $\mathcal{L}$ for the physics-informed objective of Eq. (1) of the main text.

Appendix A. Proofs

A.1 Proof of Proposition 1 (precision-bounded stopping floor)

We restate the proposition and prove it from the IEEE-754 floating-point model.

Proposition (Precision-bounded stopping floor)**.** *Let $\mathcal{L}$ be evaluated in precision p with unit roundoff u_p. The iterate difference $|\Delta\mathcal{L}_k| = |\mathcal{L}_{k+1} - \mathcal{L}_k|$ that drives the L-BFGS `tolerance_change` test satisfies the roundoff floor $|\mathrm{fl}(\Delta\mathcal{L}_k)| \gtrsim u_p|\mathcal{L}_k|$, so the test resolves genuine progress only down to* $\max(\tau,\, u_p|\mathcal{L}_k|)$.

Proof. By the IEEE-754 floating-point model (Higham 2002), every computed scalar satisfies $\mathrm{fl}(a) = a(1+\delta_a)$ with $|\delta_a| \le u_p$, equivalently $|\mathrm{fl}(a) - a| \le u_p|a|$. Writing $\mathrm{fl}(\mathcal{L}_k) = \mathcal{L}_k + e_k$ with $|e_k| \le u_p|\mathcal{L}_k|$, the computed iterate difference decomposes as

$$\mathrm{fl}(\mathcal{L}_{k+1}) - \mathrm{fl}(\mathcal{L}_k) = (\mathcal{L}_{k+1} - \mathcal{L}_k) + (e_{k+1} - e_k), \tag{5}$$

and the second term is bounded in magnitude by $u_p(|\mathcal{L}_{k+1}| + |\mathcal{L}_k|) \approx 2u_p|\mathcal{L}_k|$ near convergence, where $|\mathcal{L}_{k+1}| \approx |\mathcal{L}_k|$. This term is an irreducible noise floor: any genuine decrease $(\mathcal{L}_{k+1} - \mathcal{L}_k)$ smaller than $2u_p|\mathcal{L}_k|$ is masked by rounding error and indistinguishable from zero. The L-BFGS stopping criterion $|\mathrm{fl}(\mathcal{L}_{k+1}) - \mathrm{fl}(\mathcal{L}_k)| < \tau$ (Liu & Nocedal 1989) therefore resolves genuine progress only down to $\max(\tau,\, 2u_p|\mathcal{L}_k|)$, i.e. $\max(\tau,\, u_p|\mathcal{L}_k|)$ up to the harmless factor of two.

Substituting the roundoffs settles the precision comparison. With the default `tolerance_change` $\tau \approx 1.19 \times 10^{-7}$ and the loss magnitudes $|\mathcal{L}_k| = O(1)$–$O(10^2)$ typical of the residual-dominated PINN regimes we study, the FP32 floor $2u_{\mathrm{FP32}}|\mathcal{L}_k| \sim 10^{-7}$–$10^{-5}$ is of the same order as or exceeds τ, so the test sits at the noise floor and may terminate while reducible progress remains. In FP64 the floor drops to $2u_{\mathrm{FP64}}|\mathcal{L}_k| \sim 10^{-15}$–$10^{-13}$, nine orders of magnitude below τ, restoring the test's ability to detect progress of size τ. Hence FP64 relaxes the stopping floor that FP32 imposes, without changing the model. □

A.2 Proof of Proposition 2 (alignment as over-determination)

We build on Theorem 4.1 of Xu et al. (2025), which shows that for any finite collocation set $\chi^* \subset \Omega \times [0,T]$ there exist infinitely many functions u_θ (constructed as bump perturbations supported in small neighborhoods of χ^*) with $M[u_\theta]|_{\chi^*} = 0$ yet $M[u_\theta] \neq 0$ almost everywhere on $\Omega \times [0,T] \setminus \chi^*$.

Proposition (Alignment as over-determination)**.** *Let $\mathcal{F}_{\chi^*} = \{u_\theta : M[u_\theta]|_{\chi^*} = 0\}$ be the collocation-feasible set. The sub-sequence agreement term of Eq. (3) in the main text imposes $m = (k-1)|\chi|$ equality constraints; whenever these are independent on the hypothesis class, the feasible set shrinks to $\mathcal{F}_{\chi^*} \cap \{\text{agreements}\}$ of dimension* $\dim \mathcal{F}_{\chi^*} - m$, *removing the bump-type spurious solutions of Xu et al. (2025).*

Proof. We prove the two claims in turn.

(a) Dimension reduction. The agreement term penalizes, at each shared collocation point $(x, t + j\Delta t)$, the disagreement between the predictions issued by the sub-sequences $s_0, \ldots, s_j$ that all contain it. On its zero set this is the system of $m = (k-1)|\chi|$ equality constraints

$$u_\theta^{(s_0)}(x, t + j\Delta t) = u_\theta^{(s_j)}(x, t + j\Delta t), \qquad j = 1, \ldots, k-1,\ (x,t) \in \chi. \tag{6}$$

Assume these constraints are functionally independent on the hypothesis class—a generic condition, since each constrains a distinct pair of sub-sequence outputs at a distinct point. By the regular level-set theorem (a corollary of the implicit function theorem), each independent equality constraint lowers the dimension of the feasible manifold $\mathcal{F}_{\chi^*}$ by one. Hence the augmented feasible set $\mathcal{F}_{\chi^*} \cap \{\text{agreements}\}$ has dimension $\dim \mathcal{F}_{\chi^*} - m$.

(b) Removal of bump-type spurious solutions. Let $\bar{u}$ denote the initial-condition-propagating reference solution. A bump-type spurious u_θ from Theorem 4.1 of Xu et al. (2025) agrees with $\bar{u}$ on χ^* (so $M[u_\theta]|_{\chi^*} = 0$) but departs from $\bar{u}$ on $\Omega \times [0,T] \setminus \chi^*$. Now consider how a sub-sequence s_j, which is trained to predict the solution forward from time $t + j\Delta t$, renders a shared point $(x, t + j\Delta t)$. The propagating solution $\bar{u}$ is a fixed point of this forward prediction, so it is rendered identically by every sub-sequence that contains the point: $\bar{u}^{(s_j)} = \bar{u}^{(s_0)}$. A spurious u_θ that fails

to propagate the initial condition instead drifts across sub-sequences, so $u_\theta^{(s_j)} \neq u_\theta^{(s_0)}$ at some shared point, giving $\Delta_j > 0$ for at least one j. Such a u_θ therefore violates the agreement set and is excluded from $\mathcal{F}_{\chi^*} \cap \{\text{agreements}\}$. With m large relative to the local degrees of freedom deposited near χ^* by the bump construction, all bump-type spurious solutions are excluded, leaving $\bar{u}$ as the dominant feasible hypothesis. □

Remark 1. The independence assumption in part (a) is the only non-trivial hypothesis; it holds generically because the agreement constraints couple disjoint sub-sequence outputs at disjoint points. Part (b) formalizes the design rationale of the sub-sequence contrastive loss: alignment enforces the cross-subsequence agreement that the propagating solution satisfies but collocation-only spurious solutions do not.