PDE-LENS: SCALING END-TO-END PDE DISCOVERY THROUGH SPECTRAL REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

End-to-end PDE discovery maps an observed solution field to its governing equation. We study the end-to-end models that learn this map and find a recurring failure: they capture the slowly varying background of the field but miss the sharp fronts, narrow peaks, and oscillations that separate one candidate equation from another. We call this *high-frequency neglect* and locate the failure in the field encoder. PDE-LENS addresses it with PRISM, an encoder that operates directly on scattered observations in the spectral domain, using non-equispaced Fourier analysis. To train at scale, we build a quality-controlled dataset of 1.19 million numerically solved PDE trajectories from a library of 59 atomic terms, with LLM-assisted solver generation handling stiff or ill-posed cases. On unseen four-term equations, PDE-LENS reaches a median NMSE more than sixfold lower than PDE-FIND, the strongest non-pretrained baseline. Across 99 unseen trajectories over ten equation structures, PRISM recovers 65 within NMSE ≤ 0.01, against 50 for NeSymReS and 34 for a ViT encoder. For conventional symbolic regression, our encoder also achieved high performance. On the Feynman benchmark it also raises conventional symbolic-regression accuracy and high-frequency fidelity over E2E-SR and NeSymReS. PRISM's 256-dimensional latent representations also predict four dynamical properties with the best R^2 on three of four. Equation discovery thus yields not only the governing equation but a reusable, structured PDE latent space.

1 INTRODUCTION

Partial differential equations (PDEs) model a vast range of systems across physics, life sciences, transportation, and engineering. While solving PDEs is a long-standing challenge, a complementary problem has recently drawn growing interest: discovering the governing PDE from observational data, with downstream applications in interpretation, forecasting, and anomaly detection. Existing approaches largely fall into two families. Sparse-regression methods (Rudy et al., 2017) fit a governing equation as a sparse linear combination of a hand-built term library, while reinforcement-learning methods (Du et al., 2024) search over a predefined grammar of expressions. Both struggle to scale or to discover equations beyond predefined libraries.

A more promising direction is to train a network that maps solution fields directly to governing equations, in the encoder–decoder spirit of end-to-end symbolic regression (SR) (Kamienny et al., 2022). In principle, one could sidestep the field entirely: compute the derivative terms of a PDE sample numerically and fit them with an off-the-shelf SR network—this is how MDBench (Ziaei Bideh et al., 2026) evaluates E2E-SR on PDEs. But such a shortcut recovers only a symbolic expression and learns no reusable representation of the dynamics. Our goal is stronger: we aim for a *PDE latent space* that encodes the dynamical structure of PDE data and transfers to tasks beyond recovering equations, such as forecasting, anomaly detection, and property prediction (Meidani et al., 2024).

Yet, to the best of our knowledge, no prior work has performed large-scale pretraining for PDE discovery. The bottleneck is data: a pretraining dataset requires millions of numerically solved trajectories with reliable governing equations, and generating high-quality numerical solutions for diverse, stiff, or ill-posed PDEs has long been impractical. With the rise of AI agents, tools such as CodePDE (Li et al., 2026b) can generate and validate numerical solvers automatically, making a large-scale, quality-controlled PDE dataset feasible. We build on this capability to undertake such pretraining.

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.

This effort exposes a generalization bottleneck in existing field encoders. Although end-to-end models can accurately recover equations from training examples containing sharp fronts, narrow peaks, and rapid oscillations, their recovery on unseen fields often captures the dominant variation while missing these equation-discriminative details, even when the fields are drawn from the same distribution.

This observation led us to re-examine conventional E2E SR, where the same failure has been largely overlooked: its errors concentrate on instances carrying substantial high-frequency content (e.g. rapidly varying terms such as $1/x$). Such instances are rare in standard benchmarks and contribute little to aggregate metrics, but in PDEs they are the norm rather than the exception. Therefore, our contributions are as follows:

- We pioneer large-scale pretraining for PDE discovery and build a quality-controlled dataset of numerically solved PDE trajectories with the aid of automated solver generation based on LLMs.
- We identify *high-frequency neglect*, a failure pattern that affects existing field encoders on PDEs and, more subtly, conventional end-to-end symbolic regression on high-frequency functions. To this end, we design the PRISM encoder, which operates directly on scattered observations in the spectral domain.
- We show that the PRISM encoder substantially improves the performance of PDE discovery and conventional symbolic regression, and that it learns a structured PDE latent space.

2 Related Work

Generative Symbolic Regression Symbolic regression has moved from discrete search toward generative modeling. Traditional genetic programming scales poorly with expression dimensionality. Recent methods add neural guidance—neural-seeded GP (Mundhenk et al., 2021) and transformer-guided Monte-Carlo tree search (Yu et al., 2025; Shojaee et al., 2023)—or cast construction as an MDP for reinforcement learning (Liang & Yang, 2025; Tian et al., 2024) and LLM heuristic search (Shojaee et al., 2025). Despite this, they did not overcome the intrinsic limitations of search-based methods. By contrast, generative approaches learn the distribution over expressions: some model it as a sequence-to-sequence translation problem (Kamienny et al., 2022), some treat it as a multimodal problem (Meidani et al., 2024; Li et al., 2025), and others use probabilistic modeling (Han et al., 2025; Li et al., 2026a; Xiang et al., 2026). Generative symbolic regression based on large-scale pretraining has been extended to ODEs (same architecture) (d'Ascoli et al., 2024) but not to PDEs yet.

Methods for PDE Equation Discovery Sparse-identification methods (SINDy (Brunton et al., 2016; Rudy et al., 2017) and its variants (Anvari et al., 2025; Messenger & Bortz, 2021)) fit a governing equation as a sparse combination of a hand-built candidate library, which means a term absent from the library is unrecoverable. DISCOVER (Du et al., 2024) and DeepMoD (Both et al., 2021) introduce reinforcement learning, but they still depend on predefined library and struggle to scale. EqGPT (Xu et al., 2025) learns a generative model of PDEs from mathematical handbooks, yet its pretraining is over symbolic text rather than numerical solution fields. PROSE-PDE (Sun et al., 2025) uses multimodal pretraining across many PDE operators, but its symbolic modality merely regularizes forward forecasting over dense 1D fields rather than recovering equations. To the best of our knowledge, no prior method performs large-scale pretraining for end-to-end discovery from data, largely because generating a high-quality, large-scale PDE dataset has been impractical until tools such as CodePDE (Li et al., 2026b) made it feasible.

Spectral Bias and Fourier Domain Learning Rahaman et al. (2019) showed that neural networks preferentially fit low-frequency components during training, while Fourier feature mappings alter this bias and improve the representation of high-frequency functions (Tancik et al., 2020). Periodic activation functions offer a complementary implicit-representation strategy (Sitzmann et al., 2020). In PDE learning, Fourier Neural Operators use truncated Fourier multipliers to parameterize nonlocal integral operators (Li et al., 2021). Other operator-learning architectures augment Fourier layers with geometry-adaptive spectral maps (Li et al., 2023) and wavelet bases (Tripura & Chakraborty, 2023). Together, these studies establish spectral and multiresolution parameteriza-

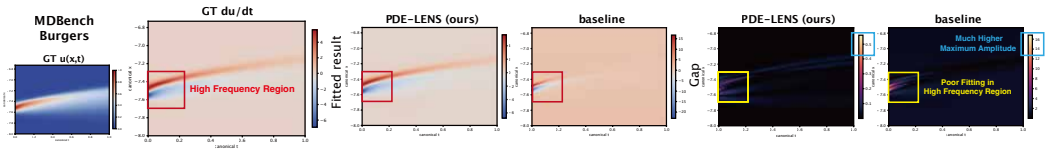


Figure 1: Comparison on MDBench Burgers: an example. From left to right: the ground-truth solution u, time derivative u_t, recovered right-hand sides for PDE-LENS and the baseline, and corresponding absolute errors $|\widehat{f} - u_t|$. Boxes highlight the sharp-front region, where the baseline incurs substantially larger errors despite capturing the broad propagation pattern.

tions as effective tools for learning operators, while also exposing the practical roles of the choice of representation for irregular PDE data.

3 The High-Frequency Neglect in End-to-End PDE Discovery

3.1 The End-to-End Large-Scale Pretraining Paradigm for PDE Discovery

We consider a scalar field $u(\mathbf{x}, t)$ over the spatiotemporal domain $\Omega \times [0, T]$, where $\mathbf{x} \in \Omega \subset \mathbb{R}^d$. Its dynamics are governed by an unknown PDE in residual form $\mathcal{G}\big(u, \partial_t u, \nabla_{\mathbf{x}} u, \nabla^2_{\mathbf{x}} u, \ldots, \mathbf{x}, t; \mathbf{c}\big) = 0$, where $\mathbf{c}$ denotes the numerical coefficients. Given samples of a PDE field $\mathcal{S} = \{(\mathbf{x}_i, t_i, u(\mathbf{x}_i, t_i))\}_{i=1}^{N}$, the goal of PDE discovery is to recover the governing PDE $\mathcal{G}^\star$ that best explains the observed solution field.

Following the encoder–decoder paradigm of end-to-end symbolic regression (E2ESR) (Kamienny et al., 2022), we formulate PDE discovery as conditional sequence generation. An encoder maps the PDE field $\mathcal{S}$ to a latent representation, and an autoregressive decoder generates the complete PDE equation, including both its symbolic structure and numerical coefficients. The encoder and decoder are trained jointly under a single objective. At inference time, the model first generates candidate equations and then separately refines their numerical coefficients (Appendix A).

We serialize each target PDE as a prefix token sequence $\mathbf{e} = (e_1, \ldots, e_L)$. Operators, field variables, and derivatives are represented as symbolic tokens, while each floating-point coefficient is encoded by sign, mantissa, and exponent tokens following general generative SR methods (Kamienny et al., 2022; Xiang et al., 2026). Let $\mathbf{z} = \mathrm{Enc}_\theta(\mathcal{S})$ denote the encoded solution field. The decoder factorizes the conditional probability as

$$p_\theta(\mathbf{e} \mid \mathcal{S}) = \prod_{\ell=1}^{L} p_\theta(e_\ell \mid e_{<\ell}, \mathbf{z}) \,. \tag{1}$$

We encode the solution field directly and train the encoder jointly with the symbolic decoder. This formulation makes equation identification the supervision for learning a field representation, whose dynamical content we subsequently examine through latent-space analysis and property prediction.

3.2 Global Fit Obscures High-Frequency Evidence

In equation discovery, the evidence distinguishing candidate equations lies not in the slowly varying background but in high-frequency structures such as sharp fronts, narrow peaks, or oscillations. Existing end-to-end models can recover these structures from training examples, yet on unseen fields their recovery captures the dominant variation while missing or misinterpreting these details. Unlike high-frequency approximation errors in function-mapping tasks (Lippe et al., 2023), this is a failure to generalize the recognition of high-frequency evidence, which we call *high-frequency neglect*.

This failure is particularly relevant to PDE discovery, where sharp transitions and oscillatory dynamics carry information about the governing equation As shown in Figure 1, the baseline's recovered right-hand side captures the dominant propagation pattern but incurs substantial errors near the sharp front. When these distinctions are lost in the field-to-latent mapping, the decoder cannot distinguish the governing equation from a smoother surrogate. We also observe this in conventional SR, which is long overlooked (Fig. 5).

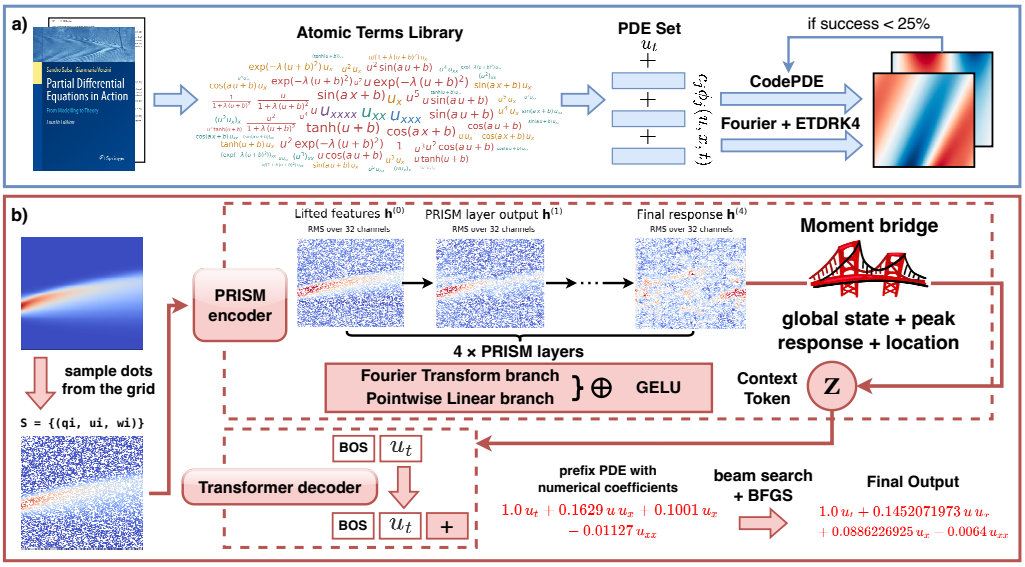


Figure 2: **PDE-LENS pipeline. (a)** Candidate PDEs are composed from a textbook-derived atomic-term library and solved under varied coefficients and initial conditions using Fourier discretization and ETDRK4, with CodePDE invoked when fewer than 25% of trajectories pass validation. **(b)** PRISM processes coordinate–value observations through spectral–pointwise layers and compresses them into a context token via a moment bridge. A Transformer decoder generates the equation, followed by beam search and BFGS coefficient refinement.

4 PDE-LENS

Our overall pipeline is shown in Figure 2. Section 4.1 will introduce our encoder which mitigates high-freqency neglect. Section 4.2 will elaborate on our data generation procedure. Remaining details are included in Section 4.3.

4.1 PRISM: The Encoder

Equation identification is a discrimination problem: the decoder needs the spectral evidence that separates one operator from another, including weak high-wavenumber and localized content (Section 3.2). PRISM's spectral layer borrows the idea of Fourier-domain learnable multipliers from FNO (Li et al., 2021) and its non-equispaced treatment from DSE-FNO (Lingsch et al., 2024), so it consumes scattered observations. PRISM combines mode-specific transformations with pointwise processing to construct a representation for equation identification. Each signed frequency mode k is transformed by a learned channel operator W_k, while a parallel pointwise branch Ah carries local feature responses through the layer. A weighted Moment Bridge then summarizes the resulting point states into a context token c. Joint training with the symbolic decoder optimizes the complete encoder for equation identification, allowing the mode-specific transformations to respond to the predictive value of spectral features rather than their energy alone.

Let $\mathcal{O} = \{(z_i, u_i, w_i)\}_{i=1}^{N}$ be scattered space–time observations with normalized coordinates $z_i = (x_i, t_i) \in [0,1]^2$ and quadrature weights w_i. Each point is lifted after a weighted z-score: $h_i^{(0)} = \rho([z_i, \widetilde{u}_i]), \widetilde{u}_i = \frac{u_i - \sum_j w_j u_j}{\left(\sum_j w_j |u_j - \sum_\ell w_\ell u_\ell|^2\right)^{1/2} + \epsilon}$. Layer ℓ computes a non-equispaced Fourier analysis over a truncated signed frequency lattice $\mathcal{K}$, $\widehat{h}^{(\ell)}(k) = \sum_{i=1}^{N} w_i h_i^{(\ell)} e^{-2\pi \mathrm{i}\langle k, z_i\rangle}$, with $k \in \mathcal{K}$, then applies a learned mode-conditioned channel operator $W_k^{(\ell)} \in \mathbb{C}^{C\times C}$, synthesizes at the observation coordinates, and adds a parallel pointwise branch:

$$h_i^{(\ell+1)} = \phi\left(\sum_{k\in\mathcal{K}} e^{2\pi \mathrm{i}\langle k, z_i\rangle} W_k^{(\ell)} \widehat{h}^{(\ell)}(k) + A^{(\ell)} h_i^{(\ell)}\right). \tag{2}$$

PRISM uses four layers, 32 hidden channels, and a 32×32 signed-mode lattice with $k_x, k_t \in \{-16, \ldots, 15\}$. The final point states pass through a weighted Moment Bridge that aggregates the

weighted mean, RMS, soft maximum, and coordinate moments into one context token:

$$c = \psi[\mu_w(h), \mathrm{rms}_w(h), \mathrm{softmax}_\tau(h), \mu_w(xh), \mu_w(th), \mu_w(z)]\,. \tag{3}$$

Together, these statistics provide a compact description of feature strength and location. The autoregressive decoder conditions on the resulting context token c to generate the equation tokens. An additional consistency loss is optionally available, which aligns representations obtained from multiple independent sample clouds of the same function.

4.2 Large-Scale PDE Data Generation

End-to-end PDE discovery requires large-scale paired data consisting of governing equations and their PDE fields. Inspired by EqGPT (Xu et al., 2025), we collect common PDE structures from mathematical textbooks and organize them into a library $\mathcal{T}$ of 59 atomic terms spanning reaction, transport, diffusion, and dispersion operators. Candidate equations are then constructed through additive composition:

$$\mathcal{G}(u) = u_t + \sum_{j=1}^{L} c_j \phi_j(u, x, t) = 0, \qquad \phi_j \in \mathcal{T}. \tag{4}$$

Given the computational budget, and to prevent the combinatorial growth in candidates with L from biasing the training distribution toward longer expressions, we process 10,656 equations with $L \leq 3$ and additionally sample 5,000 equations with $L = 4$. For each equation, we generate ten sets of internal parameters or numerical coefficients. Each parameter configuration is paired with ten initial conditions under periodic boundary conditions, resulting in 100 candidate trajectories per equation.

All candidate equations are solved over $x \in [-2.5, 2.5)$ and $t \in [0, 5)$, with each solution stored as a 500×500 spatiotemporal field. We use a Fourier pseudospectral discretization in space and ETDRK4 integration in time. The internal integration step may be smaller than the saving interval $\Delta t_{\mathrm{save}} = 0.01$. Numerical accuracy is estimated using a time-step-doubling test:

$$\varepsilon_{\mathrm{num}} = \frac{\left\| u_{\Delta t_{\mathrm{int}}} - u_{\Delta t_{\mathrm{int}}/2} \right\|_2}{\left\| u_{\Delta t_{\mathrm{int}}/2} \right\|_2 + 10^{-12}}. \tag{5}$$

A trajectory is included in the training set only if $\varepsilon_{\mathrm{num}} \leq 10^{-3}$, all solution values are finite, the maximum absolute amplitude does not exceed 10, and the spectral-tail energy ratio does not exceed 10^{-5}. For equations containing third- or fourth-order spatial derivatives, we additionally constrain the finite-horizon linear spectral amplification, the dispersive phase at the saved temporal resolution, and the dissipative sign of the fourth-order term.

If fewer than 25% of the trajectories generated by the conventional solver are valid, we invoke CodePDE (Li et al., 2026b) to generate or refine a numerical solver for that equation and evaluate its outputs using the same accuracy and stability criteria. If the refined solver still fails to produce enough valid trajectories, the equation is removed from the training set. We use DeepSeek-v4-flash for our CodePDE in the experiment. Overall, the pipeline retains 1,188,134 quality-controlled trajectories from 1,565,600 candidates.

4.3 Other Pipeline Details

Grid scaling. To handle fields sampled on different grid spacings without enlarging the training set, we introduce grid scaling to express all equations in a common coordinate system. We rewrite each PDE from its sampled source grid $(\Delta x_{\mathrm{src}}, \Delta t_{\mathrm{src}})$ to a fixed target grid $(\Delta x_{\mathrm{tgt}}, \Delta t_{\mathrm{tgt}})$. The field values themselves are unchanged. Defining $s_x = \frac{\Delta x_{\mathrm{tgt}}}{\Delta x_{\mathrm{src}}}$, $s_t = \frac{\Delta t_{\mathrm{tgt}}}{\Delta t_{\mathrm{src}}}$, a derivative term transforms as $\partial_x^p \partial_t^q u \mapsto s_x^p s_t^q \, \partial_x^p \partial_t^q u$. After renormalizing the PDE so that the coefficient of u_t remains one, a non-leading coefficient c becomes $c \mapsto c\, s_x^p s_t^{q-1}$. For example, the coefficients of u_x and u_{xx} are multiplied by s_x/s_t and s_x^2/s_t, respectively. The derivative target is transformed consistently as $(u_t)_{\mathrm{tgt}} = \frac{\Delta t_{\mathrm{src}}}{\Delta t_{\mathrm{tgt}}}(u_t)_{\mathrm{src}}$. This places token encoding, coefficient refinement and equation scoring in the same grid coordinate system.

Normalization. Apart from spacings, field amplitudes also differ wildly across equations, so each observed field is standardized per sample before being fed to the encoder. We replace u by $z =$

Table 1: NMSE on five 1D MDBench systems. The upper block contains pretrained neural models and the lower block contains non-pretrained online discovery methods.

Method	Advection	Burgers	Heat	KdV	KS	Mean NMSE ↓	Runtime (s/sample) ↓
ViT	5.00e−5	0.168	–	5.89e−3	0.985	0.3303	3.99
NeSymReS	7.19e−5	1.76e−3	1.0000	1.0000	0.9995	0.6003	6.34
PDE-LENS (ours)	8.28e−5	8.21e−4	3.85e−5	3.31e−3	0.982	0.1973	7.86
PDE-FIND	1.30e−6	7.17e−2	1.005	4.62e−4	0.797	0.3747	23.35
Bayesian	1.30e−6	6.95e−2	0.964	6.86e−4	0.797	0.3663	49.85
WSINDy	3.55e−3	1.85e−4	4.16e5	0.151	1.410	8.31e4	21.80
DeepMoD	3.50e17	8.69	–	6.48	3.08e3	8.74e16	6813.30

Table 2: Results on ten randomly sampled training-set-excluded $L = 4$ trajectories.

Method	Mean NMSE ↓	Runtime (s/sample) ↓
PDE-FIND	0.1045	35.25
Bayesian	0.1062	58.95
WSINDy	0.1359	106.64
DeepMoD	0.1443	1862.9
PDE-LENS (ours)	**0.0021**	**4.27**

$(u - \mu)/\sigma$, and the governing equation is transformed consistently by substituting $u = \mu + \sigma z$, under which each derivative rescales as $u_t = \sigma z_t$ (the affine offset μ vanishes). The PDE is then renormalized so that the coefficient of z_t remains one, matching the convention of grid scaling. The inverse transform is applied at evaluation to recover equations in physical units.

Decoder. Following Kamienny et al. (2022), we use a transformer-based decoder, two times heavier than the encoder.

Constant Optimization. Decoded coefficients are refined by solving the PDE inverse problem with BFGS; details are in Appendix A.

5 Experiments and Analysis

We follow MDBench (Ziaei Bideh et al., 2026) to report NMSE on its benchmarks, and use NRMSE (Takamoto et al., 2022) elsewhere. We cover both in-distribution and out-of-distribution settings. Runtime is pure per-sample inference for PDE-LENS and complete online discovery wall time (hyperparameter selection and fitting) for the non-pretrained methods. All non-pretrained methods observe only the first 80% of the time domain for equation selection and coefficient fitting, and the recovered equation is evaluated on the final 20%. PDE-LENS decodes the observed field once and refines the constants with BFGS on this first 80%. Appendix B includes the hardware details.

5.1 Pipeline Performance Comparison in PDE Discovery

Original MDBench. The non-pretrained baselines use the original MDBench implementation. For a controlled search space, their candidate library is replaced by the same fixed 59-atomic-term library used in our data generator. Note that among the five equations shown in Table 1, the heat-soil trajectory is OOD in its boundary and initial conditions: our pretraining dataset is limited to periodic boundary conditions and Fourier-generated periodic initial fields, whereas this MDBench case comes from a non-periodic FEniCS heat-soil setup. The other four can be considered in our training dataset distribution.

Unseen L=4 benchmark. We additionally evaluate on ten trajectories randomly sampled from the strict test partition of our four-term ($L = 4$) dataset. Table 2 reports the mean NMSE and runtime over the ten test trajectories; per-trajectory results are provided in Appendix D.

5.2 Comparing Neural Encoders within the E2E Framework

Representative neural SR methods commonly use self-attention-based set encoders to map numerical observations into conditional representations for a symbolic decoder (Biggio et al., 2021; Kamienny et al., 2022). We compare our PRISM with a NeSymReS encoder and a spatiotemporal ViT (Dosovitskiy et al., 2021) baseline, which provides a natural extension of a pointwise

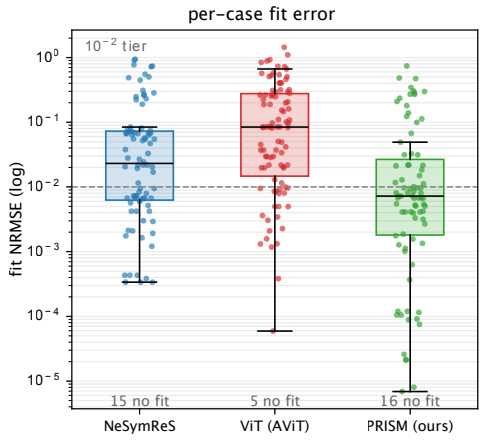


Figure 3: Field reconstruction NRMSE of the three encoders on 99 unseen PDE trajectories. Each point denotes a valid fit. The reconstruction and coefficient-refinement procedures are detailed in Appendix F. The dashed line marks the high-accuracy threshold of 10^{-2}.

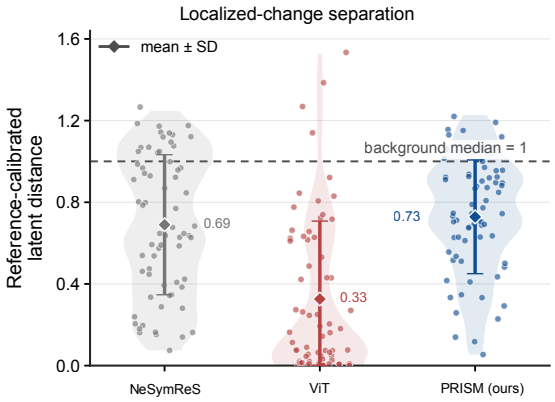


Figure 4: Separation of fine-scale dynamical changes in latent space. Each point corresponds to a pair of trajectories with similar coarse-scale fields but different high-frequency details. Distances are normalized by the median distance over all sample pairs.

Transformer to PDE fields (A standard Transformer scales poorly to PDE solution fields due to its quadratic attention mechanism).

PDE discovery task. From Table 1, we observe that neither ViT nor NeSymReS can generalize to both the Burgers and KdV test cases. They also fail entirely on the OOD heat equation. This is non-trivial: even after doubling the training budget, ViT and NeSymReS plateau at 2/5 NMSE $\leq 10^{-2}$, while ours frequently achieves 4/5 NMSE $\leq 10^{-2}$ over the course of training.

Besides, we further select 10 PDE structures spanning transport, nonlinear steepening, reaction–diffusion, and nonlinear diffusion. We sample three coefficient configurations and four initial conditions for each structure, yielding 120 candidate trajectories. Of these, 99 pass checks for resolution-refinement error, boundedness, trajectory variation, and equation-term activity, and form our unseen diagnostic set.

Figure 3 summarizes the results over all 99 trajectories. Among its 83 successful fits, PRISM achieves a median NRMSE of 7.22×10^{-3}, which is 11.7 and 3.2 times lower than those of ViT and NeSymReS respectively. On samples successfully fitted by both models in each comparison, paired Wilcoxon signed-rank tests on $\log_{10}$ NRMSE give $p = 3.4 \times 10^{-3}$ against NeSymReS and $p = 2.5 \times 10^{-9}$ against ViT. The evaluation protocol, results by equation structure, and success-rate curves are provided in Appendix F.

Table 3 complements the field-level analysis with evaluations of the recovered PDE right-hand sides. We define $S_{\mathrm{RHS}} = \max(0, 1 - \mathrm{NMSE}_{\mathrm{RHS}})$, where $S_{\mathrm{RHS}} \geq 0.99$ indicates accurate RHS recovery. To assess high-frequency fidelity, we compare the Fourier spectra of the predicted RHS and the reference time derivative. HF energy retention measures their per-mode energy overlap within the high-frequency band, while HF relative error measures the normalized discrepancy between their complex Fourier coefficients. The metric definitions and spectral evaluation protocols for PDE discovery and conventional SR are detailed in Appendix G.

Conventional symbolic regression. We further evaluate whether the PRISM encoder design is effective in conventional symbolic regression. We select the Point Transformer used in E2E-SR and the Set Transformer used in NeSymReS, the two most commonly adopted encoders in existing work as baselines. For both baselines, we follow the original hyperparameter settings reported in their respective repositories. Because our primary focus is PDE discovery, and training SR models from scratch is costly, we restrict the maximum input dimension to two in this controlled experiment.

We sample 10,000 points from each equation using the same sampling method as Kamienny et al. (2022) for both training and testing. The Set Transformer and our encoder can process all 10,000 points in a single forward pass. In contrast, the Point Transformer can process only approximately

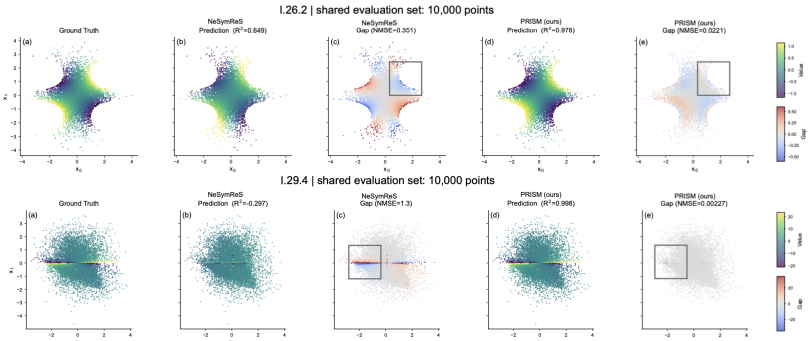


Figure 5: Two unseen Feynman equations that showcase the *high-frequency neglect* we mention. PDE-LENS outperforms all baselines. The full comparison is provided in Appendix E.

Table 3: Conventional symbolic regression and Fourier-spectrum results on the 15-equation Feynman benchmark, and the corresponding high-frequency results for PDE discovery on the 99 trajectories. HF means high frequency.

Symbolic Regression				
Method	$\#(R^2 > 0.99)$	Average $R^2 \uparrow$	HF energy retention $\uparrow$	HF relative error $\downarrow$
E2E-SR	8/15	0.699	0.467	0.551
NeSymReS	7/15	0.582	0.541	0.485
PRISM (ours)	**9/15**	**0.792**	**0.656**	**0.304**
PDE Discovery				
Method	$\#(S_{\mathrm{RHS}} \geq 0.99)$	Mean $S_{\mathrm{RHS}} \uparrow$	HF energy retention $\uparrow$	HF relative error $\downarrow$
NeSymReS	50/99	0.687	0.664	0.494
ViT	34/99	0.701	0.584	0.612
PRISM (ours)	**65/99**	**0.812**	**0.772**	**0.270**

200 points at a time because of the quadratic cost of self-attention. We therefore follow its own evaluation protocol. The models are evaluated on the challenging Feynman benchmark (Udrescu & Tegmark, 2020). We report the average R^2 and the number of equations for which $R^2 > 0.99$.

We also analyze the Fourier content of the decoded expressions. For each equation, we project the decoded expression and ground truth onto the same complex Fourier basis on shared irregular samples. The projection is fitted on one half of the samples and evaluated on the other half. To analyze the pattern shown in Figure 5, we report the same high-frequency metrics defined above. Moreover, Table 3 shows that our model achieves the highest HF energy retention and the lowest HF relative error (0.304), indicating that its decoded expressions more faithfully reproduce the high-frequency content of the target functions. Further details are provided in Appendix G.

5.3 Latent-Space Analysis

To better understand the performance differences among the three E2E models for PDE discovery, we examine the latent spaces learned by their encoders. The analysis uses 2,500 PDE fields with fixed initial and boundary conditions, while varying only the equation structure and coefficients. We extract a 256-dimensional pooled latent representation from each encoder.

Latent-Space Organization. Figure 6 projects these representations onto their first two principal components and colors them by four properties relevant to PDE dynamics: periodicity, reflection symmetry, energy trend, and propagation strength. In ViT, samples with high property values concentrate along a narrow diagonal branch, and the four properties exhibit similar color patterns. NeSymReS separates some high- and low-value samples, but several properties vary primarily along the same fan-shaped direction and concentrate near its two extremes. In contrast, PRISM produces a densely connected point cloud with smooth transitions in property values. Moreover, different properties vary along distinct directions: highly periodic samples lie toward the left, highly symmetric samples toward the upper left, strongly dissipative samples toward the lower left, and samples with stronger propagation closer to the center.

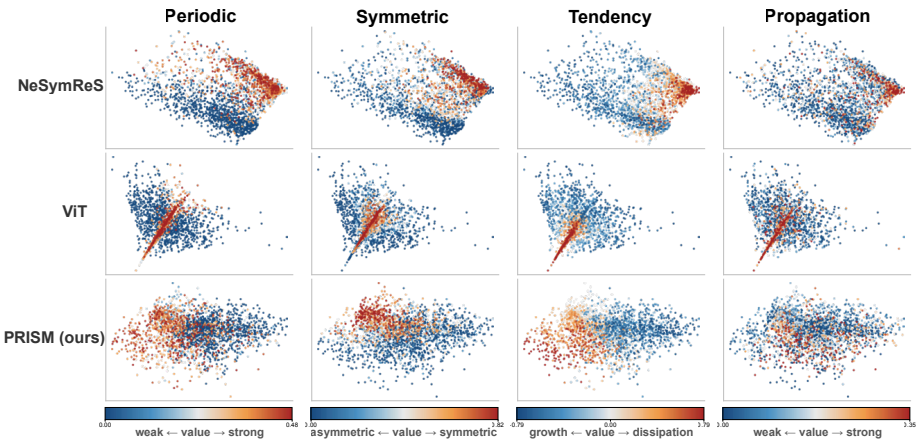


Figure 6: Dynamical properties in the latent spaces of different encoders. Each point corresponds to one PDE expression and coefficient configuration, with the initial and boundary conditions fixed. Coordinates are the first two principal components of the standardized pooled latent representations, and colors indicate periodicity, reflection symmetry, energy trend, and propagation strength. The two-dimensional projections explain 23.6%, 87.6%, and 15.3% of the latent variance for NeSymReS, ViT, and PRISM, respectively.

Table 4: Predictability of dynamical properties from the complete 256-dimensional latent representations. We report R^2 under five-fold expression-grouped cross-validation. Higher is better, and the best result in each column is shown in bold.

Encoder	Periodicity ↑	Symmetry ↑	Energy trend ↑	Propagation ↑
NeSymReS	0.635	0.695	**0.997**	0.183
ViT	0.192	0.420	0.600	0.153
PRISM (ours)	**0.836**	**0.966**	0.981	**0.907**

Predictability of Dynamical Properties. We further fit a ridge regression from the complete 256-dimensional representation to each property, using five-fold expression-grouped cross-validation. As shown in Table 4, PRISM achieves the highest R^2 on three of the four properties, with a particularly large advantage in propagation strength. These results complement the two-dimensional visualization, showing that PRISM retains multiple dynamical properties in a form that is readily accessible from the complete latent representation.

Sensitivity to Fine-Scale Dynamics. We also directly measure encoder sensitivity to fine-scale changes. We construct 67 pairs of trajectories that have similar coarse-scale solution fields but differ in their high-frequency details. For each pair, we compute the cosine distance between the standardized pooled latent representations and normalize it by the median distance. A value of 1 in Figure 4 therefore represents the encoder's background distance scale. For these coarse-scale-matched pairs, PRISM reaches a mean normalized distance of 0.73, compared with 0.69 for NeSymReS and 0.33 for ViT. PRISM preserves more information for distinguishing fine-scale dynamical changes.

6 Conclusion and Future Work

We presented PDE-LENS, an end-to-end framework that couples PDE discovery with learning representations of solution dynamics. Its PRISM encoder addresses high-frequency neglect by combining spectral and pointwise processing. Pretraining on 1.19 million trajectories yields improved numerical equation recovery over the evaluated neural encoders, while expression-grouped property prediction shows that the learned representations also capture accessible dynamical information. Experiments on conventional symbolic regression provide additional evidence for the encoder design beyond PDE discovery. Future work will extend the framework to higher-dimensional systems and more varied initial and boundary conditions.

AI USE STATEMENT

In this work, we have not used generative AI tools for any of the tasks with required disclosure: generative AI was not used to develop theoretical models or conceptual frameworks, formulate mathematical claims or assist in their proofs, propose or refine hypotheses, design or provide feedback on the research methodology or experiments, implement methods, clean or reformat data sets, or interpret results. These tasks were carried out solely by the authors.

Additionally, we used generative AI tools for one task with recommended disclosure: editing the paper to improve readability, i.e., language polishing and grammar correction of the manuscript text.

We note that our adaptation of CodePDE calls a large language model API (DeepSeek-v4-flash) as a component of the method itself to generate a proper solution method to solve the PDE and provide the generated data. This in-method API usage is a property of the system we construct, described in Section 4.2, and is not a use of generative AI for authoring this paper.

We have reviewed all AI-assisted work. All AI-polished text was reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Mehrdad Anvari, Hamidreza Marasi, and Hossein Kheiri. Implicit runge-kutta based sparse identification of governing equations in biologically motivated systems. *Scientific Reports*, 15, 09 2025.

Luca Biggio, Tommaso Bendinelli, Alexander Neitz, Aurélien Lucchi, and Giambattista Parascandolo. Neural symbolic regression that scales. *CoRR*, abs/2106.06427, 2021.

Gert-Jan Both, Subham Choudhury, Pierre Sens, and Remy Kusters. Deepmod: Deep learning for model discovery in noisy data. *Journal of Computational Physics*, 428:109985, March 2021. ISSN 0021-9991. doi: 10.1016/j.jcp.2020.109985. URL http://dx.doi.org/10.1016/j.jcp.2020.109985.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, March 2016. ISSN 1091-6490.

Stéphane d'Ascoli, Sören Becker, Philippe Schwaller, Alexander Mathis, and Niki Kilbertus. ODEFormer: Symbolic regression of dynamical systems with transformers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=TzoHLiGVMo.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Mengge Du, Yuntian Chen, and Dongxiao Zhang. Discover: Deep identification of symbolically concise open-form partial differential equations via enhanced reinforcement learning. *Phys. Rev. Res.*, 6:013182, Feb 2024.

Mingquan Feng, Yixin Huang, Yizhou Liu, Bofang Jiang, and Junchi Yan. PhysPDE: Rethinking PDE discovery and a physical HYpothesis selection benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=G3CpBCQwNh.

Xiaoxu Han, Chengzhen Ning, Jinghui Zhong, Fubiao Yang, Yu Wang, and Xin Mu. Discovering mathematical equations with diffusion language model, 2025.

Pierre-Alexandre Kamienny, Stéphane d'Ascoli, Guillaume Lample, and François Charton. End-to-end symbolic regression with transformers. *Advances in Neural Information Processing Systems*, 35:10269–10281, 2022.

Qian Li, Yuxiao Hu, Juncheng Liu, and Yuntian Chen. GenSR: Symbolic regression based on equation generative space. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Shanda Li, Tanya Marwah, Junhong Shen, Weiwei Sun, Andrej Risteski, Yiming Yang, and Ameet Talwalkar. CodePDE: An inference framework for LLM-driven PDE solver generation. *Transactions on Machine Learning Research*, 2026b. ISSN 2835-8856. URL https://openreview.net/forum?id=eG3Qy5Oux6. J2C Certification.

Yanjie Li, Jingyi Liu, Min Wu, Lina Yu, Weijun Li, Xin Ning, Wenqiang Li, Meilan Hao, Yusong Deng, and Shu Wei. Mmsr: symbolic regression is a multi-modal information fusion task. *Information Fusion*, 114:102681, 2025.

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

Zongyi Li, Nikola Borislavov Kovachki, Chris Choy, Boyi Li, Jean Kossaifi, Shourya Prakash Otta, Mohammad Amin Nabian, Maximilian Stadler, Christian Hundt, Kamyar Azizzadenesheli, and Anima Anandkumar. Geometry-informed neural operator for large-scale 3d PDEs. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=86dXbqT5Ua.

Senwei Liang and Haizhao Yang. Finite expression method for solving high-dimensional partial differential equations. *Journal of Machine Learning Research*, 26(138):1–31, 2025.

Levi Lingsch, Mike Y. Michelis, Emmanuel de Bezenac, Sirani M. Perera, Robert K. Katzschmann, and Siddhartha Mishra. Beyond regular grids: Fourier-based neural operators on arbitrary domains, 2024. URL https://arxiv.org/abs/2305.19663.

Phillip Lippe, Bastiaan S. Veeling, Paris Perdikaris, Richard E Turner, and Johannes Brandstetter. PDE-refiner: Achieving accurate long rollouts with neural PDE solvers. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Qv6468llWS.

Kazem Meidani, Parshin Shojaee, Chandan K. Reddy, and Amir Barati Farimani. SNIP: Bridging mathematical symbolic and numeric realms with unified pre-training. In *The Twelfth International Conference on Learning Representations*, 2024.

Daniel A. Messenger and David M. Bortz. Weak sindy for partial differential equations. *Journal of Computational Physics*, 443:110525, October 2021. ISSN 0021-9991. doi: 10.1016/j.jcp.2021.110525. URL http://dx.doi.org/10.1016/j.jcp.2021.110525.

Terrell N. Mundhenk, Mikel Landajuela, Ruben Glatt, Claudio P. Santiago, Daniel faissol, and Brenden K. Petersen. Symbolic regression via deep reinforcement learning enhanced genetic programming seeding. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks, 2019. URL https://arxiv.org/abs/1806.08734.

Samuel H. Rudy, Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3(4):e1602614, 2017.

Parshin Shojaee, Kazem Meidani, Amir Barati Farimani, and Chandan Reddy. Transformer-based planning for symbolic regression. *Advances in Neural Information Processing Systems*, 36: 45907–45919, 2023.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K. Reddy. LLM-SR: Scientific equation discovery via programming with large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions, 2020. URL https://arxiv.org/abs/2006.09661.

Jingmin Sun, Yuxuan Liu, Zecheng Zhang, and Hayden Schaeffer. Towards a foundation model for partial differential equations: Multi-operator learning and extrapolation, 2025. URL https://arxiv.org/abs/2404.12355.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Dan MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An extensive benchmark for scientific machine learning. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=dh_MkX0QfrK.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains, 2020. URL https://arxiv.org/abs/2006.10739.

Yuan Tian, Wenqi Zhou, Hao Dong, David S Kammer, and Olga Fink. Sym-q: Adaptive symbolic regression via sequential decision-making. *arXiv preprint arXiv:2402.05306*, 2024.

Tapas Tripura and Souvik Chakraborty. Wavelet neural operator for solving parametric partial differential equations in computational mechanics problems. *Computer Methods in Applied Mechanics and Engineering*, 404:115783, 2023. ISSN 0045-7825. doi: https://doi.org/10.1016/j.cma.2022.115783. URL https://www.sciencedirect.com/science/article/pii/S0045782522007393.

Silviu-Marian Udrescu and Max Tegmark. Ai feynman: a physics-inspired method for symbolic regression, 2020.

Chuyang Xiang, Yichen Wei, and Junchi Yan. MOD-SR: Unifying multimodal learning and direct optimization with gradient-guided diffusion model for symbolic regression. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=p06oR4sXgz.

Hao Xu, Yuntian Chen, Rui Cao, Tianning Tang, Mengge Du, Jian Li, Adrian H. Callaghan, and Dongxiao Zhang. Generative discovery of partial differential equations by learning from math handbooks. *Nature Communications*, 16(1):10255, 2025. doi: 10.1038/s41467-025-65114-2. URL https://www.nature.com/articles/s41467-025-65114-2.

Zihan Yu, Jingtao Ding, Yong Li, and Depeng Jin. Symbolic regression via MDLformer-guided search: from minimizing prediction error to minimizing description length. In *The Thirteenth International Conference on Learning Representations*, 2025.

Amirmohammad Ziaei Bideh, Aleksandra Georgievska, and Jonathan Gryak. Mdbench: Benchmarking data-driven methods for model discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 19746–19754, 2026. doi: 10.1609/aaai.v40i24.39056. URL https://ojs.aaai.org/index.php/AAAI/article/view/39056.

A Constant Optimization

For each decoded equation, we keep its symbolic structure fixed and replace all predicted numerical constants with optimizable variables $\mathbf{c} = (c_1, \ldots, c_m)$. The equation is first converted to the physical field coordinates and the target grid coordinates described above. We then fix any predicted constant multiplying u_t and divide the residual by the resulting u_t coefficient, producing the monic form

$$\mathcal{R}(u; \mathbf{c}) = u_t + \mathcal{F}(u, u_x, u_{xx}, \ldots; \mathbf{c}). \tag{6}$$

Only the remaining non-leading constants are optimized.

Spatial and temporal derivatives are computed directly from the same observed field supplied to the encoder. Depending on the boundary condition and noise setting, we use periodic spectral differentiation, local polynomial differentiation, or finite differences. The constants are fitted by minimizing the normalized residual energy

$$\mathcal{L}_{\text{const}}(\mathbf{c}) = \frac{\|\mathcal{R}(u; \mathbf{c})\|_2^2}{\|u_t\|_2^2 + \epsilon}. \tag{7}$$

This objective uses only the observed trajectory; ground-truth coefficients are never provided to the optimizer.

Following Feng et al. (2025), we solve the resulting nonlinear least-squares problem with BFGS for at most 80 iterations and tolerance 10^{-6}. To reduce sensitivity to initialization, we run BFGS from three starting points: the zero vector, the constants decoded by the model, and the decoded constants clipped to $[-10, 10]$. The finite solution with the lowest residual objective is substituted back into the expression. Candidates with non-finite or numerically invalid residuals are discarded, and the resulting equation is finally evaluated by comparing its predicted right-hand side against the reference time derivative using NMSE (Ziaei Bideh et al., 2026).

B Hardware Details

All experiments are run on a node with eight Ascend 910B3 NPUs and a 192-core Kunpeng-920 CPU, using the CANN 8.0.RC3 software stack. Both the PDE discovery model and the general symbolic regression model are trained on all eight NPUs and evaluated on a single NPU.

C Hyperparameter Settings

PDE discovery. The NeSymReS baseline uses a Set Transformer with hidden dimension 256, two layers, eight attention heads, and 50 inducing points. The ViT baseline uses 16×16 patches, embedding dimension 256, six encoder layers, and eight attention heads with axial attention. PRISM uses four spectral–pointwise layers with 32 hidden channels and frequency indices $k_x, k_t \in \{-16, \ldots, 15\}$, followed by the Moment Bridge. The common decoding and coefficient-refinement settings for the 99-trajectory comparison are given in Appendix F.

Conventional Symbolic Regression All models are trained from scratch for 100,000 optimizer steps with maximum input dimension two and output dimension one. PRISM uses four spectral–pointwise layers with 32 channels, maximum frequency 64, eight detail points, eight nearest neighbors, packet frequencies $(2, 8, 24, 64)$, and Adam with learning rate 10^{-4}, 10,000 warm-up updates, and 40,000 tokens per batch. NeSymReS follows its faithful Set Transformer configuration (512 hidden dimensions, eight heads, 50 inducing points, five encoder and decoder layers) with the same optimizer and token budget. Both models process all 10,000 Feynman points in one bag.

E2E-SR retains its original protocol: 200 input points per bag, maximum sequence length 200, 10,000 tokens per batch, and 3,000 steps per epoch. The 10,000 sampled points are therefore evaluated through repeated 200-point bags.

D UNSEEN FOUR-TERM BENCHMARK

Table 5 reports the complete per-trajectory derivative NMSE values for the unseen benchmark described in Section 5.1. PDE-FIND (Rudy et al., 2017) uses a fixed, test-independent expanded library with all 59 generator-supported atomic families and ten pre-specified internal-parameter instances per parameterized family. Its column-normalized sparse regression selects thresholds and ridge penalties using only the training/validation portion of each trajectory. Bayesian uses its complete effective search grid. For WSINDy, we retain all sparsity thresholds, basis choices, derivative orders, and polynomial orders and use 200 weak-form integration points. DeepMoD (Both et al., 2021) uses the same compact configuration on every trajectory: polynomial order four, derivative order four, thresholds $\{0.1, 0.3, 0.5\}$, and at most 10,000 optimization iterations.

Table 5: Per-trajectory NMSE on ten randomly sampled training-set-excluded $L = 4$ trajectories. Lower is better.

Trajectory	PDE-LENS	PDE-FIND	Bayesian	WSINDy	DeepMoD
L4-0868-c09-ic07	1.506×10^{-5}	3.710×10^{-3}	2.011×10^{-2}	3.770×10^{-3}	1.228×10^{-1}
L4-4118-c05-ic04	2.832×10^{-5}	7.746×10^{-3}	7.683×10^{-3}	8.113×10^{-3}	2.767×10^{-2}
L4-2795-c05-ic04	4.973×10^{-5}	5.814×10^{-3}	1.563×10^{-2}	6.375×10^{-3}	1.172×10^{-2}
L4-4540-c02-ic07	5.183×10^{-5}	2.817×10^{-4}	3.253×10^{-4}	3.928×10^{-5}	5.879×10^{-3}
L4-4065-c00-ic02	5.704×10^{-5}	1.080×10^{-4}	5.863×10^{-5}	1.145×10^{-4}	9.681×10^{-4}
L4-1445-c02-ic01	2.089×10^{-3}	9.591×10^{-1}	9.589×10^{-1}	1.208	8.777×10^{-1}
L4-4761-c07-ic03	2.961×10^{-3}	1.487×10^{-3}	9.830×10^{-4}	1.805×10^{-3}	4.026×10^{-4}
L4-2484-c02-ic02	2.987×10^{-3}	1.655×10^{-2}	1.371×10^{-2}	1.431×10^{-2}	1.674×10^{-2}
L4-1750-c01-ic05	5.218×10^{-3}	4.231×10^{-2}	3.513×10^{-2}	1.066×10^{-1}	1.541×10^{-1}
L4-3058-c06-ic02	7.708×10^{-3}	8.052×10^{-3}	9.339×10^{-3}	9.352×10^{-3}	2.250×10^{-1}

E COMPLETE COMPARISON FOR CONVENTIONAL SYMBOLIC REGRESSION

Figure 7 presents a qualitative comparison on eight of the fifteen two-dimensional Feynman equations used in our conventional symbolic regression evaluation. These examples span cases ranging from near-exact recovery by both methods to more challenging equations with substantial differences in prediction accuracy. For each equation, both methods are evaluated on the same 10,000 input points, enabling a direct comparison of their predictions and pointwise residuals. As shown, our performances are better than the baseline, showing more details of the high-frequency part in the field, at the boundaries or areas near zero.

F EVALUATION PROTOCOL AND RESULTS FOR THE ENCODER COMPARISON

Decoding and reconstruction. All three encoders are evaluated using beam search with beam size 10, a maximum decoding length of 200, and candidate reranking. For each trajectory, let u_n denote the observed field at saved time t_n, and let $\widehat{F}_{\mathbf{c}}$ denote the recovered PDE right-hand side. We evaluate this RHS on the observed field and accumulate its values with explicit Euler:

$$\widehat{u}_0(\mathbf{c}) = u_0, \qquad \widehat{u}_{n+1}(\mathbf{c}) = \widehat{u}_n(\mathbf{c}) + \Delta t\, \widehat{F}_{\mathbf{c}}[u_n].$$

The notation $\widehat{F}_{\mathbf{c}}[u_n]$ includes the spatial derivatives evaluated from the observed field. Every RHS evaluation uses u_n, so this procedure measures reconstruction consistency along the observed trajectory rather than autonomous rollout accuracy.

Field reconstruction metric. We compute

$$\mathrm{NRMSE}_{\mathrm{field}} = \frac{\sqrt{\mathrm{mean}_{n,i}\left[(\widehat{u}_{n,i}(\mathbf{c}^\star) - u_{n,i})^2\right]}}{\max\left(\sqrt{\mathrm{mean}_{n,i}\left[(u_{n,i} - \bar{u})^2\right]}, 10^{-12}\right)},$$

where $\bar{u}$ is the mean over all observed space–time points. Reconstruction-error distributions and conditional medians include only valid fits, with fitting coverage reported separately.

Results by equation structure. Figure 8 decomposes the aggregate results by equation structure. PRISM achieves the lowest conditional median NRMSE on six of the ten structures, spanning derivative-only, nonlinear-advection, reaction–diffusion, and mixed convection–diffusion equations. The largest separation occurs for the nonlinear-advection structure containing uu_x, where PRISM reaches a median NRMSE of approximately 10^{-4}, more than two orders of magnitude below both alternatives. Its advantage therefore does not arise from a single equation family, but extends across several distinct dynamical mechanisms. Because these medians are conditional on successful fits, the parenthetical counts additionally expose differences in coverage rather than conflating missing fits with numerical error.

The success curves provide a complementary view across error thresholds. PRISM retains nearly the same fitting coverage as NeSymReS, while shifting most successful fits toward substantially lower NRMSE. At the 10^{-2} threshold, more than three-fifths of its fitted trajectories meet the accuracy criterion, compared with approximately two-fifths for NeSymReS and one-quarter for ViT. ViT produces the highest overall coverage, but its curve is shifted toward larger errors. Together, the two panels show that PRISM improves high-accuracy equation recovery across multiple PDE structures.

G High-Frequency Evaluation

Evaluation targets. We evaluate the spectral fidelity of the recovered expressions. For PDE discovery, the reference signal is the time derivative $r(x,t) = u_t(x,t)$, and the predicted signal $\widehat{r}(x,t)$ is the recovered PDE right-hand side evaluated along the observed trajectory. For conventional SR, the reference and predicted signals are the values of the ground-truth and decoded expressions, respectively, evaluated at shared input locations.

Metric definitions. Let $a_{\boldsymbol{k}}$ and $b_{\boldsymbol{k}}$ denote the complex Fourier coefficients of the reference and predicted signals, estimated using the same basis and preprocessing. Let $\mathcal{H}$ denote the set of high-frequency modes. For a reference signal with nonzero energy in this band, we define HF energy retention as

$$R_{\mathrm{HF}} = \frac{\sum_{\boldsymbol{k}\in\mathcal{H}} \min\left(|a_{\boldsymbol{k}}|^2, |b_{\boldsymbol{k}}|^2\right)}{\sum_{\boldsymbol{k}\in\mathcal{H}} |a_{\boldsymbol{k}}|^2}, \tag{8}$$

and HF relative error as

$$E_{\mathrm{HF}} = \left(\frac{\sum_{\boldsymbol{k}\in\mathcal{H}} |b_{\boldsymbol{k}} - a_{\boldsymbol{k}}|^2}{\sum_{\boldsymbol{k}\in\mathcal{H}} |a_{\boldsymbol{k}}|^2}\right)^{1/2}. \tag{9}$$

Equation 8 measures the overlap of reference and predicted energy at corresponding frequency modes and lies between zero and one. Excess predicted energy does not increase the overlap beyond the reference energy at that mode. Equation 9 additionally accounts for phase discrepancies and excess spectral energy. Higher retention and lower relative error indicate better high-frequency recovery.

We compute two-dimensional Fourier transforms of the reference and predicted RHS fields on the same space–time grid. Each field is centered by subtracting its mean. For nonperiodic cases, we additionally apply a separable Tukey window with parameter 0.25 along both axes.

Let k_x and ω denote the discrete spatial and temporal frequencies, and let $k_{\max}$ and $\omega_{\max}$ be their maximum absolute values. We define

$$\mathcal{H} = \left\{(k_x, \omega) : |k_x| \geq \frac{k_{\max}}{2} \text{ or } |\omega| \geq \frac{\omega_{\max}}{2}\right\}. \tag{10}$$

Thus, a mode belongs to the high-frequency band if its spatial or temporal frequency lies in the upper half of the corresponding discrete frequency range.

We compute each metric separately for each trajectory and report the arithmetic mean over all 99 trajectories. Cases without valid recovery receive an RHS score of zero, HF energy retention of zero, and HF relative error of one. The latter corresponds to predicting zero Fourier coefficients in the high-frequency band.

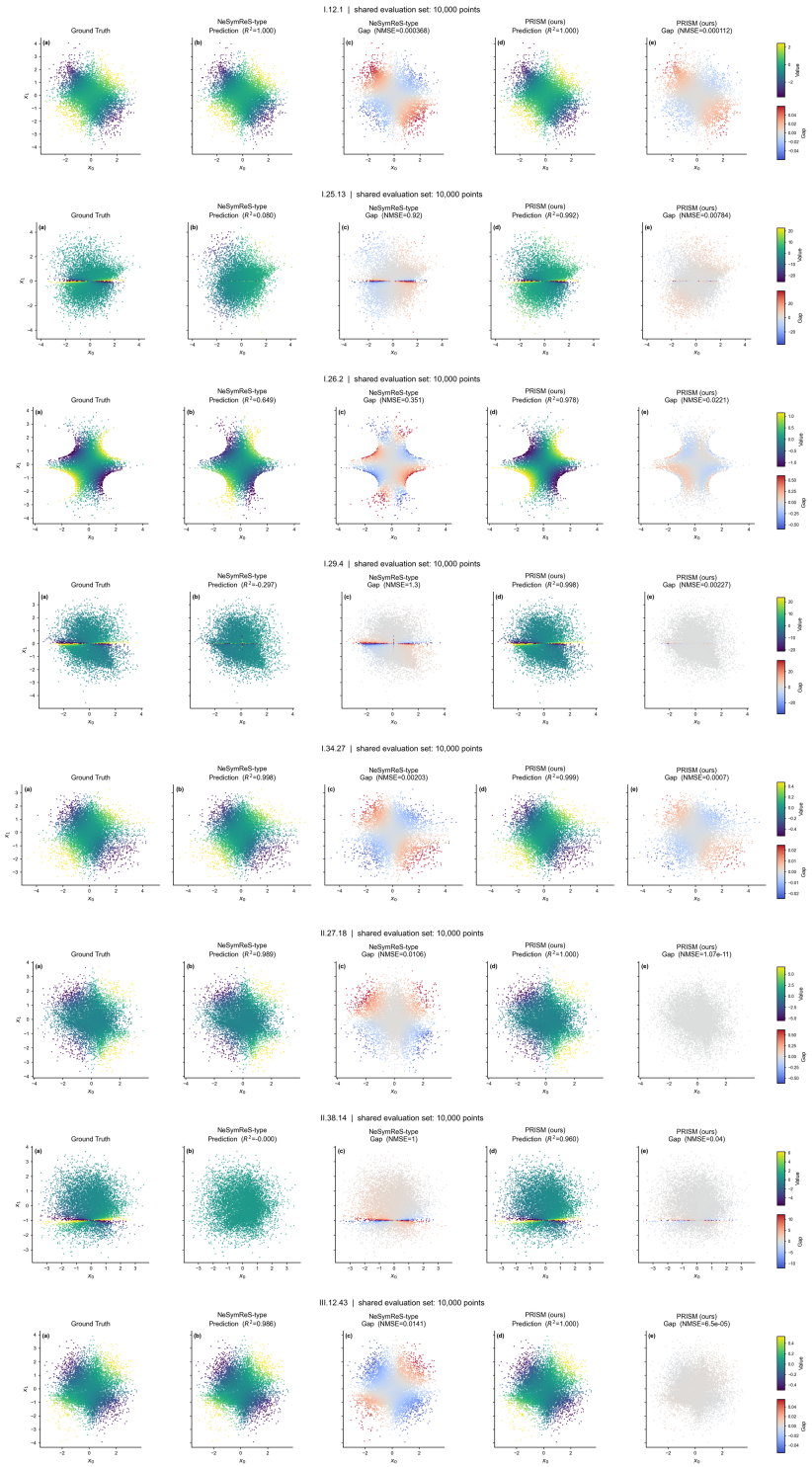


Figure 7: Qualitative comparison on eight two-dimensional Feynman equations. Each row corresponds to one equation evaluated on a shared set of 10,000 input points. From left to right, the columns show the ground-truth values, the NeSymReS-type prediction, its signed residual $\widehat{y} - y$, the PRISM prediction, and its signed residual. Prediction panels report R^2, while residual panels report NMSE. Within each row, the prediction panels use the same value scale, and the two residual panels share a symmetric color scale. PRISM more faithfully recovers the target variation in the challenging cases, while both methods closely match the ground truth on several easier equations.

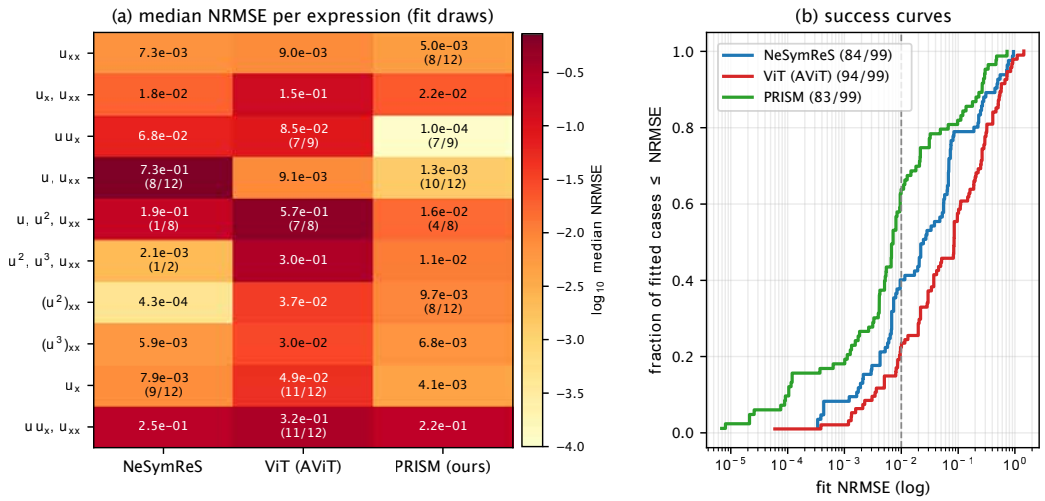


Figure 8: Performance breakdown over the 99 unseen trajectories. **Left:** Conditional median NRMSE for each equation structure, computed over trajectories for which a fitted equation is available. Parentheses report the number of fitted trajectories when the coverage is incomplete. **Right:** Empirical cumulative distributions of NRMSE over fitted trajectories. No-fit cases are excluded from the curves, while the legend reports the corresponding coverage over all 99 trajectories. The dashed line marks the high-accuracy threshold of 10^{-2}.