RESTORING CROSS-STEP FEEDBACK IN PHYSICS-INFORMED NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Physics-Informed Neural Networks (PINNs) are mesh-free solvers for partial differential equations (PDEs), yet their training often suffers from two recurring difficulties. One is **loss imbalance**, where the gradients of the PDE residual and the boundary or initial losses can differ by orders of magnitude, complicating loss weighting. The other is **poor internal propagation**, where the backbone's layer-to-layer updates can make sharp or high-frequency solution features difficult to learn. Existing remedies largely address these issues separately, but adaptive weighting can remain sensitive to **short-term fluctuations**, while stronger backbones can still underfit **difficult solution structures**. These challenges motivate investigating decayed signal accumulation along both network depth and training time. We propose **CSF-PINN** (**C**ross-**S**tep **F**eedback **P**hysics-**I**nformed **N**eural **N**etwork), which applies a common leaky-integral design at the architecture and loss-weighting levels. The Leaky-Integral Residual Block (LIRB) aggregates transformed features across depth through a recurrence equivalent to a damped hidden-state update with an initial-representation skip connection, while Integral-Controlled Weighting (ICW) combines instantaneous gradient statistics with a ratio of leaky integrals of square-root losses. Across five benchmark PDEs, CSF-PINN achieves the lowest error among all compared methods, including a 67.6% reduction in relative L_2 error over the strongest baseline on Wave.

1 INTRODUCTION

Physics-informed neural networks (PINNs) are a deep learning approach to solving partial differential equations (PDEs) (Raissi et al., 2019). They encode the governing equation together with its boundary and initial conditions as constraints in the training loss, so that the network approximates the PDE solution without a mesh (Xu et al., 2025). The framework applies equally to forward and inverse problems and has been used across fluid mechanics, solid mechanics, and electromagnetics (Matsubara & Yaguchi, 2025; Ghanem et al., 2025; Wan et al., 2026; Best Mckay et al., 2025). Training a PINN, however, is often unstable, and on challenging problems it can converge slowly or settle on an incorrect solution (Krishnapriyan et al., 2021). These difficulties arise mainly from two sources (Chen et al., 2022; Zhang et al., 2023; 2025; Zhou et al., 2025a).

P1) loss imbalance. The gradient of the PDE residual is usually much larger than the gradients of the boundary and initial terms, often by orders of magnitude (Vemuri & Denzler, 2023; Wang et al., 2022). Training is then dominated by the residual while the condition terms stay under-enforced, and the model tends to settle on a solution with a small residual that does not satisfy the boundary and initial conditions.

P2) poor internal propagation. A PINN backbone determines how intermediate features are transformed and combined across layers. Although earlier-layer information is represented in its hidden states, the resulting propagation and optimization dynamics can make sharp fronts and high-frequency solution features difficult to learn, leaving parts of the solution under-fitted.

These two problems have been addressed along two separate lines (Daw et al., 2023; Mojgani et al., 2022; Basir, 2023; Farhani et al., 2025; Kim & N'guessan, 2026). To mitigate ***P1***, one line dynamically assigns weights to the loss terms, with representative ideas including weighting based on gradient statistics such as gradient magnitude or the neural tangent kernel (Vemuri & Denzler, 2023; Wang et al., 2022), treating the weights as learnable parameters (McClenny & Braga-Neto, 2023;

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.

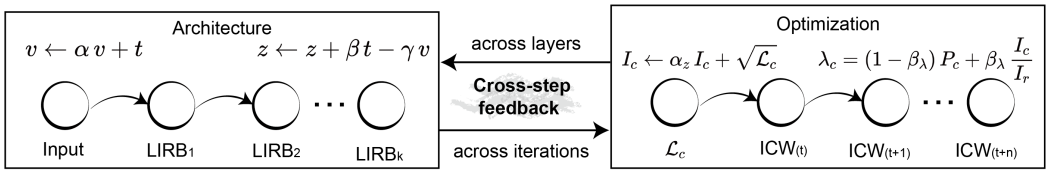

Figure 1: CSF-PINN restores cross-step feedback at two levels through one leaky-integral principle: LIRB carries an integral state across layers, and ICW carries a leaky integral of the loss history across iterations.

Xiang et al., 2022), and resolving conflicts between loss gradients from a multi-objective view (Liu et al., 2025a; Hwang & Lim, 2024). To mitigate ***P2***, the other line redesigns the backbone to enlarge its representational capacity, with representative ideas including periodic or Fourier feature representations (Tancik et al., 2020), attention-based sequential structures (Zhao et al., 2024), and networks that replace fixed activations with learnable ones (Liu et al., 2025c). Yet the remedies on each side do not fully resolve the difficulties they target. Adaptive weighting based on gradient statistics can remain sensitive to short-term fluctuations, leaving the loss imbalance in ***P1*** difficult to control. An enlarged backbone can still struggle to learn the sharp or high-frequency solution features associated with ***P2***. Moreover, the two lines typically focus on different parts of the training process: weighting methods adjust the relative influence of loss terms, whereas architectural methods modify how features are transformed and propagated. This leaves open how the two mechanisms should be designed together.

We examine ***P1*** and ***P2*** through a common design perspective: how signals are accumulated along their respective update axes. Inside the network, earlier-layer information is already represented in the hidden state, but the propagation rule determines how transformed features are retained and combined across depth. During training, instantaneous gradient statistics describe the current imbalance, while accumulated loss signals can indicate which constraints remain difficult over multiple iterations. These observations motivate cross-step feedback based on decayed signal accumulation, where a step is a layer in the network or an iteration in training (as shown in Figure 1). The shared perspective guides the design of both components without requiring the two difficulties to have an identical underlying cause.

Building on this view, we propose **CSF-PINN**, which implements Cross-Step Feedback at both levels through a common leaky-integral design. A leaky integral accumulates past signals with decay, making their contributions depend on how recently they were observed. In the architecture, the Leaky-Integral Residual Block (LIRB) aggregates transformed features across layers through a recurrence equivalent to a damped hidden-state update with an initial-representation skip connection, providing a propagation rule aimed at mitigating ***P2***. At the loss-weighting level, Integral-Controlled Weighting (ICW) combines instantaneous gradient statistics with a ratio of leaky integrals of square-root losses. The resulting weights incorporate both the current gradient imbalance and the relative persistence of condition losses, providing a history-dependent rule aimed at mitigating ***P1***.

Our main contributions are summarized as follows:

- CSF-PINN brings decay-controlled accumulation to both feature propagation and adaptive loss weighting, connecting two central aspects of PINN training through a shared recursive structure.
- LIRB aggregates transformed features across depth. An exact equivalent formulation reveals how hidden-state damping, an initial-representation skip, and feature injection jointly realize this propagation rule.
- ICW combines instantaneous gradient statistics with a ratio of accumulated loss signals. A local perturbation analysis characterizes its sensitivity to condition and reference signals and identifies the filtering behavior of its history-dependent branch.
- Across five PDE benchmarks, CSF-PINN achieves the lowest error among the compared methods, including a 67.6% reduction in relative L_2 error over the strongest baseline on Wave. Component and propagation-rule ablations examine the contributions of the proposed design.

2 Related Work

2.1 How PINNs Work?

A PINN (Raissi et al., 2019) encodes the constraints of a PDE into the training loss of a neural network (Chen et al.; Zou et al., 2026; Hong et al., 2026). Consider a PDE on a domain Ω with boundary $\partial\Omega$,

$$\begin{aligned} \mathcal{N}[u](\mathbf{x}) &= f(\mathbf{x}), \quad \mathbf{x} \in \Omega, \\ \mathcal{B}_c[u](\mathbf{x}) &= g_c(\mathbf{x}), \quad \mathbf{x} \in \partial\Omega,\ c = 1, \dots, C, \end{aligned} \tag{1}$$

where $\mathcal{N}$ is the differential operator and $\{\mathcal{B}_c\}_{c=1}^{C}$ are the boundary and initial conditions (Zhou et al., 2025b; Liu et al., 2025b; Shih et al., 2025). A PINN approximates the solution by a network u_θ and, via automatic differentiation, minimizes a composite loss:

$$\mathcal{L}(\theta) = \mathcal{L}_r(\theta) + \sum_{c=1}^{C} \lambda_c \, \mathcal{L}_c(\theta), \tag{2}$$

where $\mathcal{L}_r$ is the mean-squared PDE residual, $\mathcal{L}_c$ the mean-squared violation of the c-th condition, and the λ_c are loss weights. Methods to improve PINNs act mainly on two parts of this process: how information flows forward inside u_θ, and how the loss gradients are balanced through the λ_c.

2.2 Enhancing Information Flow in the Network

Many methods (Jiang et al., 2026b;a) enlarge what the backbone can represent. Periodic activations and Fourier features map coordinates onto a high-frequency basis to mitigate spectral bias (Tancik et al., 2020); Kolmogorov–Arnold networks use learnable activations on the edges (Liu et al., 2025c); attention-based structures capture dependencies among coordinates (Zhao et al., 2024); domain-decomposition methods assign separate networks to subregions (Jagtap & Karniadakis, 2020; Jagtap et al., 2020); and residual-style designs ease the training of deeper networks (Wang et al., 2024a; Abijuru et al., 2026). These methods modify the representation or propagation of information within the network. Our architectural component instead investigates a particular decay-weighted aggregation rule, while our weighting component acts separately on the balance among training losses.

2.3 Balancing Gradient Flow across Losses

Many methods (Wu & Wu, 2026) set the weights from gradient statistics, using the magnitude (Vemuri & Denzler, 2023), the neural tangent kernel (Wang et al., 2022), the variance (Maddu et al., 2022), or the norm (Deguchi & Asai, 2023). Others learn the weights directly (McClenny & Braga-Neto, 2023; Xiang et al., 2022), build them from accumulated residuals (Anagnostopoulos et al., 2024), assign them in causal order for time-dependent problems (Wang et al., 2024b), or resolve conflicts between loss gradients from a multi-objective view (Liu et al., 2025a; Hwang & Lim, 2024; Wang et al., 2025; Zhou et al., 2025a). These reduce the imbalance using current or historical training statistics, but operate only at the loss-weighting level and do not alter information flow inside the network.

3 Method

3.1 Overview

CSF-PINN combines LIRB for decay-weighted feature propagation with ICW for history-dependent loss weighting. Figure 2 summarizes the framework. We introduce both components and their analyses, then establish their shared accumulation structure.

3.2 Leaky-Integral Residual Block

Motivation. A standard residual update propagates information from earlier layers through its current hidden state. The question is therefore how transformed features should be accumulated

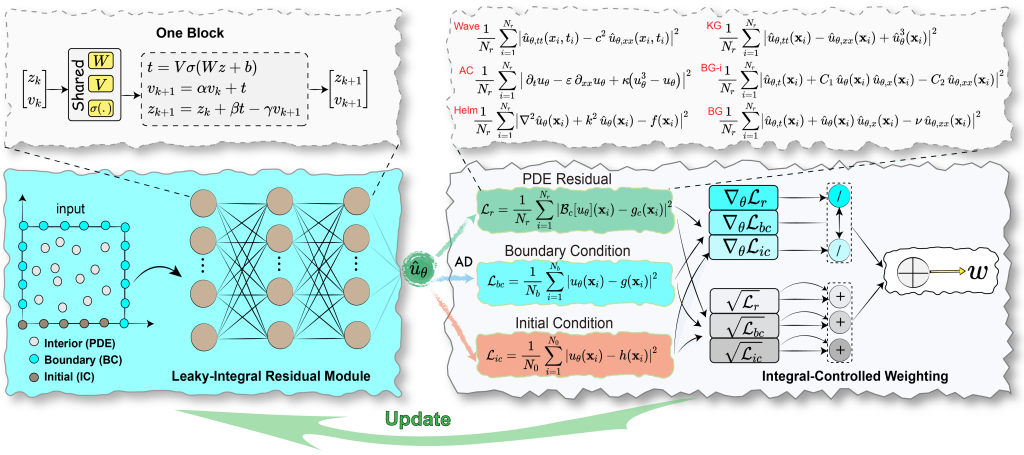

Figure 2: CSF-PINN framework. LIRB (left) accumulates transformed features across depth. Automatic differentiation constructs the PDE residual and condition losses. ICW (right) combines instantaneous gradient statistics with accumulated square-root loss ratios to determine the training weights.

across depth. To address ***P2***, we investigate a propagation rule that assigns exponentially decaying coefficients to transformed features from preceding layers. This design changes how features are retained and combined, rather than introducing an independent memory unavailable to a first-order network.

Block. LIRB uses an auxiliary integral state v to express this accumulation. At block k, let

$$T_k(z) = V_k\sigma(W_k z + b_k)$$

denote the layer-specific learnable transform. Each block updates

$$\begin{aligned} t_k &= T_k(z_k), \\ v_{k+1} &= \alpha v_k + t_k, \\ z_{k+1} &= z_k + \beta t_k - \gamma v_{k+1}. \end{aligned} \tag{3}$$

With $v_0 = 0$, the integral state unrolls as

$$v_{k+1} = \sum_{j=0}^{k} \alpha^{k-j} t_j. \tag{4}$$

Equivalent propagation form. We use a depth-shared coefficient $\alpha = \sigma(\theta_\alpha) \in (0, 1)$, with $\beta = 1$ and $\gamma = 1 - \alpha$. Under this choice,

$$\begin{aligned} z_{k+1} - z_k &= t_k - (1 - \alpha)v_{k+1} \\ &= \alpha(v_{k+1} - v_k). \end{aligned} \tag{5}$$

Summing across layers gives $z_k = z_0 + \alpha v_k$. Eliminating the auxiliary state therefore yields

$$z_{k+1} = \alpha z_k + (1 - \alpha)z_0 + \alpha T_k(z_k). \tag{6}$$

This equivalence reveals the structure of LIRB: transformed features are accumulated with exponential decay around the initial representation. The coefficient α jointly controls feature retention, hidden-state damping, and injection of the current transformation, while the initial skip anchors propagation to z_0. The integral formulation exposes the accumulated feature history, and the equivalent single-state form makes the resulting propagation mechanism explicit.

Connection to optimization-inspired networks. LIRB follows the optimization-inspired construction of (Wu et al., 2024), which interprets a forward pass as algorithmic updates with learnable transformations. Eliminating v from equation 3 gives a two-step recurrence whose homogeneous characteristic polynomial is

$$\rho(\lambda) = (\lambda - 1)(\lambda - \alpha). \tag{7}$$

For $|\alpha| < 1$, the non-unit root lies strictly inside the unit disk. Appendix B.1 establishes this root condition and, separately, the approximation capacity of the learnable-decay LIRB family. The latter follows from the equivalent propagation rule: as $\alpha \to 1^-$, LIRB uniformly approaches any fixed finite-depth residual network with the same transforms, transferring the corresponding L^p universality result to the LIRB family.

Proposition 1 (Depth-uniform state bound)**.** *Let $\alpha \in (0, 1)$ be shared across depth, with $\beta = 1$, $\gamma = 1 - \alpha$, and $v_0 = 0$. If $\|t_k\| \le M_t$ for every layer, then*

$$\|v_k\| \le \frac{M_t}{1-\alpha}, \qquad \|z_k\| \le \|z_0\| + \frac{\alpha M_t}{1-\alpha}, \tag{8}$$

independently of the network depth.

The assumptions and proof are given in Appendix B.2. This result bounds state magnitudes under uniformly bounded transformed features. It does not by itself establish stable gradients, faster optimization, or improved approximation of high-frequency solutions.

3.3 Integral-Controlled Weighting

Motivation. The weights λ_c in equation 2 determine the relative influence of the PDE residual and the boundary or initial constraints. Instantaneous gradient statistics describe the current imbalance, while accumulated loss signals summarize which constraints remain difficult over multiple updates. To address ***P1***, ICW combines these two sources of information through a history-dependent weighting rule.

Weighting. The proportional term P_c measures the current gradient imbalance between the residual and the c-th condition. Its implementation is specified in equation 74. For the residual reference channel r and each condition channel c, ICW maintains an integral state initialized at zero:

$$\begin{aligned} I_j &\leftarrow \alpha_z I_j + \sqrt{\mathcal{L}_j}, \qquad j \in \{r, 1, \dots, C\}, \\ \lambda_c &= (1 - \beta_\lambda) P_c + \beta_\lambda \frac{I_c}{I_r + \varepsilon_I}. \end{aligned} \tag{9}$$

Here $0 \le \alpha_z < 1$, $0 \le \beta_\lambda \le 1$, and $\varepsilon_I > 0$ stabilizes the reference denominator.

For fixed α_z, define $E_j = (1 - \alpha_z) I_j$ and $\delta = (1 - \alpha_z)\varepsilon_I$. The same rule becomes

$$\begin{aligned} E_j &\leftarrow \alpha_z E_j + (1 - \alpha_z)\sqrt{\mathcal{L}_j}, \\ \lambda_c &= (1 - \beta_\lambda) P_c + \beta_\lambda \frac{E_c}{E_r + \delta}. \end{aligned} \tag{10}$$

The leaky integral is therefore a scaled exponential moving average. ICW uses the ratio of accumulated signals to supplement instantaneous gradient information. A persistently large condition loss can increase the history-dependent contribution relative to the residual reference, although the resulting weight also depends on the reference signal and the proportional branch.

Local perturbation response. The normalized form permits an explicit analysis of the weight response. For compactness, write $a = \alpha_z$ and $b = \beta_\lambda$.

Proposition 2 (Local perturbation response)**.** *Let $a \in [0, 1)$, $b \in [0, 1]$, and $\delta > 0$ be fixed. Write*

$$E_c = \mu_c + u, \qquad E_r = \mu_r + v, \qquad P_c = \bar{P}_c + \eta,$$

where $\mu_c, \mu_r \ge 0$, and set $D = \mu_r + \delta$. If $|u| \le A$ and $|v| \le B < D$, then

$$\lambda_c - \bar{\lambda}_c = (1 - b)\eta + b\left(\frac{u}{D} - \frac{\mu_c v}{D^2}\right) + bR, \tag{11}$$

where $\bar{\lambda}_c = (1 - b)\bar{P}_c + b\mu_c / D$, and

$$|R| \le \frac{\mu_c B^2}{D^2(D - B)} + \frac{AB}{D(D - B)}. \tag{12}$$

For sinusoidal perturbations of the square-root loss signals, the normalized integral has steady-state transfer function

$$H_a(\omega) = \frac{1-a}{1-ae^{-i\omega}}. \tag{13}$$

Proposition 2 gives a local weight-response prediction with a bounded nonlinear remainder. Increasing a strengthens integral-branch filtering but slows adaptation; the proportional branch remains unfiltered, while small reference denominators amplify the sensitivity of the history-dependent ratio. Derivations appear in Appendix B.4, and accumulator bounds and conditional convergence in Appendix B.3.

3.4 A Unified Accumulation Perspective

LIRB and ICW share a first-order accumulation operator across depth and training iterations. Normalizing LIRB by $s_k = (1-\alpha)v_k$ gives $s_{k+1} = \alpha s_k + (1-\alpha)t_k$, matching the ICW recurrence for $E_j = (1-\alpha_z)I_j$ with square-root loss inputs.

For fixed decay $\rho \in [0, 1)$, an isolated unit input perturbation contributes $(1-\rho)\rho^j$ after j subsequent updates. For prescribed sinusoidal inputs, the steady-state amplitude gain is

$$G_\rho(\omega) = \frac{1-\rho}{\sqrt{1+\rho^2-2\rho\cos\omega}}. \tag{14}$$

Larger ρ prolongs signal retention while strengthening nonzero-frequency attenuation, revealing a common retention–filtering trade-off.

4 Experiments

4.1 Experimental Setup

We evaluate CSF-PINN on Klein–Gordon, Wave, Helmholtz, Allen–Cahn, and inverse Burgers. Baselines include nine weighting methods: EW, GWPINN (Vemuri & Denzler, 2023), NTK (Wang et al., 2022), SA-PINN (McClenny & Braga-Neto, 2023), Causal (Wang et al., 2024b), RBA (Anagnostopoulos et al., 2024), DCGD (Hwang & Lim, 2024), ConFIG (Liu et al., 2025a), and DBPINN (Zhou et al., 2025a); and two architectural methods: PINNsFormer (Zhao et al., 2024) and PirateNet (Wang et al., 2024a).

Sampling and training budgets are matched following (Zhou et al., 2025a). CSF-PINN has comparable parameter counts to the weighting baselines and fewer parameters than the architectural baselines. Training uses Adam, with L-BFGS refinement for inverse Burgers, on a single NVIDIA A100 GPU. We report relative L_2 error and MAE. Benchmark definitions appear in Appendix D.1, with implementation details in the same appendix.

4.2 Main Results

Table 1 reports results on Helmholtz, Burgers, and Wave, with Klein–Gordon and Allen–Cahn in the Appendix. CSF-PINN achieves the best performance under both metrics on all benchmarks, with particularly large gains on Wave (0.1725 vs. 0.5316) and Helmholtz (0.0506 vs. 0.1392). Consistently, Figure 3 shows smaller point-wise errors, steadier convergence, and closer agreement with the reference on Wave. Additional qualitative results for the other benchmarks are provided in the Appendix. Figure 4 shows predicted-versus-reference scatter plots on Wave. From the weakest method to CSF-PINN, the points concentrate onto the diagonal and their color cools, with CSF-PINN the tightest along $y = x$.

4.3 Ablation and Training Analysis

Comparison with two-line combinations. Table 2 combines architecture-based backbones with SA-PINN or DBPINN weighting. CSF-PINN achieves the lowest error on all three benchmarks, showing that simply combining the two lines does not consistently recover its gains.

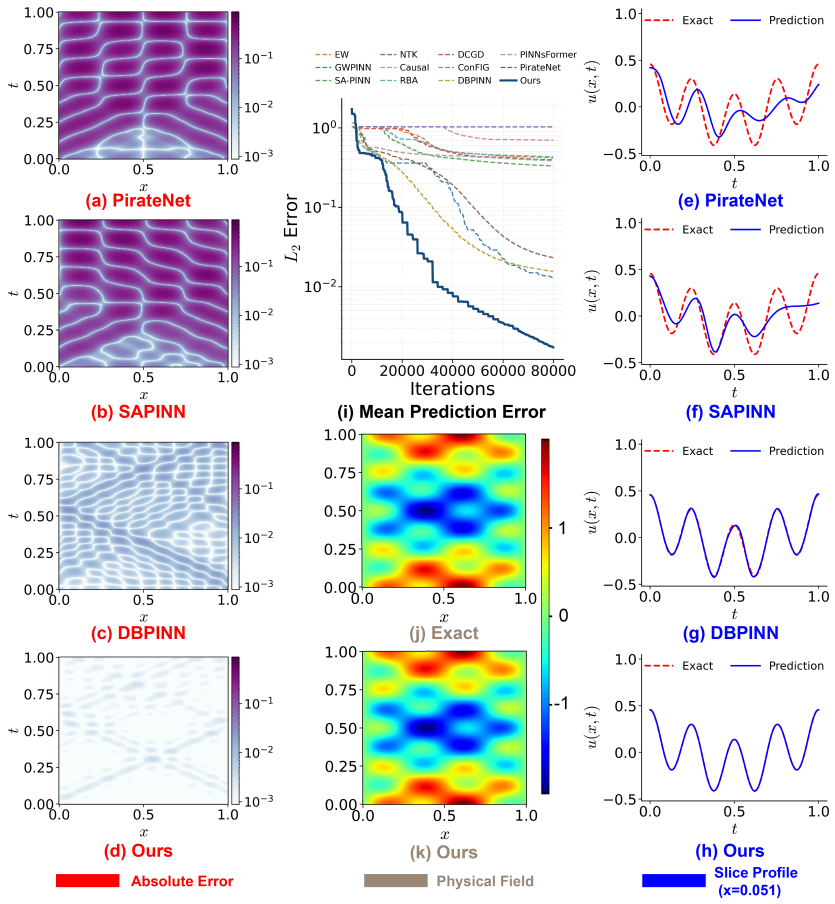

Figure 3: Results on Wave. (a)–(d) point-wise absolute error for PirateNet, SA-PINN, DBPINN, and CSF-PINN; (i) mean prediction-error curves; (j)–(k) reference solution and CSF-PINN prediction; (e)–(h) slice profiles against the reference.

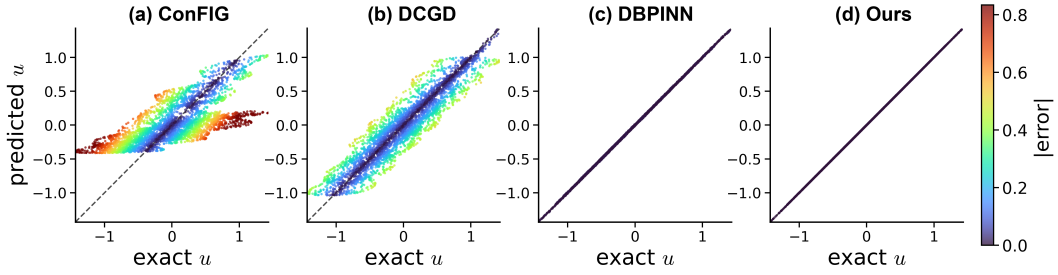

Figure 4: Predicted versus reference values on Wave for four methods, ordered from worst to best by relative L_2 error and colored by point-wise absolute error.

Effect of each component. Table 3 ablates LIRB and ICW. Adding either one to the plain backbone with equal weighting lowers the error on every PDE, and adding both lowers it further, giving the best result in all cases. The improvement from the two together exceeds that of either alone.

Effect of the ICW design. Table 4 evaluates the two design choices in ICW. Setting $\beta_\lambda = 0$ retains only the gradient-statistics term, while setting $\alpha_z = 1$ replaces decayed accumulation with pure accumulation. Both variants yield higher errors than full ICW on all three benchmarks, supporting the complete weighting design.

Table 1: Relative L_2 error (%) and MAE (%) on Helmholtz, Burgers, and Wave. Each entry reports the mean with the standard deviation as a subscript. Best in **bold**. OOM: Out of Memory; N/A: Not Applicable. Results on the remaining benchmarks are given in the Appendix.

Method	Helmholtz		Burgers		Wave	
	L_2	MAE	L_2	MAE	L_2	MAE
PINN (EW)	$2.8907_{\pm 1.6823}$	$0.8241_{\pm 0.7101}$	$4.5162_{\pm 2.9477}$	$0.8474_{\pm 0.5568}$	$38.8128_{\pm 7.7683}$	$17.1709_{\pm 6.3058}$
PINNsFormer	$1.5330_{\pm 0.3333}$	$0.4419_{\pm 0.2376}$	$2.2198_{\pm 1.4564}$	$0.2708_{\pm 0.1862}$	OOM	
PirateNet	$1.3273_{\pm 0.7751}$	$0.2644_{\pm 0.1241}$	$2.5943_{\pm 1.3595}$	$0.2878_{\pm 0.1915}$	$5.6790_{\pm 1.6022}$	$1.5243_{\pm 0.4128}$
GWPINN	$0.5639_{\pm 0.2176}$	$0.1476_{\pm 0.0937}$	$0.6381_{\pm 0.2760}$	$0.1190_{\pm 0.0358}$	$1.3207_{\pm 0.4237}$	$0.5685_{\pm 0.3094}$
SA-PINN	$4.2435_{\pm 1.1562}$	$1.3293_{\pm 0.6302}$	$2.2797_{\pm 1.0095}$	$0.2946_{\pm 0.0936}$	$30.9458_{\pm 7.9634}$	$13.3977_{\pm 3.2191}$
NTK	$0.2030_{\pm 0.0911}$	$0.0502_{\pm 0.0158}$	$0.7700_{\pm 0.1096}$	$0.1250_{\pm 0.0402}$	$0.5316_{\pm 0.2385}$	$0.1439_{\pm 0.0951}$
Causal	N/A		$4.8110_{\pm 1.3961}$	$0.8041_{\pm 0.1094}$	$65.1899_{\pm 24.2735}$	$25.8551_{\pm 14.5064}$
RBA	$5.9242_{\pm 2.9553}$	$1.7106_{\pm 1.0858}$	$6.3036_{\pm 1.3693}$	$1.2430_{\pm 0.4929}$	$39.3238_{\pm 20.6769}$	$17.3425_{\pm 4.4740}$
DCGD	$0.5812_{\pm 0.1722}$	$0.1454_{\pm 0.1216}$	$2.5829_{\pm 1.0975}$	$0.3167_{\pm 0.1202}$	$36.7132_{\pm 7.9218}$	$16.1391_{\pm 5.6469}$
ConFIG	$0.2246_{\pm 0.1003}$	$0.0750_{\pm 0.0146}$	$2.9575_{\pm 0.5123}$	$0.6598_{\pm 0.3313}$	$72.3621_{\pm 25.6453}$	$32.2923_{\pm 14.1058}$
DBPINN	$0.1392_{\pm 0.0477}$	$0.0396_{\pm 0.0095}$	$0.4288_{\pm 0.0338}$	$0.0861_{\pm 0.0277}$	$0.5542_{\pm 0.0863}$	$0.1866_{\pm 0.0748}$
Ours	$\mathbf{0.0506}_{\pm 0.0239}$	$\mathbf{0.0184}_{\pm 0.0041}$	$\mathbf{0.3871}_{\pm 0.0694}$	$\mathbf{0.0578}_{\pm 0.0214}$	$\mathbf{0.1725}_{\pm 0.0384}$	$\mathbf{0.0757}_{\pm 0.0393}$

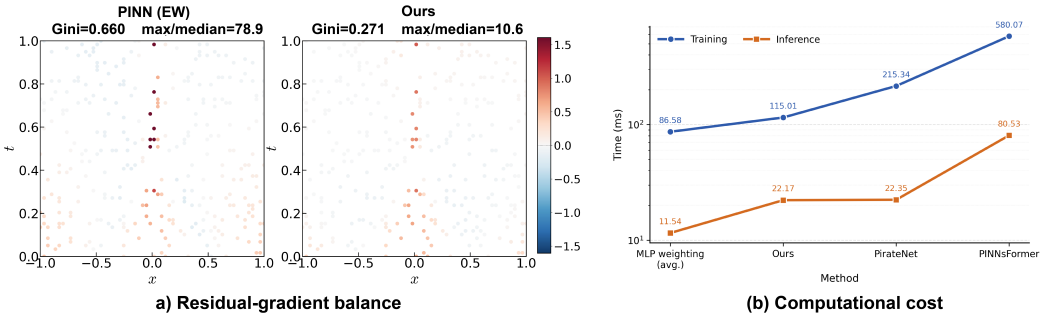

Figure 5: Training dynamics and computational cost. (a) Residual-gradient distribution on Burgers for PINN (EW) and CSF-PINN; color denotes the log-ratio to the panel median. (b) Training and inference time for CSF-PINN, MLP-based weighting methods, PirateNet, and PINNsFormer, with MLP weighting reported as the average.

Table 2: Comparison with architecture-plus-weighting combinations on Helmholtz, Burgers, and Wave, relative L_2 error (%) and MAE (%). Best in **bold**.

Method	Helmholtz		Burgers		Wave	
	L_2	MAE	L_2	MAE	L_2	MAE
PINNsFormer+SA	2.3412	0.7165	2.0781	0.2455	OOM	
PirateNet+SA	0.4589	0.0842	3.5318	0.4162	0.9009	0.1077
PINNsFormer+DB	0.4631	0.0969	1.2793	0.1628	OOM	
PirateNet+DB	0.0936	0.0345	0.9835	0.1022	0.6844	0.0983
Ours	**0.0506**	**0.0184**	**0.3871**	**0.0578**	**0.1725**	**0.0757**

Training Dynamics Figure 5(a) shows a more balanced residual-gradient distribution under CSF-PINN: the Gini coefficient drops from 0.660 to 0.271 and the max-to-median ratio from 78.9 to 10.6. Figure 5(b) shows that CSF-PINN is $5.04\times$ faster in training and $3.63\times$ faster in inference than PINNsFormer, and $1.87\times$ faster to train than PirateNet with comparable inference time. Figure 6 further shows convergence toward PDE-dependent loss weights; full weight trajectories are given in the Appendix.

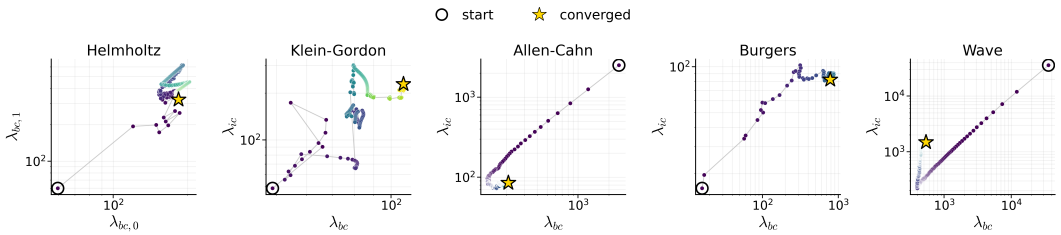

Figure 6: Adaptive loss-weight trajectories across PDEs. Each point is one training iteration, with ◯ and ☆ marking the initial and converged weights. Both axes are log-scaled.

Table 3: Ablation on the two components on Helmholtz, Burgers, and Wave, relative L_2 error (%) and MAE (%). Adding LIRB or ICW alone improves over the baseline, and the two together are best on every PDE. Best in **bold**.

LIRB	ICW	Helmholtz		Burgers		Wave	
		L_2	MAE	L_2	MAE	L_2	MAE
✗	✗	2.8907	0.8241	4.5162	0.8474	38.8128	17.1709
✓	✗	1.0374	0.2849	1.9188	0.1847	10.3845	2.2884
✗	✓	0.1138	0.0236	0.4949	0.1035	0.5093	0.1034
✓	✓	**0.0506**	**0.0184**	**0.3871**	**0.0578**	**0.1725**	**0.0757**

Table 4: ICW design ablation using relative L_2 error (%). Removing either the integral term or its decay degrades performance; the full leaky design is best.

Variant	Helmholtz	Burgers	Wave
No integral (β_λ=0)	0.1847	3.8489	0.8975
Pure accum. (α_z=1)	0.0931	0.9778	0.4756
ICW (leaky)	**0.0506**	**0.3871**	**0.1725**

Composability with Weak Formulations Without modifying LIRB or ICW, we replace the strong-form residual with local conservation moments on an inviscid Burgers entropy-shock problem. CSF-PINN reduces global L_1 error and shock-location MAE by 5.9% and 37.3%, respectively, over a matched weak-form PINN. Detailed settings and results are provided in the Appendix.

5 Limitations

CSF-PINN currently focuses on fixed-PDE PINN training. Extending the framework to parameterized and operator-learning settings remains future work. Our theoretical analysis characterizes collocation-sampling generalization within a fixed PDE. Finally, although we evaluate several representative PDE benchmarks and include weak-form and physics-prior compatibility studies, larger-scale problems, higher-dimensional domains, and complex geometries remain to be explored.

6 Conclusion

CSF-PINN combines decay-weighted architectural propagation with history-dependent loss weighting. LIRB admits an equivalent single-state formulation involving hidden-state damping and an initial-representation skip connection, while ICW combines gradient statistics with accumulated loss signals. The analysis characterizes the architectural state bound and the local perturbation response of the weighting rule. Across five benchmark PDEs, CSF-PINN achieves the lowest error among the compared methods, with ablations supporting the empirical contributions of both components. These findings motivate further investigation of decay-based propagation and weighting, including their dependence on problem structure and training dynamics.

AI USE STATEMENT

Generative AI tools were used solely to assist with language editing and to improve the readability of the manuscript. All AI-assisted edits were reviewed by the authors. We take responsibility for the final content of this work, including all text, claims, results, and artifacts.

REFERENCES

Jephte Abijuru, Mayank Nagda, Phil Ostheimer, Sebastian Josef Vollmer, Marius Kloft, and Sophie Fellenz. Physics-informed residual flows. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Sokratis J Anagnostopoulos, Juan Diego Toscano, Nikolaos Stergiopulos, and George Em Karniadakis. Residual-based attention in physics-informed neural networks. *Computer Methods in Applied Mechanics and Engineering*, 421:116805, 2024.

Shamsulhaq Basir. Investigating and mitigating failure modes in physics-informed neural networks (pinns). *Communications in Computational Physics*, 33(5):1240–1269, 2023.

Maricela Best Mckay, Avleen Kaur, Chen Greif, and Brian Wetton. Near-optimal sketchy natural gradients for physics-informed neural networks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 4005–4019. PMLR, 13–19 Jul 2025.

Donglin Chen, Xiang Gao, Chuanfu Xu, Siqi Wang, Shizhao Chen, Jianbin Fang, and Zheng Wang. Flowdnn: a physics-informed deep neural network for fast and accurate flow prediction. *Frontiers of Information Technology & Electronic Engineering*, 23(2):207–219, 2022.

Huaguan Chen, Yang Liu, and Hao Sun. PINP: physics-informed neural predictor with latent estimation of fluid flows. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*.

Arka Daw, Jie Bu, Sifan Wang, Paris Perdikaris, and Anuj Karpatne. Mitigating propagation failures in physics-informed neural networks using retain-resample-release (R3) sampling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7264–7302. PMLR, 23–29 Jul 2023.

Shota Deguchi and Mitsuteru Asai. Dynamic & norm-based weights to normalize imbalance in back-propagated gradients of physics-informed neural networks. *Journal of Physics Communications*, 7(7):075005, 2023.

Ghazal Farhani, Nima Hosseini Dashtbayaz, Alexander Kazachek, and Boyu Wang. A simple remedy for failure modes in physics informed neural networks. *Neural Networks*, 183:106963, 2025.

Paul Ghanem, Ahmet Demirkaya, Tales Imbiriba, Alireza Ramezani, Zachary Danziger, and Deniz Erdogmus. Learning physics informed neural odes with partial measurements. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 16799–16807, 2025.

Haoqin Hong, Ding Fan, Fubin Dou, Zhili Zhou, Haoran Sun, Congcong Zhu, and Jingrun Chen. Physics-informed deformable Gaussian splatting: Towards unified constitutive laws for time-evolving material field. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4725–4733, 2026.

Youngsik Hwang and Dong-Young Lim. Dual cone gradient descent for training physics-informed neural networks. *Advances in Neural Information Processing Systems*, 37:98563–98595, 2024.

Ameya D Jagtap and George Em Karniadakis. Extended physics-informed neural networks (xpinns): A generalized space-time domain decomposition based deep learning framework for nonlinear partial differential equations. *Communications in Computational Physics*, 28(5), 2020.

Ameya D Jagtap, Ehsan Kharazmi, and George Em Karniadakis. Conservative physics-informed neural networks on discrete domains for conservation laws: Applications to forward and inverse problems. *Computer Methods in Applied Mechanics and Engineering*, 365:113028, 2020.

Feilong Jiang, Xiaonan Hou, Jianqiao Ye, and Min Xia. Mask-pinns: Mitigating internal covariate shift in physics-informed neural networks. *Neural Networks*, pp. 109236, 2026a.

Zheng Jiang, Wei Wang, Gaowei Zhang, and Yi Wang. SSTODE: Ocean-atmosphere physics-informed neural ODEs for sea surface temperature prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 38661–38669, 2026b.

Bum Jun Kim and Gnankan Landry Regis N'guessan. Per-loss adapters for gradient conflict in physics-informed neural networks. *arXiv preprint arXiv:2605.10136*, 2026.

Aditi Krishnapriyan, Amir Gholami, Shandian Zhe, Robert Kirby, and Michael Mahoney. Characterizing possible failure modes in physics-informed neural networks. *Advances in neural information processing systems*, 34:26548–26560, 2021.

Qiang Liu, Mengyu Chu, and Nils Thuerey. Config: Towards conflict-free training of physics informed neural networks. In *International Conference on Learning Representations*, volume 2025, pp. 59531–59566, 2025a.

Xinyang Liu, Junming Hou, Chenxu Wu, Xiaofeng Cong, zihao chen, Shangqi Deng, Junling Li, Liang-Jian Deng, and Bo Liu. Physics-informed neural operator for pansharpening. In *Advances in Neural Information Processing Systems*, volume 38, pp. 95622–95652. Curran Associates, Inc., 2025b.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljacic, Thomas Hou, and Max Tegmark. Kan: Kolmogorov–arnold networks. 2025:70367–70413, 2025c.

Suryanarayana Maddu, Dominik Sturm, Christian L Müller, and Ivo F Sbalzarini. Inverse dirichlet weighting enables reliable training of physics informed neural networks. *Machine Learning: Science and Technology*, 3(1):015026, 2022.

Takashi Matsubara and Takaharu Yaguchi. Number theoretic accelerated learning of physics-informed neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 595–603, 2025.

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

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*, volume 2. MIT press Cambridge, 2012.

Rambod Mojgani, Maciej Balajewicz, and Pedram Hassanzadeh. Lagrangian pinns: A causality-conforming solution to failure modes of physics-informed neural networks. *arXiv preprint arXiv:2205.02902*, 2022.

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.

Frank Shih, Zhenghao Jiang, and Faming Liang. Uncertainty quantification for physics-informed neural networks with extended fiducial inference. In *Advances in Neural Information Processing Systems*, volume 38, pp. 75361–75397. Curran Associates, Inc., 2025.

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. volume 33, pp. 7537–7547, 2020.

Sai Karthikeya Vemuri and Joachim Denzler. Gradient statistics-based multi-objective optimization in physics-informed neural networks. *Sensors*, 23(21):8665, 2023.

Han Wan, Qi Wang, Yuan Mi, Rui Zhang, and Hao Sun. PIMRL: Physics-informed multi-scale recurrent learning for burst-sampled spatiotemporal dynamics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1096–1104, 2026.

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

Sifan Wang, Bowen Li, Yuhan Chen, and Paris Perdikaris. Piratenets: Physics-informed deep learning with residual adaptive networks. *Journal of Machine Learning Research*, 25(402):1–51, 2024a.

Sifan Wang, Shyam Sankaran, and Paris Perdikaris. Respecting causality for training physics-informed neural networks. *Computer Methods in Applied Mechanics and Engineering*, 421: 116813, 2024b.

Sifan Wang, Ananyae Kumar Bhartari, Bowen Li, and Paris Perdikaris. Gradient alignment in physics-informed neural networks: A second-order optimization perspective. In *Advances in Neural Information Processing Systems*, volume 38, pp. 186682–186732, 2025.

Guoquan Wu and Zhe Wu. A multi-objective optimization framework for adaptive weighting in physics-informed machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26885–26893, 2026.

Zhoutong Wu, Mingqing Xiao, Cong Fang, and Zhouchen Lin. Designing universally-approximating deep neural networks: A first-order optimization approach. *IEEE transactions on pattern analysis and machine intelligence*, 46(9):6231–6246, 2024.

Zixue Xiang, Wei Peng, Xu Liu, and Wen Yao. Self-adaptive loss balanced physics-informed neural networks. *Neurocomputing*, 496:11–34, 2022.

Xianliang Xu, Ye Li, and Zhongyi Huang. Refined generalization analysis of the deep ritz method and physics-informed neural networks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 69377–69431. PMLR, 13–19 Jul 2025.

Sen Zhang, Xiao-Wei Guo, Chao Li, Ran Zhao, Canqun Yang, Wei Wang, and Yanxu Zhong. Multi-viscosity physics-informed neural networks for generating ultra high resolution flow field data. *International Journal of Computational Fluid Dynamics*, 37(4):279–297, 2023.

Zhuo Zhang, Xiong Xiong, Sen Zhang, Wei Wang, Xi Yang, Shilin Zhang, and Canqun Yang. A pseudo-time stepping and parameterized physics-informed neural network framework for navier–stokes equations. *Physics of Fluids*, 37(3), 2025.

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

Chenhong Zhou, Jie Chen, Zaifeng Yang, and Ching Eng Png. Dual-balancing for physics-informed neural networks. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2025a.

ChuYu Zhou, Tianyu Li, Chenxi Lan, Rongyu Du, Guoguo Xin, Wei Li, Guoqing Wang, Xun Liu, and Hangzhou Yang. Hybrid boundary physics-informed neural networks for solving navier-stokes equations with complex boundary. In *Advances in Neural Information Processing Systems*, volume 38, pp. 137350–137386. Curran Associates, Inc., 2025b.

Junhong Zou, Wei Qiu, Zhenxu Sun, Xiaomei Zhang, Zhaoxiang Zhang, and Xiangyu Zhu. AdaField: Generalizable surface pressure modeling with physics-informed pre-training and flow-conditioned adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1676–1684, 2026.

APPENDIX

A NOTATION

For convenience, Table 5 summarizes the main notation used throughout the paper.

Table 5: Main notation used in CSF-PINN.

Symbol	Meaning
General PINN notation	
$\mathbf{x}$	Input coordinate
$\Omega, \partial\Omega$	Domain and boundary
u, u_θ	Exact and predicted solutions
θ	Network parameters
$\mathcal{N}$	PDE operator
$\mathcal{B}_c$	Condition operator
$\mathcal{L}_r$	Residual loss
$\mathcal{L}_c$	Condition loss
λ_c	Condition weight
LIRB notation	
z_k	Hidden state
t_k	Transformed feature
v_k	Integral state
T_k	Learnable transform
W_k, V_k, b_k	Transform parameters
α	State-decay coefficient
β, γ	LIRB coefficients
M_t	Bound on $\|t_k\|$
ICW notation	
P_c	Proportional term
I_c	Condition integral
I_r	Residual integral
α_z	Integral-decay coefficient
β_λ	Mixing coefficient
ε_I	Numerical stabilizer
$q_j^{(s)}$	Integral input signal
Λ_c	Weight upper bound

B THEORETICAL DETAILS

This section analyzes the architectural and weighting components of CSF-PINN. We first establish the root condition and approximation capacity of LIRB, followed by depth-uniform state bounds. We then characterize the ICW accumulators and derive the local perturbation response of the weighting rule.

B.1 ROOT CONDITION AND APPROXIMATION CAPACITY OF LIRB

We examine two properties of LIRB. First, we verify the root condition of its homogeneous recurrence. Second, we establish an approximation result through its connection to residual networks. The universality argument uses the equivalent single-state formulation and the residual-network case of (Wu et al., 2024).

Lemma 3 (Root condition for LIRB)**.** *For the recurrence in equation 3, with depth-shared coefficients and $\gamma \neq 0$, eliminating the auxiliary state gives*

$$z_{k+1} - (1+\alpha)z_k + \alpha z_{k-1} = (\beta - \gamma)t_k - \alpha\beta t_{k-1}. \tag{15}$$

The homogeneous characteristic polynomial is $\rho(\lambda) = (\lambda - 1)(\lambda - \alpha)$. Hence the root condition holds for $|\alpha| < 1$.

Proof. Substituting $\gamma v_k = z_{k-1} - z_k + \beta t_{k-1}$ into the updates in equation 3 yields the two-step recurrence. Its homogeneous roots are 1 and α. For $|\alpha| < 1$, the only root on the unit circle is the simple root 1. □

Corollary 4 (Conditional universality of LIRB)**.** *Let $d \geq 2$, $1 \leq m \leq d$, and $1 \leq p < \infty$. Let σ be a non-affine activation satisfying Assumption 1 of (Wu et al., 2024). Consider the family of LIRB networks with arbitrary finite depth L, initialized by $z_0 = x \in [0,1]^d$ and $v_0 = 0$, with $\beta = 1$ and $\gamma = 1 - \alpha$. Their equivalent propagation rule is*

$$z_{k+1} = \alpha z_k + (1-\alpha)x + \alpha T_k(z_k), \qquad k = 0, \ldots, L-1, \tag{16}$$

where $T_k(z) = V_k\sigma(W_k z + b_k)$. The module parameters satisfy $W_k \in \mathbb{R}^{q\times d}$, $V_k \in \mathbb{R}^{d\times q}$, and $b_k \in \mathbb{R}^q$, with fixed module width $q \geq 3$. The coefficient $\alpha \in (0,1)$ is shared across depth and is a free parameter of the network. With an affine readout $h(x) = Az_L(x) + c$, where $A \in \mathbb{R}^{m\times d}$ and $c \in \mathbb{R}^m$, this family is dense in $C([0,1]^d, \mathbb{R}^m)$ under the L^p norm.

Proof. Fix a target function $f \in C([0,1]^d, \mathbb{R}^m)$ and $\varepsilon > 0$. Under the stated activation assumptions, the residual-network case of Theorem 1 in (Wu et al., 2024) provides a finite-depth network

$$\begin{aligned} y_0(x) &= x, \\ y_{k+1}(x) &= y_k(x) + T_k(y_k(x)), \qquad k = 0, \ldots, L-1, \\ h_{\mathrm{res}}(x) &= Ay_L(x) + c, \end{aligned} \tag{17}$$

such that $\|f - h_{\mathrm{res}}\|_{L^p} < \varepsilon/2$, where all L^p norms below are over $[0,1]^d$ with values in $\mathbb{R}^m$.

Construct an LIRB network with the same transforms, depth, and readout:

$$\begin{aligned} z_0^{(\alpha)}(x) &= x, \\ z_{k+1}^{(\alpha)}(x) &= \alpha z_k^{(\alpha)}(x) + (1-\alpha)x + \alpha T_k(z_k^{(\alpha)}(x)), \\ h_\alpha(x) &= Az_L^{(\alpha)}(x) + c. \end{aligned} \tag{18}$$

Extend this recurrence continuously to $\alpha = 1$ for the purpose of taking a limit. At this endpoint, it coincides with the residual recurrence, so $h_1 = h_{\mathrm{res}}$.

Since σ is continuous and the depth is finite, $h_\alpha(x)$ is jointly continuous in (α, x) on the compact set $[1/2, 1] \times [0,1]^d$. Uniform continuity therefore implies

$$\lim_{\alpha\to 1^-} \sup_{x\in[0,1]^d} \|h_\alpha(x) - h_{\mathrm{res}}(x)\| = 0. \tag{19}$$

Hence an $\alpha \in (0,1)$ can be chosen sufficiently close to 1 such that $\|h_\alpha - h_{\mathrm{res}}\|_{L^p} < \varepsilon/2$. The triangle inequality gives

$$\begin{aligned} \|f - h_\alpha\|_{L^p} &\leq \|f - h_{\mathrm{res}}\|_{L^p} + \|h_{\mathrm{res}} - h_\alpha\|_{L^p} \\ &< \varepsilon. \end{aligned} \tag{20}$$

Every selected $\alpha \in (0,1)$ is realizable by the sigmoid parameterization with the finite value $\theta_\alpha = \log(\alpha/(1-\alpha))$. □

This result establishes the approximation capacity of the learnable-decay LIRB family with fixed module width and variable depth. The experiments evaluate finite-depth instances of this family.

B.2 Bounded Propagation of LIRB

The following result uses the coefficients implemented in the main method and establishes a depth-uniform bound on both states.

Proposition 5 (Depth-uniform LIRB states). *Let $\beta = 1$, $\gamma = 1 - \alpha$, $0 < \alpha < 1$, and $v_0 = 0$. Suppose*

$$\|t_k\| = \|T_k(z_k)\| \le M_t \qquad \textit{for every } k. \tag{21}$$

Then

$$v_k = \sum_{j=0}^{k-1} \alpha^{k-1-j} t_j, \tag{22}$$

$$\|v_k\| \le M_t \frac{1-\alpha^k}{1-\alpha} \le \frac{M_t}{1-\alpha}, \tag{23}$$

$$z_k = z_0 + \alpha(v_k - v_0), \tag{24}$$

$$\|z_k\| \le \|z_0\| + \frac{\alpha M_t(1-\alpha^k)}{1-\alpha} \le \|z_0\| + \frac{\alpha M_t}{1-\alpha}. \tag{25}$$

Both states are therefore bounded independently of the depth.

Proof. Unrolling $v_{k+1} = \alpha v_k + t_k$ from $v_0 = 0$ gives equation 22. The triangle inequality and the finite geometric sum give equation 23. Next,

$$\begin{aligned} z_{k+1} - z_k &= t_k - (1-\alpha)v_{k+1} \\ &= v_{k+1} - \alpha v_k - (1-\alpha)v_{k+1} \\ &= \alpha(v_{k+1} - v_k). \end{aligned} \tag{26}$$

Summing equation 26 from 0 to $k-1$ gives the exact identity equation 24; combining it with equation 23 proves equation 25. □

A sufficient condition for the bounded-transform hypothesis is a bounded activation, fixed module width, and uniformly bounded output matrices V_k. More generally, the proposition applies whenever $\|T_k(z_k)\| \le M_t$ holds uniformly across depth. The result is therefore stated under bounded-parameter assumptions. Likewise, the numerical constant deteriorates as $\alpha \to 1$, even though it remains finite for every learned $\alpha \in (0, 1)$.

Interpretation. The proposition bounds the magnitudes of the LIRB states under bounded transformed features. The identity equation 24 also shows that the auxiliary state is determined by the current representation and z_0, consistent with the single-state implementation in equation 6. The bound does not establish improved approximation accuracy or optimization convergence. Table 13 evaluates the complete propagation rule against residual controls, rather than isolating a benefit from an independent auxiliary memory.

B.3 Bounded and Convergent Integral of ICW

For the residual reference channel r and each condition channel c, write

$$\begin{aligned} I_j^{(s)} &= \alpha_z I_j^{(s-1)} + q_j^{(s)}, \qquad q_j^{(s)} = \sqrt{\mathcal{L}_j^{(s)}}, \\ \lambda_c^{(s)} &= (1-\beta_\lambda)P_c^{(s)} + \beta_\lambda \frac{I_c^{(s)}}{I_r^{(s)} + \varepsilon_I}, \end{aligned} \tag{27}$$

where $j \in \{r, 1, \dots, C\}$, $0 < \alpha_z < 1$, $0 \le \beta_\lambda \le 1$, and $\varepsilon_I > 0$. The stabilizer is necessary for a deterministic bounded-ratio statement when the residual integral becomes small.

Proposition 6 (Bounded leaky integral). *Let $I_j^{(0)} = 0$ and suppose $0 \le q_j^{(s)} \le M_j$ for every update. Then*

$$I_j^{(T)} = \sum_{s=1}^{T} \alpha_z^{T-s} q_j^{(s)} \le M_j \frac{1-\alpha_z^T}{1-\alpha_z} \le \frac{M_j}{1-\alpha_z}. \tag{28}$$

Proof. The first equality follows by recursively substituting the update from $I_j^{(0)} = 0$. Bounding every $q_j^{(s)}$ by M_j and summing the geometric series proves the inequalities. □

Proposition 7 (Convergence under a settling loss signal)**.** *If $q_j^{(s)} \to \ell_j$ as $s \to \infty$, then*

$$I_j^{(s)} \longrightarrow I_j^\star = \frac{\ell_j}{1-\alpha_z}. \tag{29}$$

If $q_j^{(s)} = \ell_j$ for every s, the convergence is geometric:

$$\left| I_j^{(T)} - I_j^\star \right| = \frac{\ell_j}{1-\alpha_z} \alpha_z^T. \tag{30}$$

Proof. Let $e_s = I_j^{(s)} - I_j^\star$. Since $I_j^\star = \alpha_z I_j^\star + \ell_j$,

$$e_s = \alpha_z e_{s-1} + (q_j^{(s)} - \ell_j). \tag{31}$$

For any $\epsilon > 0$, choose S such that $|q_j^{(s)} - \ell_j| \le (1-\alpha_z)\epsilon/2$ for $s \ge S$. Unrolling from S gives

$$|e_T| \le \alpha_z^{T-S}|e_S| + \sum_{s=S+1}^{T} \alpha_z^{T-s}|q_j^{(s)} - \ell_j| \le \alpha_z^{T-S}|e_S| + \frac{\epsilon}{2}. \tag{32}$$

The first term is below $\epsilon/2$ for all sufficiently large T, proving equation 29. When the input is constant, direct evaluation of the geometric sum gives equation 30. □

Corollary 8 (Bounded ICW weights)**.** *Suppose $0 \le P_c^{(s)} \le P_{\max,c}$ and $q_c^{(s)} \le M_c$. Then*

$$0 \le \lambda_c^{(s)} \le (1-\beta_\lambda) P_{\max,c} + \beta_\lambda \frac{M_c}{(1-\alpha_z)\varepsilon_I} =: \Lambda_c \tag{33}$$

for every update s.

Proof. Proposition 6 bounds the numerator by $M_c/(1-\alpha_z)$, while $I_r^{(s)} + \varepsilon_I \ge \varepsilon_I$. Substitution into equation 27 gives the result. □

A bounded ratio additionally requires either a positive stabilizer or a positive lower bound on the reference integral. Equation equation 33 makes the precise assumption used later by the collocation-generalization theorem.

Interpretation. These results characterize the ICW accumulator: bounded driving signals yield bounded integral states, and convergent driving signals yield convergent states. Table 4 evaluates the roles of the history-dependent contribution and decay through no-integral and pure-accumulation variants. Appendix B.4 further characterizes the local sensitivity of the resulting weights to perturbations in the condition and reference signals.

B.4 Local Perturbation Response of ICW

Proof of Proposition 2. Using $D = \mu_r + \delta$, direct subtraction gives

$$\begin{aligned} &\frac{\mu_c + u}{D+v} - \left(\frac{\mu_c}{D} + \frac{u}{D} - \frac{\mu_c v}{D^2} \right) \\ &\qquad = \frac{\mu_c v^2}{D^2(D+v)} - \frac{uv}{D(D+v)} =: R. \end{aligned} \tag{34}$$

Substitution into equation 10 establishes equation 11. Since $D + v \ge D - B > 0$, the triangle inequality yields equation 12. In particular, the remainder is second order in jointly small numerator and denominator perturbations.

Steady-state frequency response. Write the normalized integral recurrence as

$$E_j^{(s)} = aE_j^{(s-1)} + (1-a)q_j^{(s)}, \qquad q_j^{(s)} = \sqrt{\mathcal{L}_j^{(s)}}. \tag{35}$$

For a prescribed sinusoidal input

$$q_j^{(s)} = \mu_j + \mathrm{Re}\big(A_j e^{i\omega s}\big), \tag{36}$$

the stationary periodic solution is

$$E_j^{(s)} = \mu_j + \mathrm{Re}\big(H_a(\omega)A_j e^{i\omega s}\big), \qquad H_a(\omega) = \frac{1-a}{1-ae^{-i\omega}}. \tag{37}$$

This follows by substituting a complex sinusoidal particular solution into the recurrence. Any finite initial-state mismatch contributes a transient proportional to a^s.

Suppose the condition signal, reference signal, and proportional branch have complex perturbation amplitudes A_c, A_r, and A_P at the same frequency. Proposition 2 gives the first-order weight amplitude

$$A_\lambda^{(1)} = (1-b)A_P + bH_a(\omega)\left(\frac{A_c}{D} - \frac{\mu_c A_r}{D^2}\right). \tag{38}$$

The nonlinear pointwise error is bounded by b times equation 12, with $A = |H_a(\omega)A_c|$ and $B = |H_a(\omega)A_r| < D$. The mean inputs and perturbation amplitudes are chosen so that the prescribed square-root loss signals remain nonnegative.

Attenuation and response time. The magnitude response satisfies

$$|H_a(\omega)|^2 = \left[1 + \frac{2a(1-\cos\omega)}{(1-a)^2}\right]^{-1}. \tag{39}$$

For fixed nonzero frequency modulo 2π, this decreases with a, because $a/(1-a)^2$ increases on $[0,1)$. For a step from μ to $\mu+\Delta$, with $E^{(0)} = \mu$ and the new input applied from update 1, the state after n updates is

$$E^{(n)} = \mu + \Delta(1-a^n). \tag{40}$$

Thus, stronger attenuation is accompanied by slower adaptation. This step formula describes the filtered state; it is not a general response-time formula for a ratio whose numerator and denominator both change.

Scope. The response analysis treats the driving signals as prescribed and the decay and mixing coefficients as fixed. In PINN training, the signals depend on network parameters and interact with the weight updates. The analysis therefore does not establish closed-loop training stability. It also does not imply that ICW always fluctuates less than proportional-only weighting, because the proportional branch is unfiltered and the branches may have correlated or cancelling perturbations.

C Fixed-PDE Collocation Generalization

The stability results above concern the internal layer and iteration states. We next study whether a physics loss evaluated on finitely many collocation points controls its population counterpart on unseen points. The argument is a standard uniform Rademacher-complexity analysis applied to PINN residual classes (Mohri et al., 2012). This analysis characterizes finite-sample collocation generalization of the trained model.

Let $j = 0$ denote the interior residual and let $j = 1, \ldots, C$ denote boundary, initial, or observation channels. For each channel, let

$$Z_{j,1}, \ldots, Z_{j,n_j} \overset{\text{i.i.d.}}{\sim} \mu_j, \qquad \mathcal{F}_j = \{F_{j,\theta} : \theta \in \Theta\}. \tag{41}$$

Here $F_{j,\theta}$ includes the appropriate differential, boundary, initial, or observation operator applied to u_θ. Let $a_0 = 1$ and $a_c = \lambda_c$. For deterministic spatial weights w_j, define

$$\mathcal{R}_{\boldsymbol{\lambda}}(\theta) = \sum_{j=0}^{C} a_j \mathbb{E}_{\mu_j}[w_j(Z)\ell_j(F_{j,\theta}(Z))]\,, \tag{42}$$

$$\widehat{\mathcal{R}}_{\boldsymbol{\lambda}}(\theta) = \sum_{j=0}^{C} \frac{a_j}{n_j} \sum_{i=1}^{n_j} w_j(Z_{j,i})\ell_j(F_{j,\theta}(Z_{j,i})). \tag{43}$$

Theorem 9 (Uniform collocation-generalization bound)**.** *Assume $0 \le w_j(z) \le W_j$, $\ell_j(0) = 0$, and that ℓ_j is L_j-Lipschitz and bounded by M_j on the range of $\mathcal{F}_j$. Suppose $0 \le \lambda_c \le \Lambda_c$ and write $\Lambda_0 = 1$. Then, with probability at least $1-\delta$, simultaneously for all $\theta \in \Theta$ and all admissible $\boldsymbol{\lambda}$,*

$$\mathcal{R}_{\boldsymbol{\lambda}}(\theta) \le \widehat{\mathcal{R}}_{\boldsymbol{\lambda}}(\theta) + \sum_{j=0}^{C} \Lambda_j \Bigg[2W_j L_j \mathfrak{R}_{n_j}(\mathcal{F}_j) + 3W_j M_j \sqrt{\frac{\log(2(C+1)/\delta)}{2n_j}} \Bigg], \tag{44}$$

where $\mathfrak{R}_n$ denotes expected Rademacher complexity.

Proof. Fix channel j and define

$$\mathcal{G}_j = \{z \mapsto w_j(z)\ell_j(F_{j,\theta}(z)) : \theta \in \Theta\}\,. \tag{45}$$

Each member of $\mathcal{G}_j$ is bounded in absolute value by $W_j M_j$. The standard one-sided uniform Rademacher inequality gives, with failure probability at most $\delta/(C+1)$,

$$\mathbb{E}_{\mu_j}[w_j\ell_j(F_{j,\theta})] \le \frac{1}{n_j}\sum_{i=1}^{n_j} w_j(Z_{j,i})\ell_j(F_{j,\theta}(Z_{j,i})) + 2\mathfrak{R}_{n_j}(\mathcal{G}_j) + 3W_j M_j \sqrt{\frac{\log(2(C+1)/\delta)}{2n_j}} \tag{46}$$

simultaneously for all θ.

For a fixed sample, the coordinate map $s \mapsto w_j(Z_{j,i})\ell_j(s)$ vanishes at zero and is at most $W_j L_j$-Lipschitz. The contraction inequality therefore gives

$$\mathfrak{R}_{n_j}(\mathcal{G}_j) \le W_j L_j \mathfrak{R}_{n_j}(\mathcal{F}_j). \tag{47}$$

A union bound makes equation 46 hold for all $C+1$ channels. On this event, multiply each channel inequality by its realized nonnegative coefficient a_j. The empirical term retains a_j, whereas the deviation term is bounded using $a_j \le \Lambda_j$. Summing over channels proves equation 44.

The channel-wise event is uniform in θ and is formed before the eventual choice of $\boldsymbol{\lambda}$. It therefore contains the trained parameter and any data-dependent ICW weights satisfying the stated bounds. □

Corollary 8 supplies one sufficient choice of Λ_c when the proportional term is bounded and the integral ratio uses the positive stabilizer. Alternatively, an explicitly clipped ICW implementation directly supplies the required coefficient box. The theorem therefore uses the bounded-coefficient condition supplied by either the stabilized ICW ratio or explicit clipping.

C.1 An Explicit Residual-Class Bound

The differential operator changes across PDEs, so Theorem 9 retains the residual-class complexity explicitly. The next proposition shows that it is finite under a compact-parameter assumption.

Proposition 10 (Compact parameter class)**.** *Let $\Theta \subset \mathbb{R}^P$ lie in a Euclidean ball of radius B_θ. Suppose, for one residual channel,*

$$|F_\theta(z)| \le B_F, \qquad |F_\theta(z) - F_{\theta'}(z)| \le L_\theta \|\theta - \theta'\|_2 \tag{48}$$

uniformly in z. Then, for every $\epsilon > 0$,

$$\mathfrak{R}_n(\mathcal{F}) \le \epsilon + B_F \sqrt{\frac{2P}{n} \log\left(1 + \frac{2B_\theta L_\theta}{\epsilon}\right)}. \tag{49}$$

Choosing $\epsilon = B_F/\sqrt{n}$ gives $\mathfrak{R}_n(\mathcal{F}) = O(\sqrt{P \log n/n})$ when the other constants are fixed.

Proof. An ϵ/L_θ-net of the radius-B_θ parameter ball has cardinality at most

$$N \le \left(1 + \frac{2B_\theta L_\theta}{\epsilon}\right)^P. \tag{50}$$

The Lipschitz condition maps this parameter net to an ϵ-net of $\mathcal{F}$ in empirical supremum norm. Replacing a function by its closest net element changes the empirical Rademacher supremum by at most ϵ. Every evaluation vector of a net function has Euclidean norm at most $B_F\sqrt{n}$. Massart's finite-class lemma then gives

$$\mathfrak{R}_n(\mathcal{F}) \le \epsilon + B_F \sqrt{\frac{2 \log N}{n}}. \tag{51}$$

Substitution of the covering-number bound proves equation 49. □

For a compact input domain, compact parameter set, and a smooth activation, the network, its finite-order derivatives, and any fixed finite-order PDE residual are continuous on a compact set. Thus B_F and L_θ are finite. The compactness assumption is imposed through a bounded parameter set and is independent of the LIRB recurrence.

C.2 From Physics Risk to Solution Error

Converting population physics risk into solution error additionally requires a stability estimate for the specific PDE.

Corollary 11 (Solution error under PDE stability)**.** *Suppose the fixed PDE has a unique admissible solution $u^\star$ and satisfies*

$$\|u - u^\star\|_X \le C_{\mathrm{stab}} \sum_{j=0}^{C} \|F_j(u)\|_{L^p(\mu_j)}. \tag{52}$$

Assume $w_j \ge w_j^- > 0$, $a_j \ge a_j^- > 0$, and $\ell_j(s) = |s|^p$. Then

$$\|u_\theta - u^\star\|_X \le C_{\mathrm{stab}} \mathcal{R}_{\boldsymbol{\lambda}}(\theta)^{1/p} \sum_{j=0}^{C} (a_j^- w_j^-)^{-1/p}. \tag{53}$$

Proof. All risk contributions are nonnegative, so for every channel,

$$a_j^- w_j^- \|F_j(u_\theta)\|^p_{L^p(\mu_j)} \le \mathcal{R}_{\boldsymbol{\lambda}}(\theta). \tag{54}$$

Taking pth roots, summing over channels, and applying equation 52 proves the claim. □

Combining Corollary 11 with Theorem 9 yields a high-probability solution-error bound whenever equation 52 is available. The norm and stability constant are PDE dependent. Coercive elliptic problems admit energy-norm estimates, whereas scalar conservation laws require L_1 stability together with entropy admissibility.

C.3 Specializations Used in the Auxiliary Studies

Smooth L_1 loss. For the nonsmooth Burgers experiment,

$$\rho_{1,\epsilon}(s) = \sqrt{s^2 + \epsilon^2} - \epsilon \tag{55}$$

satisfies $\rho_{1,\epsilon}(0) = 0$ and $|\rho'_{1,\epsilon}(s)| \le 1$. It therefore meets the Lipschitz assumption with $L = 1$. Moreover,

$$|s| \le \rho_{1,\epsilon}(s) + \epsilon. \tag{56}$$

For normalized channel measures and $p = 1$, replacing $|s|$ by $\rho_{1,\epsilon}$ adds at most $C_{\mathrm{stab}}(C+1)\epsilon$ to the stability conversion.

Empirically normalized distance weights. Let $0 < w^- \le w(z) \le W$ and define

$$\begin{aligned} \mathcal{R}^{\mathrm{sn}}(\theta) &= \frac{\mathbb{E}[w(Z)\ell(F_\theta(Z))]}{\mathbb{E}[w(Z)]}, \\ \widehat{\mathcal{R}}^{\mathrm{sn}}(\theta) &= \frac{\sum_{i=1}^n w(Z_i)\ell(F_\theta(Z_i))}{\sum_{i=1}^n w(Z_i)}. \end{aligned} \tag{57}$$

The one-channel event in Theorem 9 implies

$$\begin{aligned} \mathcal{R}^{\mathrm{sn}}(\theta) \le \frac{W}{w^-}\widehat{\mathcal{R}}^{\mathrm{sn}}(\theta) + \frac{1}{w^-}\Bigg[& 2WL\mathfrak{R}_n(\mathcal{F}) \\ & + 3WM\sqrt{\frac{\log(2/\delta)}{2n}}\Bigg]. \end{aligned} \tag{58}$$

Indeed, the unnormalized empirical numerator equals $\overline{w}_n\widehat{\mathcal{R}}^{\mathrm{sn}}$ with $\overline{w}_n \le W$, while $\mathbb{E}[w] \ge w^-$. For the experimental weight $w(x,t) = 0.1 + \exp[-|x - t/2|/0.12]$, one has $w^- = 0.1$ and $W = 1.1$.

Weak conservation moments. Let $Z_1, \dots, Z_n$ be sampled from the normalized collocation distribution μ on Ω. For fixed compactly supported test functions $\phi_1, \dots, \phi_Q$, absorb the known domain-volume factor $\kappa = |\Omega|$ into

$$h_{q,\theta}(z) = \kappa\left[u_\theta(z)\partial_t\phi_q(z) + f(u_\theta(z))\partial_x\phi_q(z)\right], \tag{59}$$

Let $M_q(\theta) = \mathbb{E}[h_{q,\theta}(Z)]$ and $\widehat{M}_q(\theta) = n^{-1}\sum_i h_{q,\theta}(Z_i)$. If $|h_{q,\theta}| \le B_q$ and $\mathcal{H}_q = \{h_{q,\theta} : \theta \in \Theta\}$, then with probability at least $1 - \delta$, simultaneously for all θ,

$$\begin{aligned} \frac{1}{Q}\sum_{q=1}^Q \rho_{1,\epsilon}(M_q(\theta)) \le \frac{1}{Q}\sum_{q=1}^Q \rho_{1,\epsilon}(\widehat{M}_q(\theta)) \\ + \frac{2}{Q}\sum_{q=1}^Q \mathfrak{R}_n(\mathcal{H}_q) \\ + \frac{3}{Q}\sum_{q=1}^Q B_q\sqrt{\frac{\log(4Q/\delta)}{2n}}. \end{aligned} \tag{60}$$

To prove this, apply the two-sided uniform Rademacher inequality to each $\mathcal{H}_q$, take a union bound over q, and use that $\rho_{1,\epsilon}$ is 1-Lipschitz. Boundary and initial channels, on which ICW acts in this experiment, are covered directly by Theorem 9.

Finite-group symmetry projection. Let a finite group G act on the input domain and let $\chi : G \to \{-1, 1\}$ be a character. Define

$$(\Pi_\chi f)(z) = \frac{1}{|G|}\sum_{g\in G}\chi(g)f(g^{-1}z). \tag{61}$$

For any sample $S = (z_1, \dots, z_n)$,

$$\widehat{\mathfrak{R}}_S(\Pi_\chi \mathcal{F}) \le \frac{1}{|G|} \sum_{g \in G} \widehat{\mathfrak{R}}_{g^{-1}S}(\mathcal{F}). \tag{62}$$

Indeed, substitute equation 61 into the empirical Rademacher supremum, upper-bound the supremum of the group sum by the sum of the suprema, and use the symmetry of the Rademacher signs to remove $\chi(g)$. If the sampling distribution is G-invariant, taking expectation over S gives

$$\mathfrak{R}_n(\Pi_\chi \mathcal{F}) \le \mathfrak{R}_n(\mathcal{F}). \tag{63}$$

Thus exact finite-group projection does not enlarge expected class complexity.

Scope. The theorem establishes fixed-PDE collocation generalization, where sampled physics risk controls population physics risk at unseen points drawn from the same channel distributions. Extension to parameterized PDE families would require a conditional model and an explicit sampling distribution over PDE parameters.

D Experimental Details

D.1 Benchmark Details

This section specifies the governing equations, domains, boundary and initial conditions, and reference solutions for the five benchmarks.

Klein–Gordon. We consider the nonlinear equation

$$u_{tt} - u_{xx} + u^3 = f(x,t), \qquad (x,t) \in [0,1] \times [0,1]. \tag{64}$$

The forcing term is chosen so that the exact solution is

$$u^\star(x,t) = x\cos(5\pi t) + (xt)^3. \tag{65}$$

Dirichlet boundary values are prescribed by $u^\star$ at $x = 0$ and $x = 1$, with initial conditions $u(x,0) = x$ and $u_t(x,0) = 0$.

Wave. We consider

$$u_{tt} - 4u_{xx} = 0, \qquad (x,t) \in [0,1] \times [0,1], \tag{66}$$

with homogeneous Dirichlet boundaries $u(0,t) = u(1,t) = 0$ and initial conditions

$$u(x,0) = \sin(\pi x) + \tfrac{1}{2}\sin(4\pi x), \qquad u_t(x,0) = 0. \tag{67}$$

The exact solution contains mixed-frequency components:

$$u^\star(x,t) = \sin(\pi x)\cos(2\pi t) + \tfrac{1}{2}\sin(4\pi x)\cos(8\pi t). \tag{68}$$

Helmholtz. We consider the time-independent equation

$$u_{xx} + u_{yy} + u = q(x,y), \qquad (x,y) \in [-1,1]^2, \tag{69}$$

with homogeneous Dirichlet conditions on all four boundaries. The source term is

$$q(x,y) = (1 - 17\pi^2)\sin(\pi x)\sin(4\pi y), \tag{70}$$

giving the exact solution $u^\star(x,y) = \sin(\pi x)\sin(4\pi y)$. There is no initial condition. The boundaries at $x = \pm 1$ and $y = \pm 1$ form two groups, yielding two boundary-loss terms.

Allen–Cahn. We consider the reaction–diffusion equation

$$u_t - 0.0001\, u_{xx} + 5u^3 - 5u = 0, \qquad (x,t) \in [-1,1] \times [0,1], \tag{71}$$

with periodic boundary conditions

$$u(-1,t) = u(1,t), \qquad u_x(-1,t) = u_x(1,t), \tag{72}$$

and initial condition $u(x,0) = x^2\cos(\pi x)$. The reference solution of (McClenny & Braga-Neto, 2023) is used for evaluation.

Table 6: Dataset and training schedule per PDE (N^f interior, N^{bc} boundary, N^{ic} initial, N^d observation points), following the standard settings (Zhou et al., 2025a).

PDE	N^f	N^{bc}	N^{ic}	N^d	Optim.-Epochs
Klein–Gordon	10000	2000	1000	–	Adam-20k
Wave	20000	2500	2500	–	Adam-80k
Helmholtz	20000	400	–	–	Adam-30k
Allen–Cahn	20000	100	512	–	Adam-80k
Burgers (inv.)	10000	200	100	300	Adam-5k+L-BFGS

Burgers (inverse). We consider

$$u_t + C_1 u u_x - C_2 u_{xx} = 0, \qquad (x,t) \in [-1,1] \times [0,1], \tag{73}$$

with initial condition $u(x,0) = -\sin(\pi x)$ and homogeneous Dirichlet boundaries $u(-1,t) = u(1,t) = 0$. The unknown coefficients C_1 and C_2 are identified using additional observations sampled from the reference solution (Raissi et al., 2019). Their true values are $C_1 = 1$ and $C_2 = 0.01/\pi$.

D.2 Implementation Details

All methods are implemented in PyTorch and run on a single NVIDIA A100 GPU. For each PDE, CSF-PINN uses the standard collocation, boundary, and initial point settings adopted in prior work (Zhou et al., 2025a). The backbone width H, bottleneck size q, and block count K are chosen to keep the parameter count comparable to the MLP weighting baselines and below the architecture-based baselines. The LIRB coefficient α is learnable through a sigmoid reparameterization, initialized at $\alpha = 1/3$ (with $\beta = 1$ and $\gamma = 1 - \alpha$), and the block count is fixed to $K = 6$. Optimization uses Adam, with L-BFGS as a final refinement on the Burgers inverse problem. Table 6 lists the dataset and training schedule, and Table 7 details the backbone architecture and parameter count of every method.

Proportional term of ICW. The proportional part P_c in equation 9 measures the gradient imbalance between the residual and the c-th condition from the current-iteration back-propagated gradient statistics. Following the gradient-statistics rule of (Wang et al., 2021), it uses the maximum residual gradient against the mean condition gradient,

$$P_c \;=\; \frac{\max\left|\nabla_\theta \mathcal{L}_r\right|}{\overline{\left|\nabla_\theta \mathcal{L}_c\right|}}, \tag{74}$$

where $\max|\cdot|$ and $\overline{|\cdot|}$ are taken over the entries of the back-propagated gradient.

E Additional Results

This section is divided into two parts: *Quantitative Results*, which report the relative L_2 error and MAE on the two held-out benchmarks (Klein–Gordon and Allen–Cahn) together with the corresponding combination and ablation studies; and *Qualitative Results*, which provide the full per-PDE visualizations for all five benchmarks.

E.1 Quantitative Results

The main text reports the relative L_2 error and MAE on Helmholtz, Burgers, and Wave, and defers Klein–Gordon and Allen–Cahn to this Appendix. The four tables below mirror, on these two benchmarks, the four tables of the main text: the full baseline comparison (Table 9), the architecture-plus-weighting combination (Table 10), the two-component ablation (Table 11), and the ICW-design ablation (Table 12). The two benchmarks probe different regimes: Klein–Gordon is a smooth nonlinear wave equation on which several methods already reach sub-percent error, whereas Allen–Cahn is a stiff reaction–diffusion problem with sharp transitions on which many methods fail outright.

Table 7: Backbone architecture, key hyper-parameters, and parameter count of each method on every PDE. The nine λ-rule baselines share one MLP backbone per PDE and differ only in the weighting scheme. The CSF-PINN backbone is configured to remain within 20% of the parameter count of the shared MLP backbone used by the weighting baselines on each PDE, while remaining smaller than both architecture-based baselines. On the Burgers inverse problem, every model additionally has two trainable PDE coefficients.

PDE	Method	Backbone	Key hyper-params	Params
Helmholtz	λ-rule	MLP $[2, 50^4, 1]$	depth 5, width 50	7,851
	PINNsFormer	Transformer	d_m=24, N=1	11,417
	PirateNet	Fourier+gated res.	h=32, B=3	11,652
	CSF-PINN	LIRB	H=50, q=15, K=6	**9,291**
Klein–Gordon	λ-rule	MLP $[2, 50^3, 1]$	depth 4, width 50	5,301
	PINNsFormer	Transformer	d_m=20, N=1	7,513
	PirateNet	Fourier+gated res.	h=28, B=3	8,964
	CSF-PINN	LIRB	H=50, q=10, K=6	**6,261**
Allen–Cahn	λ-rule	MLP $[2, 128^4, 1]$	depth 5, width 128	50,049
	PINNsFormer	Transformer	d_m=42, N=2	54,389
	PirateNet	Fourier+gated res.	h=60, B=4	51,305
	CSF-PINN	LIRB	H=128, q=32, K=6	**49,857**
Burgers (inv.)	λ-rule	MLP $[2, 20^8, 1]$	depth 9, width 20	3,023
	PINNsFormer	Transformer	d_m=16, N=1	5,003
	PirateNet	Fourier+gated res.	h=22, B=3	5,594
	CSF-PINN	LIRB	H=20, q=14, K=6	**3,527**
Wave	λ-rule	MLP $[2, 500^5, 1]$	depth 6, width 500	1,004,001
	PINNsFormer	Transformer	d_m=176, N=3	1,384,769
	PirateNet	Fourier+gated res.	h=300, B=4	1,264,505
	CSF-PINN	LIRB	H=500, q=200, K=6	**1,203,201**

Full baseline comparison. Table 9 compares CSF-PINN against all eleven baselines. CSF-PINN attains the lowest error on both benchmarks under both metrics. On Klein–Gordon it reaches a relative L_2 error of 0.2719% and MAE of 0.0911%, improving on the strongest baseline DBPINN (0.7295% L_2) by a factor of about 2.7, and on the next best ConFIG (0.7831%) and NTK (0.8385%) by a similar margin. Causal reaches 30.06%, RBA 20.23%, SA-PINN 15.68%, while the equally weighted PINN and DCGD reach 8.90% and 10.25%, respectively. The two architecture baselines, PINNsFormer (4.19%) and PirateNet (4.29%), improve over the equally weighted PINN but still trail the strongest weighting methods.

On the stiff Allen–Cahn problem the spread is wider: EW (50.19%), RBA (56.46%), and DCGD (51.94%) all exceed 50% relative L_2 error, while SA-PINN (7.14%), Causal (7.42%), NTK (9.02%), GWPINN (9.41%), and DBPINN (9.09%) cluster between 7% and 10%. The strongest baseline here is ConFIG at 5.06%, compared with 3.91% for CSF-PINN. CSF-PINN also exhibits relatively small standard deviations on both benchmarks (± 0.0711 on Klein–Gordon and ± 1.2214 on Allen–Cahn). The strongest baseline differs across the two regimes, whereas CSF-PINN achieves the lowest error on both.

Comparison with a two-line combination. Table 10 pairs an architecture-based method with a weighting-based one, repeating on these two benchmarks the two-line combination study of the main text. Simply stacking the two remedies is not enough: PINNsFormer+SA-PINN reaches 6.89% on Klein–Gordon and 6.68% on Allen–Cahn, and PirateNet+SA-PINN reaches 0.32% and 4.42% respectively. CSF-PINN improves on both combinations on each benchmark (0.27% and 3.91%). The contrast is sharpest on Klein–Gordon, where PINNsFormer+SA-PINN remains an order of magnitude worse, and the strongest combination PirateNet+SA-PINN is still behind CSF-PINN despite using a heavier backbone.

Effect of each component. Table 11 ablates the two components against the equally weighted baseline (8.90% on Klein–Gordon, 50.19% on Allen–Cahn). Adding LIRB alone lowers the error

Table 8: Coefficient identification on the inverse Burgers problem. Absolute errors are reported in units of 10^{-3}. Best in **bold**.

Method	$\hat{C}_1$	$\hat{C}_2$	$\lvert\hat{C}_1 - C_1\rvert \times 10^3$	$\lvert\hat{C}_2 - C_2\rvert \times 10^3$
EW	0.989576	0.006414	10.4238	3.2310
GWPINN	0.999245	0.003543	0.7552	0.3600
SA-PINN	0.996714	0.004591	3.2863	1.4084
NTK	1.001475	0.003620	1.4747	0.4370
Causal	0.979868	0.006541	20.1325	3.3580
RBA	0.973208	0.008328	26.7916	5.1452
DCGD	0.997180	0.004783	2.8198	1.5997
ConFIG	0.993093	0.005225	6.9072	2.0421
DBPINN	1.000315	0.003436	0.3152	0.2531
Ours	**1.000121**	**0.003420**	**0.1206**	**0.2368**

to 3.84% and 6.48%, and adding ICW alone lowers it to 0.67% and 6.04%. Combining the two gives the best result on both benchmarks (0.27% and 3.91%). Two observations follow. First, each component helps on its own on both PDEs, so neither the architecture-side nor the optimization-side feedback is redundant. Second, the components are complementary: on Klein–Gordon ICW carries most of the gain while LIRB adds a further refinement, whereas on the stiff Allen–Cahn problem LIRB and ICW each remove only part of the error in isolation (6.48% and 6.04%) and their combination drops well below either (3.91%), so the joint improvement exceeds what either part achieves alone.

Effect of the ICW design. Table 12 evaluates the two design choices in ICW. Removing the integral contribution gives 0.95% error on Klein–Gordon and 9.49% on Allen–Cahn. Pure accumulation gives 0.78% and 5.48%, respectively. Without temporal discounting, early loss contributions can retain influence and delay adaptation to changes in relative loss levels. Full ICW achieves the lowest error on both benchmarks.

Coefficient Identification on Inverse Burgers We further evaluate the identified coefficients on the inverse Burgers problem. The true values are $C_1 = 1$ and $C_2 = 0.01/\pi$. Table 8 reports the identified values and absolute errors.

CSF-PINN yields the lowest absolute error for both coefficients among the compared methods.

E.2 Qualitative Results

This subsection provides the full per-PDE visualizations referenced in the main text. For each benchmark we show, against the weighting baselines, a slice profile and a point-wise absolute-error map; against the architecture baselines, a combined error-map-and-slice comparison; and, where applicable, a predicted-versus-reference scatter plot. The adaptive-weight trajectories are gathered at the end.

Klein–Gordon. Figures 9–12 report the Klein–Gordon results. Across the weighting baselines, CSF-PINN tracks the reference most closely and leaves the smallest error, while Causal, RBA, and SA-PINN diverge over the later part of the domain.

Wave. Figures 13–15 report the Wave results. The stiff, high-frequency solution defeats most baselines in phase and amplitude over time, while CSF-PINN stays aligned with the reference.

Helmholtz. Figures 16–19 report the Helmholtz results. CSF-PINN gives the most uniform and lowest error over the domain.

Allen–Cahn. Figures 20–23 report the Allen–Cahn results. Several weighting baselines collapse toward zero or negative predictions as t grows; CSF-PINN follows the reference.

Burgers (inverse). Figures 24–27 report the inverse Burgers results.

Table 9: Relative L_2 error (%) and MAE (%) on Klein–Gordon and Allen–Cahn. Each entry reports the mean with the standard deviation as a subscript. Best in **bold**.

Method	Klein–Gordon		Allen–Cahn	
	L_2	MAE	L_2	MAE
PINN (EW)	$8.8999_{\pm 2.2306}$	$2.9695_{\pm 1.3540}$	$50.1906_{\pm 14.0739}$	$18.8249_{\pm 7.9602}$
PINNsFormer	$4.1858_{\pm 1.0948}$	$1.3680_{\pm 0.8523}$	$9.7861_{\pm 3.4842}$	$2.4256_{\pm 0.6787}$
PirateNet	$4.2940_{\pm 1.2707}$	$1.4565_{\pm 0.7692}$	$5.8911_{\pm 1.0952}$	$1.0433_{\pm 0.4004}$
GWPINN	$1.4131_{\pm 0.4988}$	$0.5037_{\pm 0.2495}$	$9.4062_{\pm 2.3423}$	$2.3015_{\pm 0.7248}$
SA-PINN	$15.6759_{\pm 4.4472}$	$4.9050_{\pm 1.7611}$	$7.1439_{\pm 2.4865}$	$1.6042_{\pm 0.4266}$
NTK	$0.8385_{\pm 0.1024}$	$0.2656_{\pm 0.0782}$	$9.0172_{\pm 2.4317}$	$2.1769_{\pm 1.0296}$
Causal	$30.0567_{\pm 8.0393}$	$7.1537_{\pm 3.2616}$	$7.4171_{\pm 1.6176}$	$1.7292_{\pm 0.7845}$
RBA	$20.2295_{\pm 6.0135}$	$6.7901_{\pm 3.0671}$	$56.4592_{\pm 20.9296}$	$21.3826_{\pm 9.7528}$
DCGD	$10.2469_{\pm 3.5132}$	$3.3905_{\pm 1.4769}$	$51.9379_{\pm 12.2860}$	$20.7305_{\pm 8.4785}$
ConFIG	$0.7831_{\pm 0.0857}$	$0.2439_{\pm 0.0337}$	$5.0555_{\pm 1.1743}$	$1.1863_{\pm 0.4361}$
DBPINN	$0.7295_{\pm 0.1789}$	$0.2373_{\pm 0.0936}$	$9.0872_{\pm 4.7799}$	$2.7925_{\pm 1.8682}$
Ours	$\mathbf{0.2719}_{\pm 0.0711}$	$\mathbf{0.0911}_{\pm 0.0133}$	$\mathbf{3.9106}_{\pm 1.2214}$	$\mathbf{0.6872}_{\pm 0.2958}$

Table 10: Comparison with architecture-plus-weighting combinations on Klein–Gordon and Allen–Cahn, relative L_2 error (%) and MAE (%). Best in **bold**.

Method	Klein–Gordon		Allen–Cahn	
	L_2	MAE	L_2	MAE
PINNsFormer+SA-PINN	6.8903	2.1628	6.6845	1.9470
PirateNet+SA-PINN	0.3246	0.1689	4.4206	0.7489
Ours	**0.2719**	**0.0911**	**3.9106**	**0.6872**

Table 11: Ablation on the two components on Klein–Gordon and Allen–Cahn, relative L_2 error (%) and MAE (%). Best in **bold**.

LIRB	ICW	Klein–Gordon		Allen–Cahn	
		L_2	MAE	L_2	MAE
✗	✗	8.8999	2.9695	50.1906	18.8249
✓	✗	3.8421	0.9485	6.4824	1.9453
✗	✓	0.6724	0.1293	6.0376	1.1385
✓	✓	**0.2719**	**0.0911**	**3.9106**	**0.6872**

Table 12: Ablation on the ICW design on Klein–Gordon and Allen–Cahn, relative L_2 error (%). Best in **bold**.

Variant	Klein–Gordon	Allen–Cahn
No integral (β_λ=0)	0.9482	9.4884
Pure accum. (α_z=1)	0.7833	5.4838
ICW (leaky)	**0.2719**	**3.9106**

Adaptive-weight trajectories. Figure 28 traces the adaptive loss weights over training for every PDE, together with the discovered coefficients on the inverse Burgers problem. The weights settle to PDE-specific steady states, and on the inverse problem the estimated coefficients converge to their true values $C_1 = 1$ and $C_2 = 0.01/\pi$.

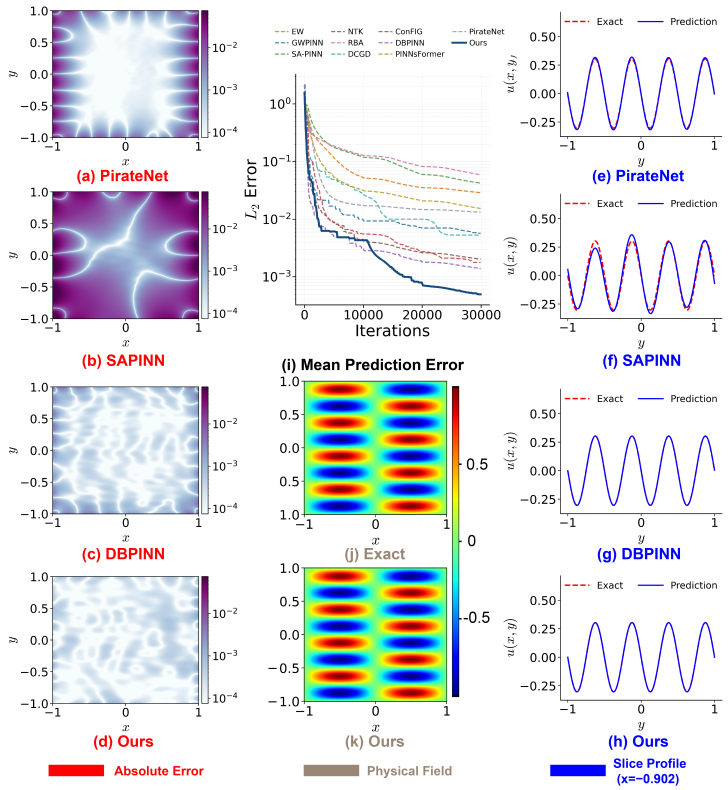

Figure 7: Results on Helmholtz, with the same panel layout as Figure 3. CSF-PINN gives the smallest point-wise error and the prediction closest to the reference.

Relative contributions of LIRB and ICW. The larger standalone gain of ICW is consistent with the broader PINN literature, where loss imbalance is recognized as a major source of training failure and adaptive loss balancing can substantially improve accuracy (Wang et al., 2022; McClenny & Braga-Neto, 2023; Zhou et al., 2025a). Our claim is therefore not that the architecture and weighting components contribute equally in magnitude. Rather, they act on complementary axes: ICW addresses imbalance across loss channels, whereas LIRB introduces persistent state across network depth. Importantly, the gain from LIRB remains substantial after ICW is already applied. On Wave, replacing a state-reset architecture by the persistent-state LIRB reduces relative L_2 error from 0.8542% to 0.1725%.

E.3 Further Ablations

Is persistent state more than a residual pathway? Recent work suggests that residual pathways themselves can explain part of the architectural improvement observed in PINNs. For example, (Abijuru et al., 2026) show through controlled ablations that replacing attention or state-space modules with simpler mappings can preserve, and in some cases improve, accuracy. Their results point to residual pathways as an important source of stability, and their ResPINN follows the standard residual form $z_{k+1} = z_k + t_k$.

We compare standard residual, state-reset, and full LIRB updates while keeping ICW fixed. Under $\beta = 1$ and $\gamma = 1 - \alpha$, resetting the auxiliary state at every block yields $z_{k+1} = z_k + \alpha T_k(z_k)$. Full LIRB instead follows the equivalent propagation rule in equation 6, which includes hidden-state damping and an initial-representation skip connection.

Table 13 shows that full LIRB reduces relative L_2 error by 63.5% on Helmholtz and 79.8% on Wave compared with the state-reset control. This supports the complete propagation rule over the

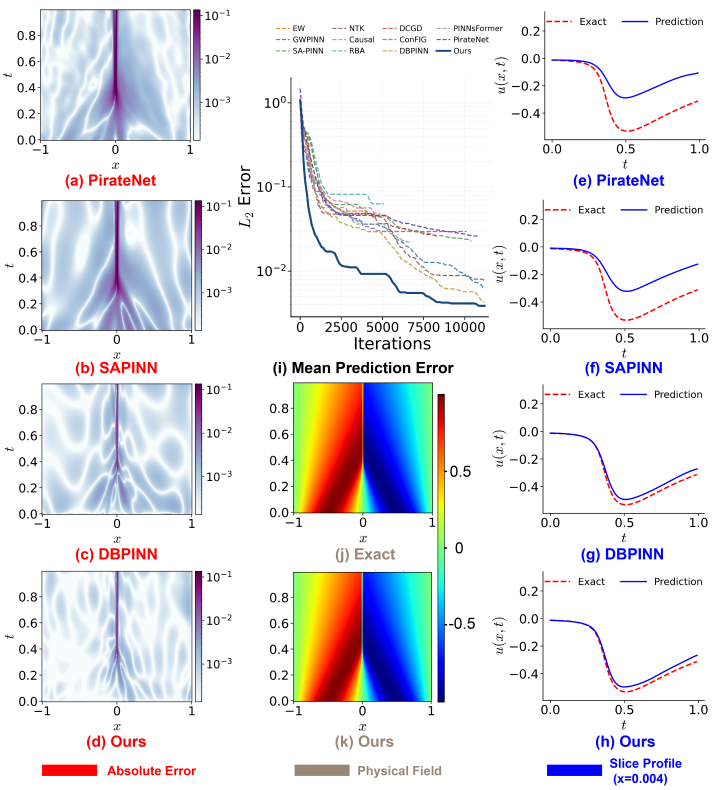

Figure 8: Results on the inverse Burgers problem, with the same panel layout as Figure 3. CSF-PINN gives the smallest point-wise error, and the identified coefficients match the true values.

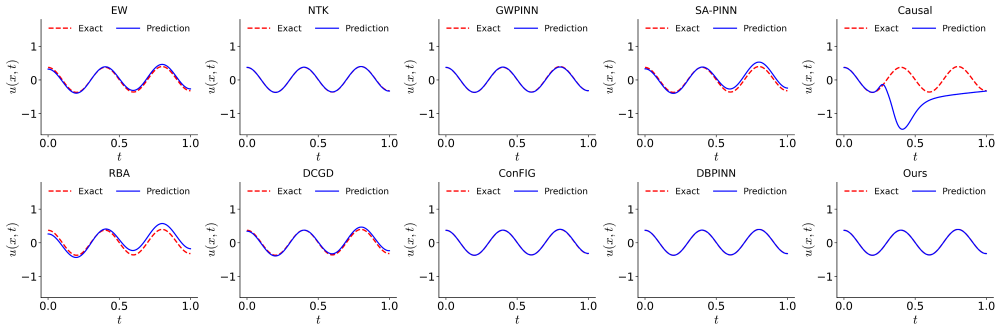

Figure 9: Klein–Gordon: slice profiles of the prediction (solid) against the reference (dashed) for the nine weighting baselines and CSF-PINN (Ours). Most baselines underestimate the late-domain growth and Causal diverges, whereas CSF-PINN overlaps the reference throughout.

scaled residual alternative, but does not separately identify the contributions of damping and the initial-representation skip connection.

Disentangling damping and the initial skip connection. The equivalent form in equation 6 combines hidden-state damping with an initial-representation skip connection. To examine their roles, we compare four propagation rules on Klein–Gordon and Allen–Cahn while keeping ICW and the remaining training configuration matched within each PDE. Each variant uses its own depth-shared learnable coefficient $\alpha = \sigma(\theta_\alpha)$, initialized at $1/3$.

Table 14 examines the contributions of hidden-state damping and the initial-representation skip under ICW weighting. Damping alone improves both metrics on Klein–Gordon; on Allen–Cahn, it re-

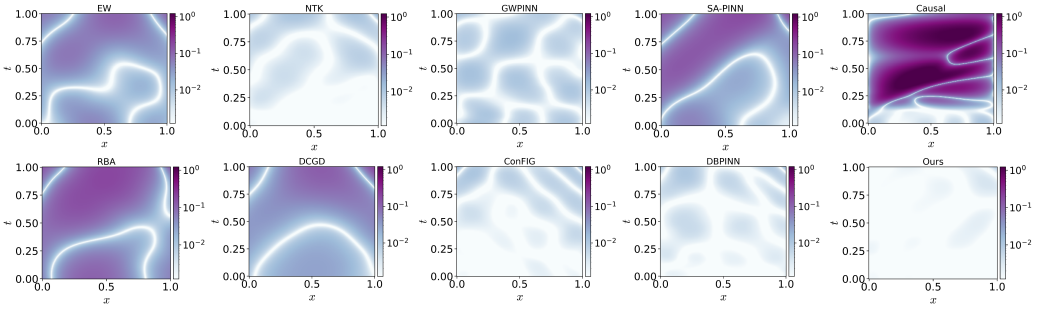

Figure 10: Klein–Gordon: point-wise absolute-error maps for the nine weighting baselines and CSF-PINN (Ours), on a shared log scale. CSF-PINN leaves the smallest error across the domain; Causal is the weakest.

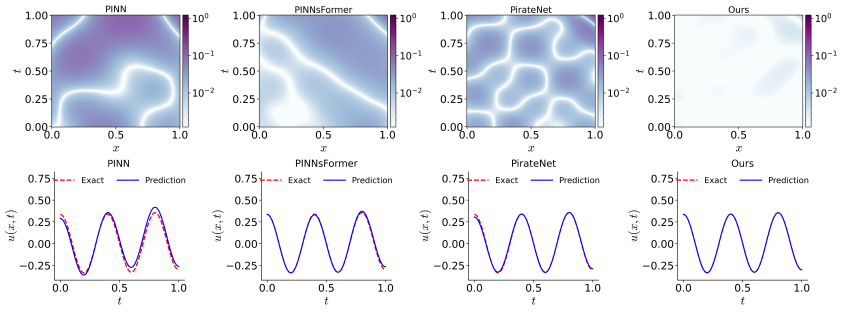

Figure 11: Klein–Gordon: comparison against the architecture baselines. Top row, point-wise absolute-error maps; bottom row, slice profiles. CSF-PINN attains the smallest error and the closest agreement with the reference.

duces MAE but increases relative L_2 error. The initial-skip-only variant improves both metrics over the scaled residual update on both PDEs, highlighting the value of retaining the initial representation. Full LIRB achieves the lowest errors across all four comparisons. Relative to initial-skip-only, it further reduces L_2 error and MAE by 9.15% and 8.35% on Klein–Gordon, and by 0.48% and 6.02% on Allen–Cahn. These results support the full propagation design, with the additional benefit varying across PDEs and error metrics.

E.4 Loss-Term Contribution Analysis

To explain *why* restoring cross-step feedback improves accuracy, we examine how the adaptive weighting redistributes each loss term's share of the training objective. For every loss term i (residual $\mathcal{L}_r$, boundary $\mathcal{L}_{bc}$, initial $\mathcal{L}_{ic}$, and, on the Burgers inverse problem, data $\mathcal{L}_d$) we record its value $\mathcal{L}_i$ and its adaptive weight λ_i at convergence, and define its contribution as the normalized weighted loss

$$\text{contribution}_i \;=\; \frac{\lambda_i \mathcal{L}_i}{\sum_j \lambda_j \mathcal{L}_j}, \tag{75}$$

i.e. the share that term i actually occupies in the total weighted objective. Figure 29 compares these shares under equal weighting (w/o, all $\lambda_i = 1$) and under our adaptive scheme (w/).

Under equal weighting the objective is typically dominated by a single term while the remaining constraints are suppressed, and the dominant term differs across PDEs. On Wave, the residual contributes only 18% while the boundary term occupies 81%; our scheme rebalances the shares to 54% residual, 25% boundary, and 21% initial. On Allen–Cahn, the boundary contribution is suppressed to nearly zero under equal weighting, with the objective split between residual (48%) and initial (52%); our scheme raises the long-neglected boundary term to 76%. On Burgers, equal weighting splits the objective roughly evenly between the residual (46%) and the data term (47%), whereas our

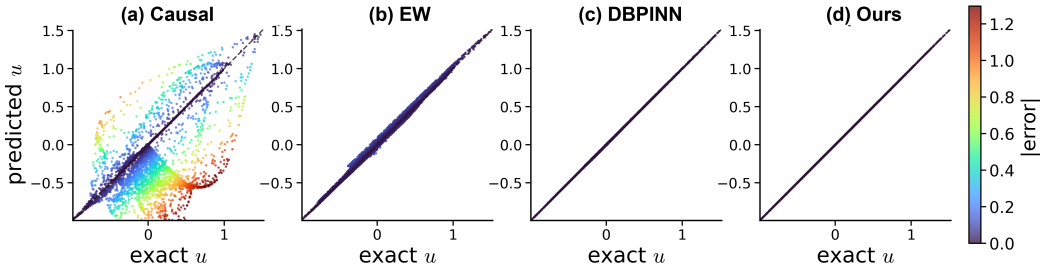

Figure 12: Klein–Gordon: predicted versus reference values, ordered from worst to best by relative L_2 error and colored by point-wise absolute error. The point cloud concentrates onto the diagonal $y = x$ from Causal to CSF-PINN, with CSF-PINN the tightest.

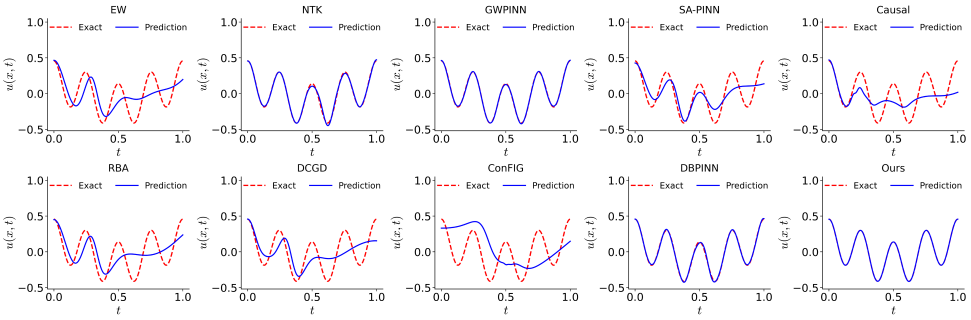

Figure 13: Wave: slice profiles for the nine weighting baselines and CSF-PINN (Ours). The baselines drift in phase and amplitude over the later part of the interval, whereas CSF-PINN remains aligned with the reference.

scheme concentrates it on the residual (78%), focusing training on the governing equation. Across all five benchmarks, the adaptive weights lift the contribution of constraints that equal weighting leaves under-enforced, which is the mechanism behind the accuracy gains, and is consistent with the residual-gradient rebalancing and the adaptive-weight trajectories reported above.

F Composability and Auxiliary Structure Studies

The main experiments and component ablations evaluate the proposed cross-step feedback mechanism. Here we examine its composability with weak formulations and equation-level structure.

F.1 Weak-Form Extension for a Nonsmooth Solution

We consider the inviscid Burgers Riemann problem

$$\begin{aligned} u_t + \partial_x \left(\frac{u^2}{2} \right) &= 0, \\ u(x, 0) &= \mathbf{1}_{\{x<0\}}, \\ (x, t) &\in [-1, 1] \times [0, 1]. \end{aligned} \tag{76}$$

The reference entropy solution contains a discontinuity on the Rankine–Hugoniot trajectory $x_s(t) = t/2$. Because pointwise derivatives are not appropriate at this discontinuity, the interior objective uses compactly supported test functions $\{\phi_q\}_{q=1}^{Q}$ and conservation moments

$$M_q(\theta) = \int_\Omega \left[u_\theta \, \partial_t \phi_q + \frac{u_\theta^2}{2} \, \partial_x \phi_q \right] \mathrm{d}x \, \mathrm{d}t. \tag{77}$$

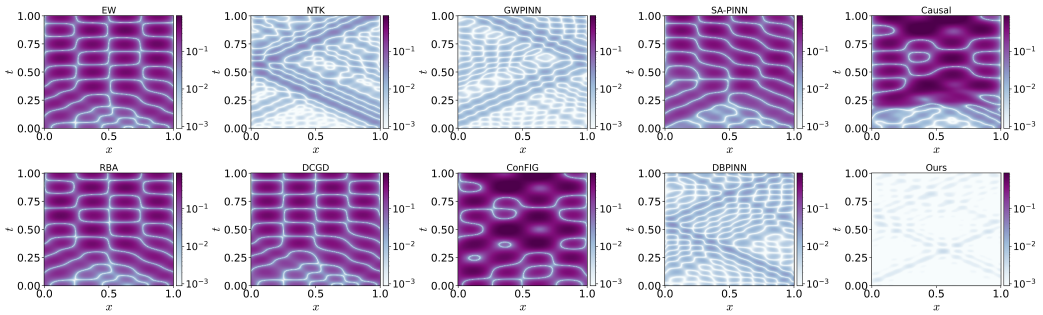

Figure 14: Wave: point-wise absolute-error maps for the nine weighting baselines and CSF-PINN (Ours), on a shared log scale. CSF-PINN gives the cleanest field.

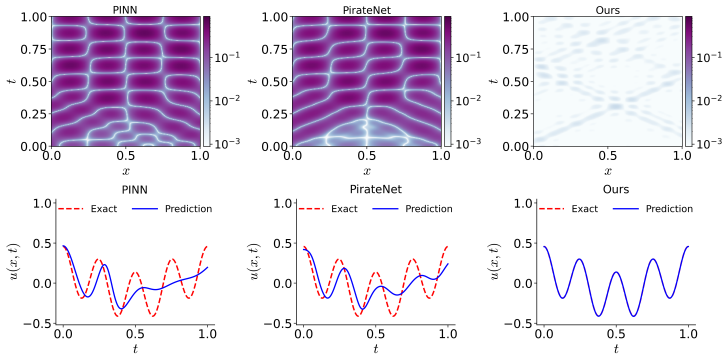

Figure 15: Wave: comparison against the architecture baselines (PINNsFormer runs out of memory and is omitted). Top row, point-wise absolute-error maps; bottom row, slice profiles. CSF-PINN tracks the reference where the baselines deviate.

The selected test functions vanish on the boundary of their support, while the initial and boundary data are enforced separately. All discrepancies use the differentiable L_1 surrogate

$$\rho_{1,\epsilon}(z) = \sqrt{z^2 + \epsilon^2} - \epsilon, \qquad \epsilon = 10^{-4}. \tag{78}$$

The weak residual is

$$\mathcal{L}_{\text{weak}} = \frac{1}{Q} \sum_{q=1}^{Q} \rho_{1,\epsilon}(M_q(\theta)). \tag{79}$$

The reference field is the entropy solution, while training uses a finite set of sampled weak conservation identities. We therefore evaluate this setting as a weak-form entropy-shock benchmark.

Matched surrounding formulation. For a controlled comparison, both the ordinary weak-form PINN and Ours use the same weak moments, collocation realization, condition losses, optimizer, physics-prior coefficients, and continuation budget. The shared structural terms are

$$\mathcal{L}_{\text{reflection}} = \mathbb{E}\, \rho_{1,\epsilon}(u_\theta(st - d, t) + u_\theta(st + d, t) - 1)\,, \tag{80}$$

$$\mathcal{L}_{\text{midpoint}} = \mathbb{E}\, \rho_{1,\epsilon}\big(u_\theta(st, t) - \tfrac{1}{2}\big)\,, \tag{81}$$

$$\mathcal{L}_{\text{decay}} = \mathbb{E}\left[\left(1 - e^{-|x - st|/\tau}\right) |\partial_x u_\theta(x, t)|\right], \tag{82}$$

with $s = 1/2$ and $\tau = 0.1$. They are derived directly from the flux and Riemann states and require no evaluation-grid labels. Both models use

$$0.15\mathcal{L}_{\text{reflection}} + 0.05\mathcal{L}_{\text{midpoint}} + 0.01\mathcal{L}_{\text{decay}}. \tag{83}$$

Both methods use the same auxiliary terms, so the comparison isolates the effect of the LIRB–ICW design.

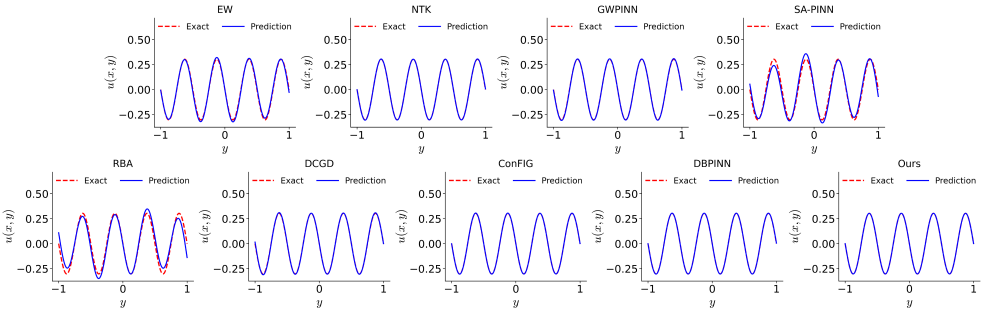

Figure 16: Helmholtz: slice profiles $u(x, y)$ for the eight weighting baselines and CSF-PINN (Ours). CSF-PINN matches the four-peak structure of the reference most closely.

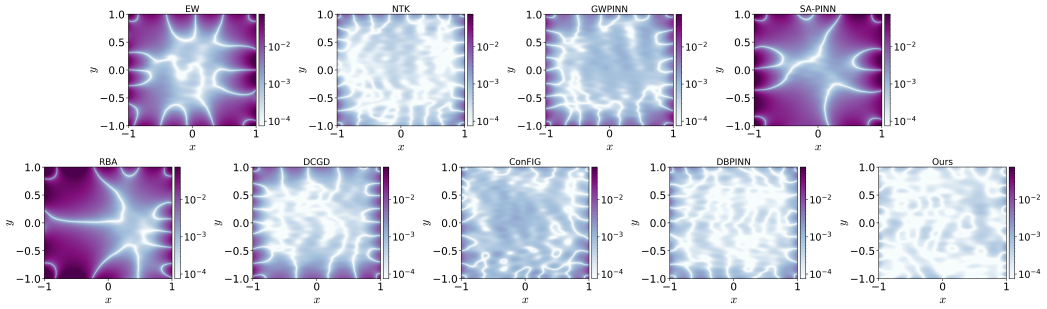

Figure 17: Helmholtz: point-wise absolute-error maps for the eight weighting baselines and CSF-PINN (Ours), on a shared log scale. CSF-PINN leaves the most uniform and lowest error; EW, SA-PINN, and RBA retain large error near the boundaries.

Both models are first optimized for 3000 weak-form epochs and 1500 prior-refinement epochs. A final 800-epoch stage applies the same log-linear output-temperature continuation

$$u_\theta(x,t) = \mathrm{sigmoid}(T(\eta) g_\theta(x,t)), \qquad T(\eta) = \exp(\eta \log 18), \quad 0 \le \eta \le 1. \tag{84}$$

The ordinary baseline is a four-layer MLP with 18,799 parameters; Ours has 18,833 parameters. Both use 4096 residual samples, 512 boundary samples, 1024 initial samples, and 1024 prior samples. The final stage uses Adam with initial learning rate 5×10^{-5} and the same cosine decay for both models. The ordinary PINN retains unit boundary and initial weights. Ours retains ICW on those two original channels. The ICW equations are unchanged, while this extension constrains the normalized condition weights to $[0.7, 1.3]$ for stable continuation. The final values are $\lambda_{\mathrm{bc}} = 0.7$ and $\lambda_{\mathrm{ic}} = 1.3$. LIRB itself is unchanged.

Table 15 shows reductions of 5.9% in global L_1, 5.9% in near-shock L_1, and 37.3% in shock-location MAE. These results establish that the original LIRB–ICW mechanism remains usable after replacing the pointwise residual by weak conservation moments. The larger strong-to-weak gain mainly reflects the change in the surrounding formulation.

F.2 Auxiliary Shock-Distance Weighting

For the same conservation law, the Rankine–Hugoniot speed is

$$s = \frac{f(u_L) - f(u_R)}{u_L - u_R} = \frac{1}{2}. \tag{85}$$

This gives the physics-derived distance and positive residual weight

$$\begin{aligned} d(x,t) &= |x - st|, \\ w(x,t) &= w_{\min} + \exp[-d(x,t)/\tau], \\ w_{\min} &= 0.1, \qquad \tau = 0.12. \end{aligned} \tag{86}$$

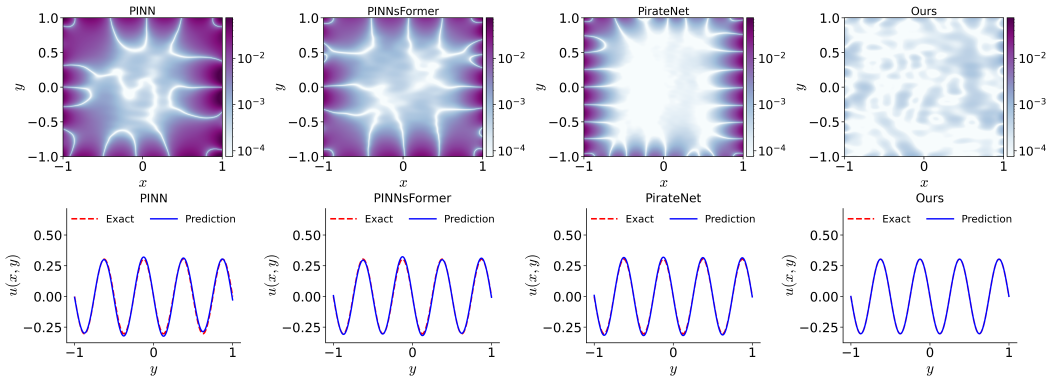

Figure 18: Helmholtz: comparison against the architecture baselines. Top row, point-wise absolute-error maps; bottom row, slice profiles. CSF-PINN gives the smallest point-wise error and the prediction closest to the reference.

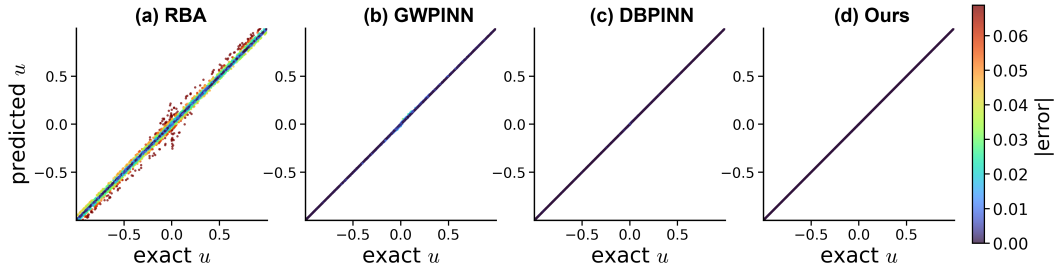

Figure 19: Helmholtz: predicted versus reference values, ordered from worst to best by relative L_2 error and colored by point-wise absolute error. CSF-PINN is the tightest along $y = x$.

The strong-form diagnostic multiplies the smooth-L_1 residual by $w/\overline{w}$. The positive floor prevents the far field from being discarded, and empirical normalization keeps the average residual scale comparable.

The spatial weight reduces global L_1 by 78.8%, away-shock L_1 by 92.6%, and shock-location MAE by 78.4%. This experiment quantifies the benefit of an analytically available structural prior and illustrates its compatibility with CSF-PINN.

F.3 Compatibility with Equation Symmetry

The Helmholtz reference solution $u^\star(x, y) = \sin(\pi x)\sin(4\pi y)$ is odd in both coordinates. We apply the deterministic Reynolds projection

$$(\Pi_{\mathrm{oo}} u_\theta)(x, y) = \frac{1}{4}\big[u_\theta(x, y) - u_\theta(-x, y) - u_\theta(x, -y) + u_\theta(-x, -y)\big]. \tag{87}$$

It introduces no trainable parameters and leaves the original optimizer unchanged.

Proposition 12 (Error monotonicity of exact projection)**.** *Assume the evaluation measure is invariant under coordinate reflections and $u^\star$ belongs to the odd–odd subspace. Then*

$$\|\Pi_{\mathrm{oo}} u_\theta - u^\star\|_{L_2} \le \|u_\theta - u^\star\|_{L_2}. \tag{88}$$

Proof. Under a reflection-invariant measure, Π_{oo} is the orthogonal projection onto the odd–odd subspace. Since $\Pi_{\mathrm{oo}} u^\star = u^\star$ and orthogonal projections are nonexpansive,

$$\|\Pi_{\mathrm{oo}} u_\theta - u^\star\|_{L_2} = \|\Pi_{\mathrm{oo}}(u_\theta - u^\star)\|_{L_2} \le \|u_\theta - u^\star\|_{L_2}. \tag{89}$$

□

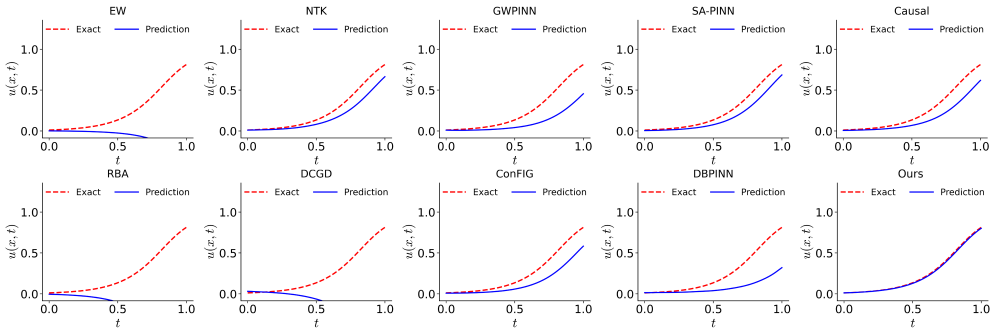

Figure 20: Allen–Cahn: slice profiles for the nine weighting baselines and CSF-PINN (Ours). EW, RBA, and DCGD collapse toward negative values; the remaining baselines underestimate the late-time growth; CSF-PINN tracks the reference.

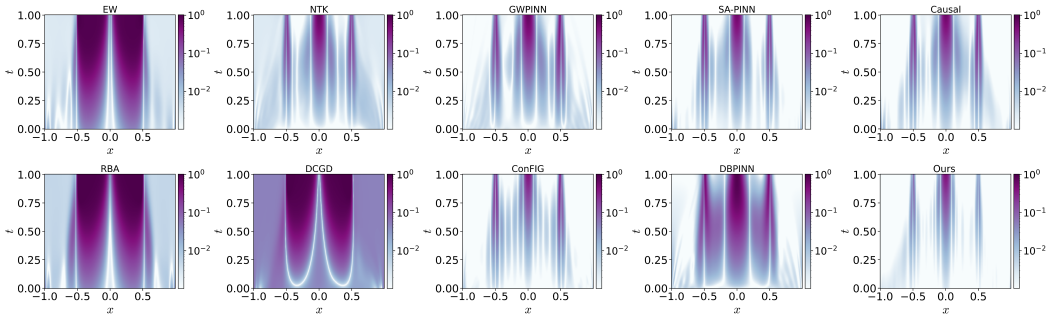

Figure 21: Allen–Cahn: point-wise absolute-error maps for the nine weighting baselines and CSF-PINN (Ours), on a shared log scale. CSF-PINN gives the cleanest field; EW, RBA, and DCGD retain broad high-error regions.

The paired projection lowers relative L_2 by 26.1% and reduces both symmetry defects by more than four orders of magnitude. The value 0.05013% is the unrounded error of the checkpoint used in this paired projection study. The 0.0506% value in the main table is the reported aggregate mean, so the two numbers have different reporting roles. Proposition 12 explains why a valid orthogonal projection cannot increase L_2 error when the target has the prescribed symmetry. The experiment therefore evaluates compatibility with a prescribed symmetry constraint.

F.4 Spatial Localization Diagnostics

The next two figures analyze the broader strong-to-weak formulation sweep before the final matched temperature-continuation stage. These figures provide spatial error diagnostics for the formulation sweep. Let $d(x,t) = |x - t/2|$ and partition the domain as

$$\begin{aligned} \Omega_{\text{near}} &= \{d < 0.05\}, \\ \Omega_{\text{trans}} &= \{0.05 \le d < 0.15\}, \\ \Omega_{\text{far}} &= \{d \ge 0.15\}. \end{aligned} \tag{90}$$

For the strong formulation, the near-shock and transition regions contribute 49% and 48% of the total error. After weak-form refinement, 96% of a much smaller total error lies near the shock. The larger near-shock fraction arises because the broad nonlocal error has been substantially reduced.

Figure 36 should be read together with the integrated global L_1 values. Beyond $d \approx 0.2$, the strong-form prediction is nearly exactly constant, so its conditional far-field error can be smaller. Its broad error around the incorrectly transported shock nevertheless dominates the domain integral. The figure diagnoses the spatial distribution of error, while the integrated L_1 metric determines the global ranking.

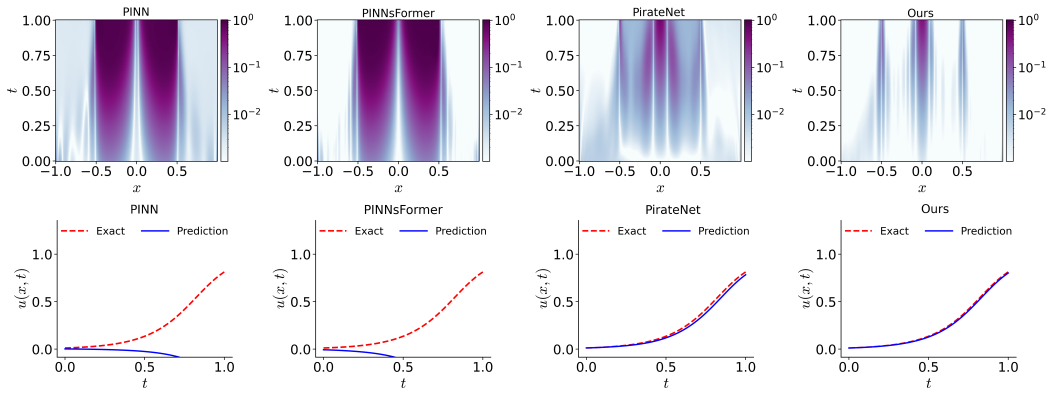

Figure 22: Allen–Cahn: comparison against the architecture baselines. Top row, point-wise absolute-error maps; bottom row, slice profiles. PINN and PINNsFormer fail to capture the solution, PirateNet is close, and CSF-PINN is best.

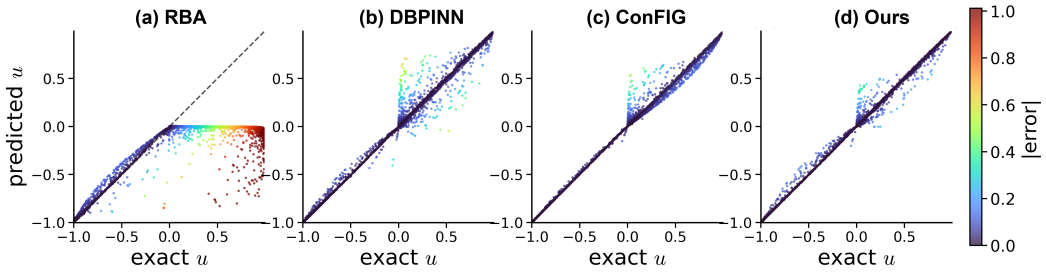

Figure 23: Allen–Cahn: predicted versus reference values, ordered from worst to best by relative L_2 error and colored by point-wise absolute error. RBA collapses to a horizontal band for $u > 0$; CSF-PINN is the tightest along $y = x$.

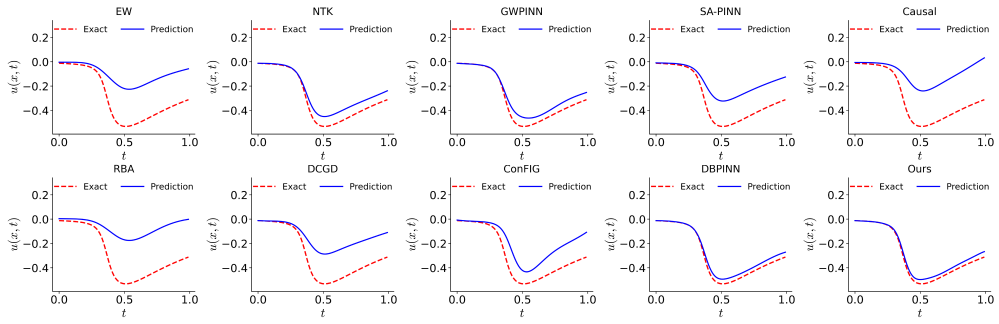

Figure 24: Burgers (inverse): slice profiles for the nine weighting baselines and CSF-PINN (Ours). CSF-PINN matches the depth of the trough near the shock most closely.

Table 13: Comparison of propagation rules under the same ICW weighting. The standard residual update is $z_{k+1} = z_k + t_k$, the state-reset update is $z_{k+1} = z_k + \alpha t_k$, and full LIRB follows equation 6. Relative L_2 error (%) and MAE (%).

Architecture	Helmholtz		Wave	
	L_2	MAE	L_2	MAE
Residual block + ICW	0.1039	0.0287	3.8492	0.8953
State-reset + ICW	0.1385	0.0384	0.8542	0.2325
LIRB + ICW	**0.0506**	**0.0184**	**0.1725**	**0.0757**

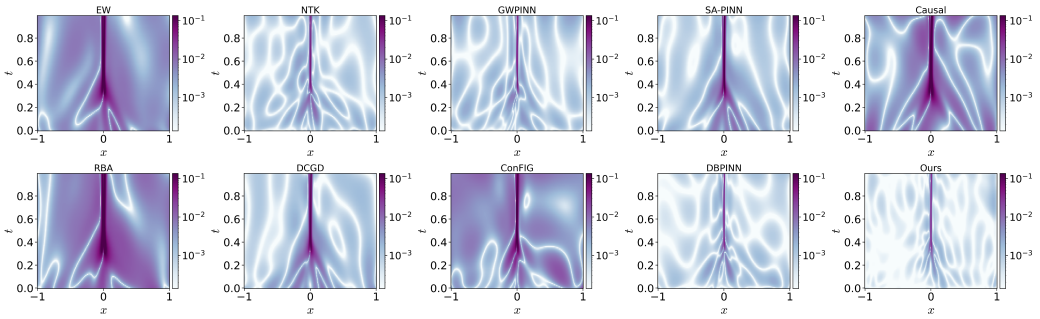

Figure 25: Burgers (inverse): point-wise absolute-error maps for the nine weighting baselines and CSF-PINN (Ours), on a shared log scale. Error concentrates along the shock at $x = 0$; CSF-PINN leaves the cleanest background.

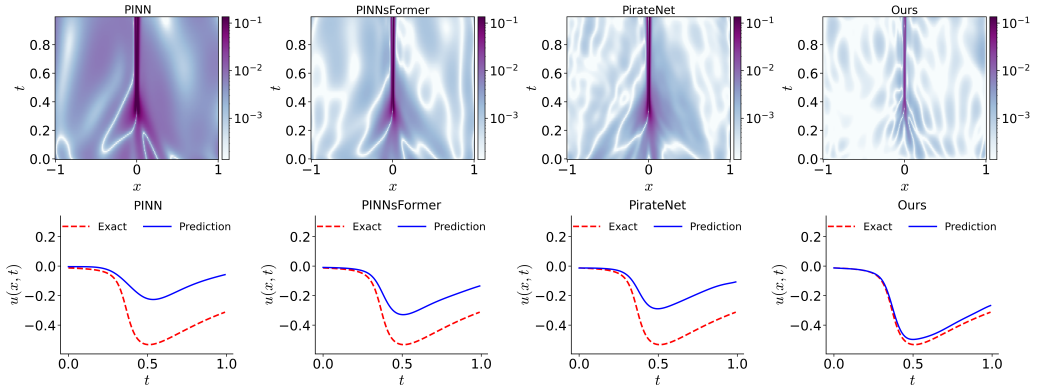

Figure 26: Burgers (inverse): comparison against the architecture baselines. Top row, point-wise absolute-error maps; bottom row, slice profiles. CSF-PINN gives the smallest error away from the shock and the closest slice.

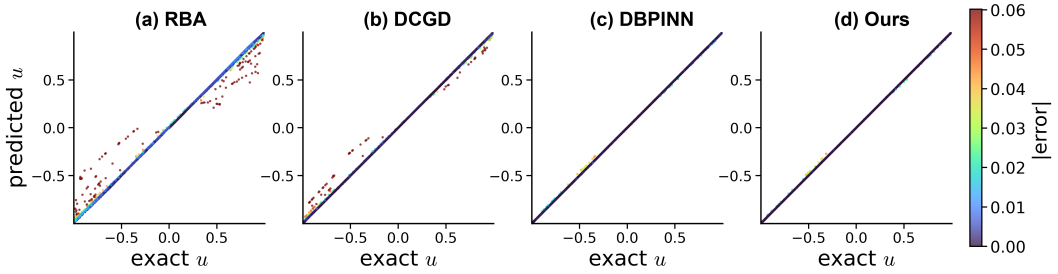

Figure 27: Burgers (inverse): predicted versus reference values, ordered from worst to best by relative L_2 error and colored by point-wise absolute error (note the small color range). RBA shows scattered outliers; CSF-PINN is the cleanest.

Table 14: Ablation of the LIRB propagation rule on Klein–Gordon and Allen–Cahn, with $t_k = T_k(z_k)$. Each variant uses a depth-shared learnable α, initialized at $1/3$ and optimized separately. Transform architectures, ICW settings, sampling, and training budgets are matched within each PDE. Relative L_2 error and MAE are reported in percent. Best results are in **bold**.

Propagation rule	Update: z_{k+1}	Klein–Gordon		Allen–Cahn	
		L_2	MAE	L_2	MAE
Scaled residual	$z_k + \alpha t_k$	0.4228	0.1849	4.1856	1.1948
Damping only	$\alpha z_k + \alpha t_k$	0.4012	0.1532	4.5728	1.0377
Initial skip only	$z_k + (1-\alpha) z_0 + \alpha t_k$	0.2993	0.0994	3.9294	0.7312
Full LIRB	$\alpha z_k + (1-\alpha) z_0 + \alpha t_k$	**0.2719**	**0.0911**	**3.9106**	**0.6872**

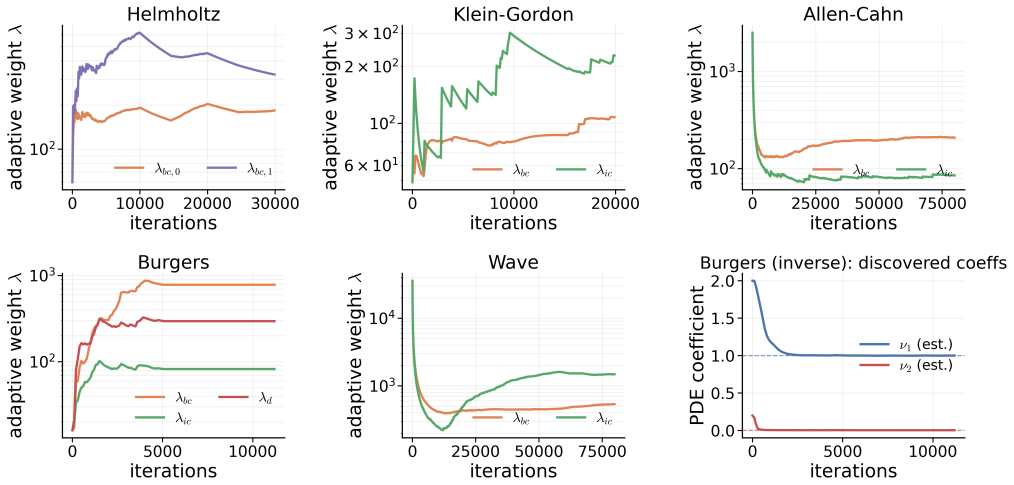

Figure 28: Adaptive loss weights versus iteration, one panel per PDE. Helmholtz shows its two boundary weights $\lambda_{bc,0}, \lambda_{bc,1}$; the other panels show λ_{bc} and λ_{ic} (Burgers additionally shows λ_d). The bottom-right panel shows the discovered coefficients on the inverse Burgers problem converging to the true values (dashed lines).

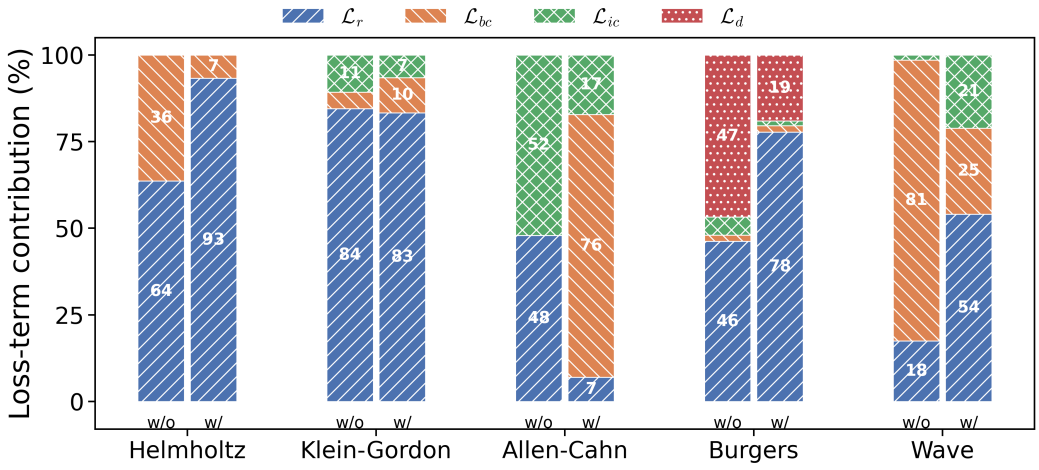

Figure 29: Loss-term contribution to the training objective, computed as the normalized weighted loss $\lambda_i \mathcal{L}_i / \sum_j \lambda_j \mathcal{L}_j$ at convergence, for equal weighting (w/o) versus our adaptive scheme (w/).

Table 15: Fully matched weak-form comparison on the nonsmooth Burgers problem. All three reported quantities are absolute-error averages, so the evaluation is L_1-based. Both models use the same weak moments, collocation points, physics priors, optimizer, and continuation protocol.

Method	Global $L_1 \downarrow$	Near-shock $L_1 \downarrow$	Shock MAE $\downarrow$
Weak-form PINN	6.640×10^{-4}	1.333×10^{-2}	2.311×10^{-4}
Ours	$\mathbf{6.245\times10^{-4}}$	$\mathbf{1.254\times10^{-2}}$	$\mathbf{1.448\times10^{-4}}$

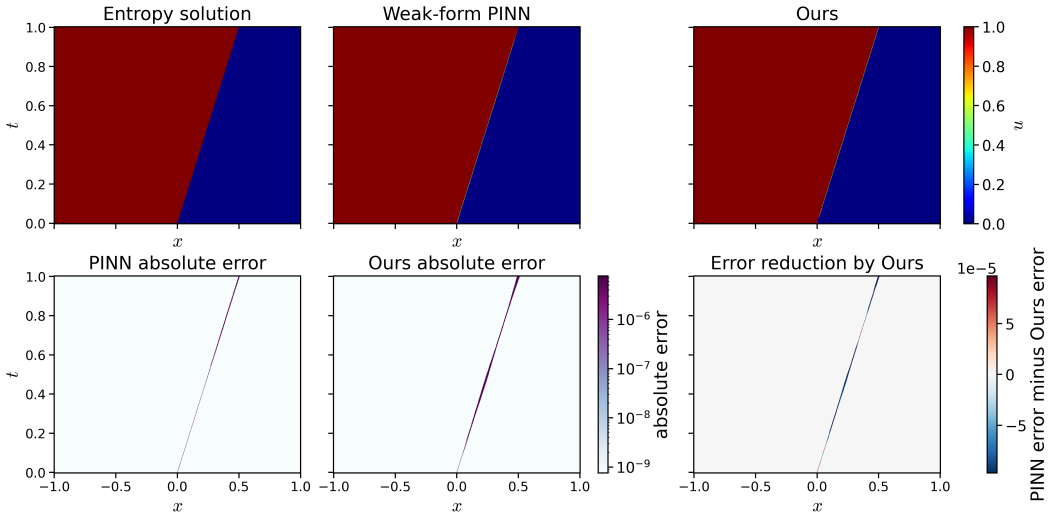

Figure 30: Matched weak-form fields and errors. The top row shows the entropy solution, the ordinary weak-form PINN, and Ours. The lower row compares absolute errors on a shared logarithmic scale and visualizes the pointwise error difference. The remaining error is concentrated around the moving discontinuity.

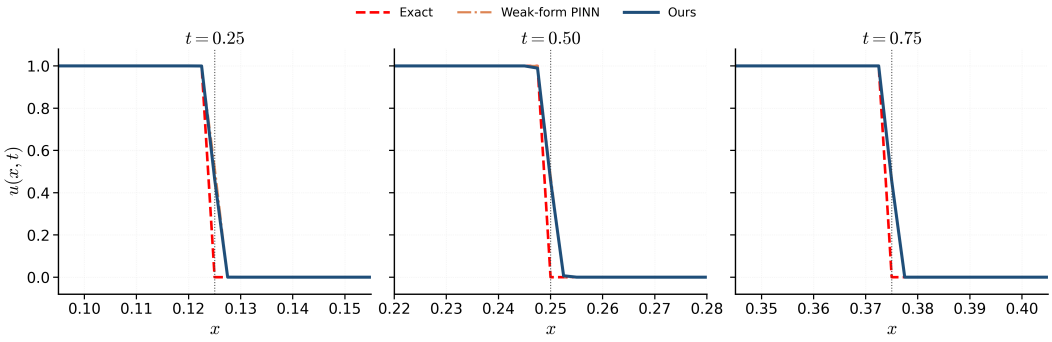

Figure 31: Shock profiles at $t \in \{0.25, 0.50, 0.75\}$ under the fully matched weak-form protocol. Both predictions follow the Rankine–Hugoniot trajectory. Ours has the smaller transition displacement and shock-location MAE.

Table 16: Matched on/off analysis of the physics-derived shock-distance weight. The network, initialization, collocation realization, optimizer, and 8000 epochs are fixed. Only the residual weight changes.

Variant	Global L_1	Near-shock L_1	Away-shock L_1	Shock MAE
Without distance prior	1.096×10^{-1}	4.695×10^{-1}	9.072×10^{-2}	2.317×10^{-1}
With distance prior	$\mathbf{2.321\times10^{-2}}$	$\mathbf{3.384\times10^{-1}}$	$\mathbf{6.694\times10^{-3}}$	$\mathbf{4.995\times10^{-2}}$

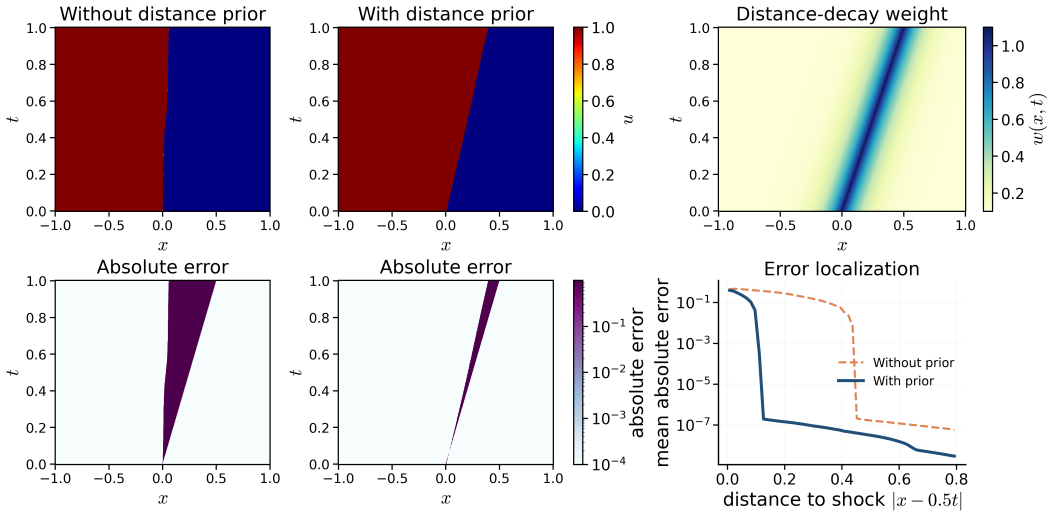

Figure 32: Effect of the shock-distance weight. The first two columns compare fields and absolute errors without and with the prior. The upper-right panel shows the prescribed weight, and the lower-right panel reports conditional error versus distance to the shock.

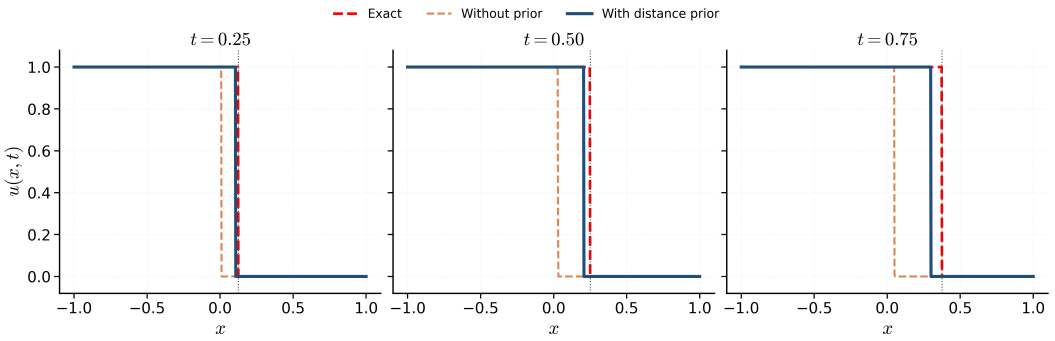

Figure 33: Shock slices for the matched distance-prior analysis. Weighting the residual near the analytically known trajectory substantially improves shock transport.

Table 17: Odd–odd Helmholtz compatibility under the original main-table configuration. Relative L_2 is appropriate because the solution is smooth.

Variant	Rel. L_2 (%) ↓	Odd-x defect	Odd-y defect
Ours	0.05013	2.840×10^{-3}	3.868×10^{-3}
Ours + odd–odd projection	**0.03703**	$\mathbf{1.192\times10^{-7}}$	$\mathbf{1.788\times10^{-7}}$

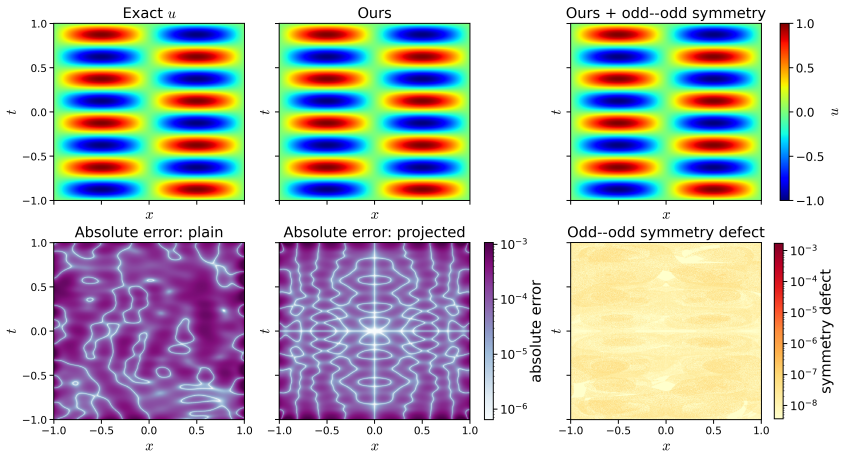

Figure 34: Odd–odd symmetry compatibility. The top row compares the exact field, Ours, and the projected output. The lower row shows absolute errors and the symmetry defect. Projection enforces the prescribed symmetry to approximately floating-point precision.

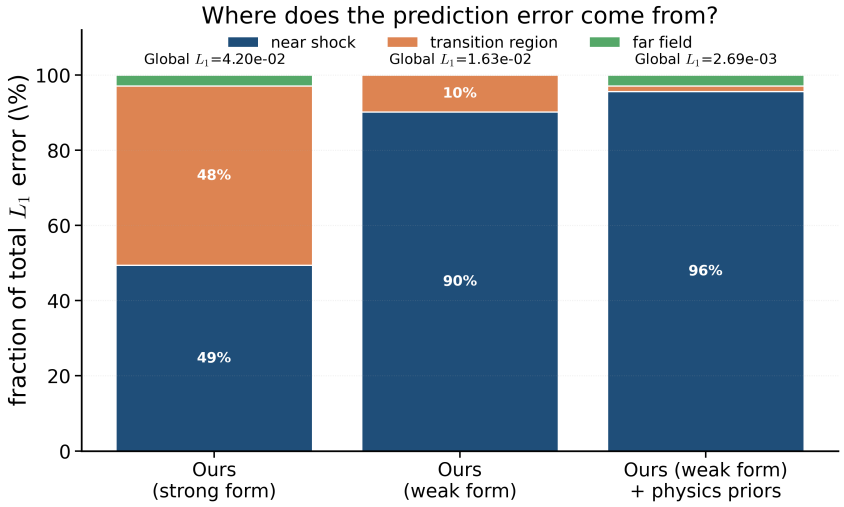

Figure 35: Fraction of total L_1 error contributed by the shock neighborhood, transition region, and far field during the formulation sweep. The global L_1 value falls from 4.20×10^{-2} to 2.69×10^{-3}, while the remaining error becomes increasingly localized near the discontinuity.

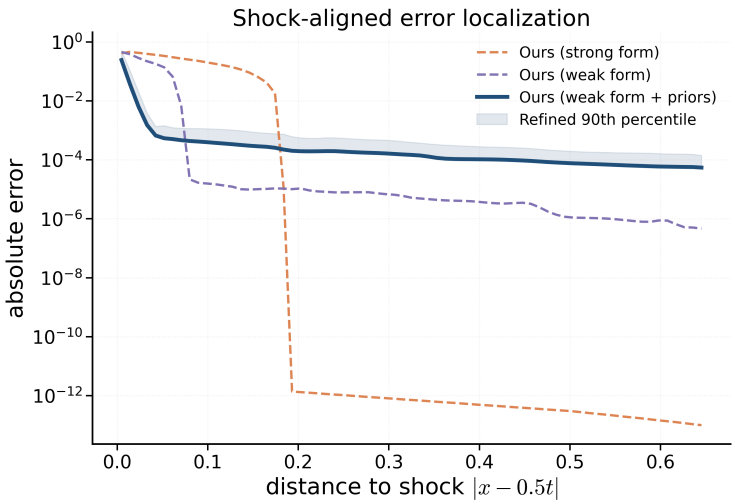

Figure 36: Mean absolute error conditioned on distance to the exact shock. The strong-form solution can attain machine-level error in the constant far field yet retains a broad high-error interval around a displaced shock. The refined weak solution substantially reduces the shock-neighborhood error. The shaded band reports its 90th-percentile pointwise error.