Mimicking the Physicist's Eye: A VLM-centric Approach For Physics Formula Discovery

Jiaqi Liu[1,3] **Songning Lai**[2] **Pengze Li**[3,4] **Di Yu**[3,5] **Wenjie Zhou**[6,9]
Yiyang Zhou[1] **Peng Xia**[1] **Zijun Wang**[7] **Xi Chen**[4] **Shixiang Tang**[8]
Lei Bai[3] **Wanli Ouyang**[3,8] **Mingyu Ding**[1] **Huaxiu Yao**[1] **Aoran Wang**[3*]

[1]University of North Carolina at Chapel Hill [2]HKUST (Guangzhou)
[3]Shanghai AI Laboratory [4]Fudan University [5]Tsinghua University [6]Nankai University
[7]UC Santa Cruz [8]The Chinese University of Hong Kong [9]Shanghai Innovation Institute

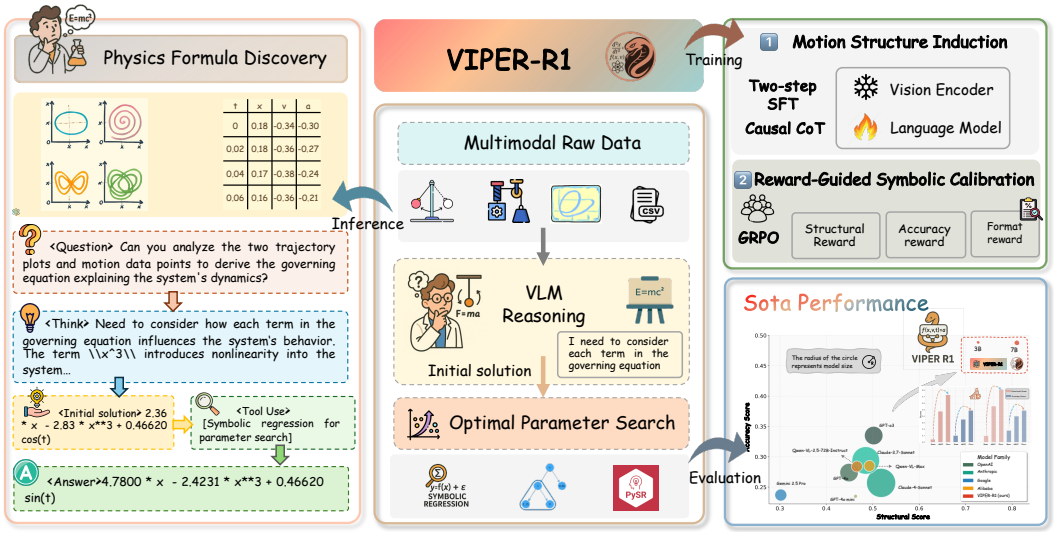


Figure 1: Overview of **VIPER-R1**, a multimodal framework for visual formula discovery from low-dimensional dynamical evidence. The model is trained via Motion Structure Induction (MSI) with Causal CoT supervision and Reward-Guided Symbolic Calibration (RGSC) for structural refinement. At inference time, an external symbolic regression module for Symbolic Residual Realignment (SR^2) can optionally reduce residual numerical mismatch. The model improves structural and exact-match accuracy on the PhysSymbol mechanics benchmark.

Abstract

Automated discovery of governing equations remains challenging when models rely only on numerical tokens and ignore the phase-space and trajectory evidence that is informative in low-dimensional dynamics. We propose **V**isual **I**nduction for **P**hysics-based **E**quation **R**easoning (**VIPER-R1**), a multimodal framework for visual formula discovery from phase portraits, trajectory plots, and aligned measurements. VIPER-R1 uses a supervised Motion Structure Induction (MSI) curriculum for joint rationale-and-equation generation, and is then refined by **Reward-Guided Symbolic Calibration (RGSC)**, a GRPO-based stage optimized with a structure-aware reward that emphasizes *topological correctness* over coefficient matching. At inference time, we optionally apply **Symbolic Residual Realignment (SR^2)**, an external symbolic regression module that models only the residual numerical

*Corresponding author: wangaoran@pjlab.org.cn

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

mismatch of the base hypothesis. We release **PhysSymbol**, a 10,000-instance two-part multimodal corpus; the quantitative evaluation in this paper uses its Part 1 mechanics split, which provides paired visualizations, trajectory data, and expert C-CoT traces. On this benchmark, VIPER-R1-7B achieves the best structural score ($S_{\text{struct}} = 0.812$), is effectively tied with VIPER-R1-3B on exact-match accuracy ($S_{\text{acc}} = 0.487$ vs. 0.488), and, after optional SR^2 refinement, reaches a Post-SR^2 MSE of 0.032, nearly $3\times$ lower than the best VLM baseline. These results show that VIPER-R1 improves symbolic structure induction in focused low-dimensional dynamics by combining paired visual evidence with structure-aware calibration to produce stronger first-order hypotheses for downstream refinement.

Code and model: `https://github.com/Jiaaqiliu/VIPER-R1`.

1 Introduction

The automated discovery of governing equations for low-dimensional physical systems from observational data remains a central challenge at the intersection of artificial intelligence and the natural sciences [34, 36, 10]. This endeavor is important for accelerating scientific analysis by uncovering compact dynamical laws from trajectories, simulations, and controlled measurements [15, 26]. Recent advances have established two parallel yet distinct research tracks: sophisticated symbolic regression (SR) algorithms that navigate immense combinatorial spaces to identify fitting equations [13, 7], and the emergence of Large Language Models (LLMs) demonstrating remarkable ability to perform in-context symbolic reasoning from textual data [17, 8, 29]. While both approaches have laid critical foundations, they often operate without the visual evidence that scientists routinely use when analyzing low-dimensional dynamical systems.

This limitation can be seen as a form of "sensory deprivation," where reliance on uni-modal symbolic data blinds models to the rich visual representations that physicists routinely exploit. Human scientific reasoning is inherently multimodal: physicists interpret visual patterns in phase portraits to infer conservation laws, recognize decay envelopes to hypothesize damping forces, and identify superposition effects to constrain theoretical possibilities [31]. Such visual intuition provides powerful pre-symbolic heuristics for navigating the vast space of candidate theories.

Recent advances in LLM-based scientific discovery partly address these issues. LLM-SR [29] generates equation hypotheses from embedded scientific knowledge, while Scientific Generative Agents [17] pair LLM-based generation with simulation validation. Yet these methods still suffer from "sensory deprivation," lacking the ability to incorporate visual evidence. Furthermore, concerns about memorization versus genuine discovery [37, 30] underscore the need for approaches that perform authentic data-driven reasoning rather than recalling known formulas.

By neglecting the crucial visual perceptual channel, existing methods are fundamentally constrained. They often resort to computationally expensive searches through vast equation spaces [35], exhibit brittle token-matching behaviors, and struggle to exploit the hypothesis-forming cues available in phase space and temporal evolution. This limitation becomes particularly pronounced in low-dimensional dynamical systems, where visual patterns provide information that is difficult to recover from purely numerical inputs alone.

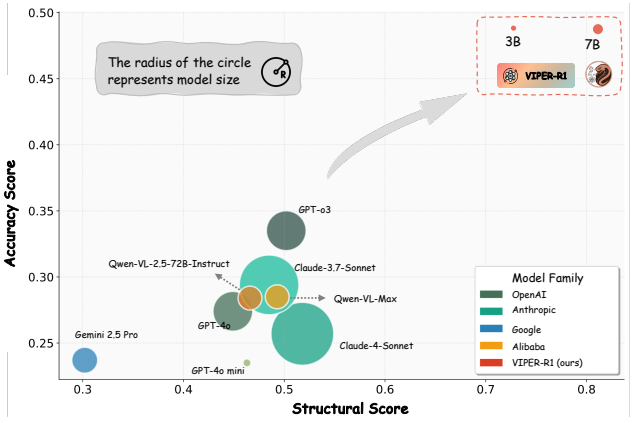


Figure 2: VIPER-R1 against representative VLM baselines on PhysSymbol.

To bridge the gap between perceptual evidence and abstract symbolic formalism, we introduce **VIPER-R1**, a **V**isual **I**nduction framework for **P**hysics-based **E**quation **R**easoning. VIPER-R1 improves symbolic structure induction in focused low-dimensional dynamics by combining paired visual evidence

with structure-aware calibration, thereby producing stronger first-order hypotheses for downstream refinement.

VIPER-R1 follows a staged training-and-refinement pipeline inspired by the workflow from qualitative pattern recognition to quantitative law formulation. First, **Motion Structure Induction (MSI)** uses an MSI-1 path for joint rationale-and-equation training under *causal* CoT (C-CoT) supervision, plus a separate MSI-2 branch retained as an oracle-rationale formulation branch for analysis. Second, **Reward-Guided Symbolic Calibration (RGSC)** applies reinforcement learning with **Group Relative Policy Optimization (GRPO)** [28] to further calibrate the hypothesis-generation policy using a structure-aware reward that explicitly favors *topological correctness* of the discovered law over coefficient matching. Finally, during inference, VIPER-R1 can invoke an external symbolic regression tool for **Symbolic Residual Realignment ($\mathbf{SR}^2$)**, which models only the residual mismatch of the base symbolic hypothesis and yields an end-to-end refined formula.

To support this research, we release **PhysSymbol**, a 10,000-instance two-part multimodal corpus. Part 1, the mechanics split used for the paper's training and quantitative evaluation, provides paired phase-space visualizations, trajectory measurements, ground-truth laws, and expert-annotated C-CoT traces that link visual motifs to symbolic terms. Part 2 is released as an auxiliary research resource.

Our contributions are summarized as follows:

- We propose **VIPER-R1**, a multimodal framework for visual formula discovery in low-dimensional dynamical systems that integrates phase-space plots, trajectory measurements, and symbolic derivation.
- We develop a hierarchical strategy consisting of **Motion Structure Induction (MSI)** for supervised hypothesis formation, **Reward-Guided Symbolic Calibration (RGSC)** via GRPO for structure-focused refinement, and an external inference-time refinement stage, **Symbolic Residual Realignment ($\mathbf{SR}^2$)**, which models residual mismatch with a symbolic regression tool.
- We introduce **PhysSymbol**, a 10,000-instance two-part multimodal physics corpus whose Part 1 mechanics split provides paired visualizations and expert-annotated C-CoT traces for the main experiments.

2 Related Work

Symbolic regression has progressed from genetic programming [12] to recursive and evolutionary systems such as AI Feynman [34] and PySR [7], as well as neural and hybrid methods that combine sequence models with RL, MCTS, or guided search [5, 11, 25, 33, 21, 18]. A persistent difficulty is still the combinatorial search for physically meaningful expressions [35, 29]. In parallel, LLM-based scientific discovery has explored skeleton generation, in-context reasoning, simulator-guided optimization, and scientific concept libraries [29, 19, 17, 8], while also raising memorization concerns that motivate contamination-resistant evaluation [37, 20, 30]. Multimodal models now support scientific figure understanding and even equation discovery from videos or visual scenes [40, 32, 16, 39, 24, 3, 6, 14], but they typically do not operate on the paired analytical plots that physicists use to inspect low-dimensional dynamics. VIPER-R1 targets this gap by training a VLM to infer symbolic laws directly from paired $x(t)$ and $v(x)$ evidence, with external symbolic regression used only as optional residual refinement.

3 Methodology

Our proposed framework consists of two supervised MSI branches, an RGSC stage, and an inference-time external refinement module, as illustrated in Figure 3. MSI-1 and MSI-2 both use C-CoT supervision, but they serve different purposes: MSI-1 learns joint rationale-and-equation generation, while MSI-2 is retained as an oracle-rationale formulation upper bound that isolates formulation quality under gold rationales and is not part of the deployed inference path. During inference, an external symbolic regression module can optionally refine residual numerical mismatch of the resulting hypothesis.

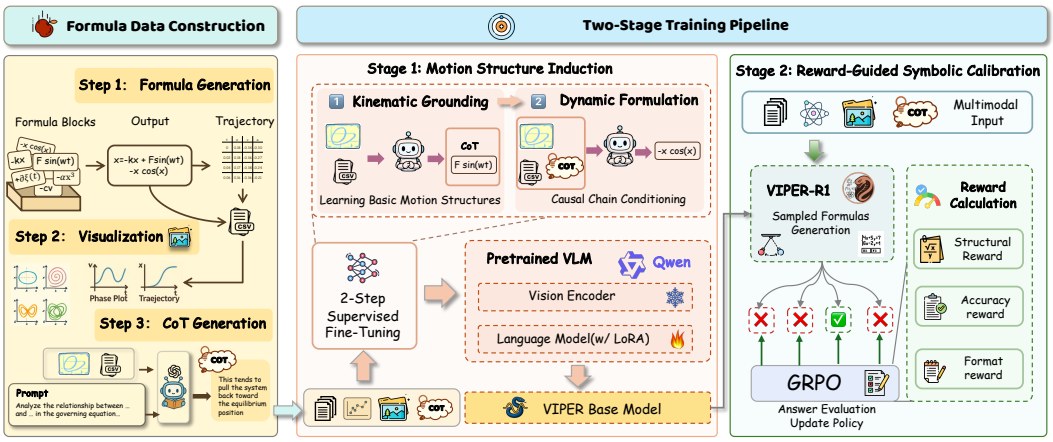


Figure 3: **Framework of VIPER-R1**. VIPER-R1 introduces a training-and-refinement framework for visual formula discovery. First, Motion Structure Induction (MSI) provides two supervised branches: MSI-1 learns joint rationale-and-equation generation, and MSI-2 learns rationale-conditioned symbolic formulation as an oracle-rationale formulation upper bound retained for analysis rather than for the deployed inference path. Reward-Guided Symbolic Calibration (RGSC) then sharpens the policy toward topologically correct laws. At inference time, an external SR^2 module can optionally refine residual numerical mismatch.

3.1 Problem Definition

We study *multimodal physical law discovery*, referring to learning a mapping from observations to a symbolic governing equation. Given empirical evidence $\mathcal{E} = \{\mathcal{V}, \mathcal{D}\}$ consisting of visual depictions $\mathcal{V}$ (for kinematic systems, a phase portrait I_{phase}, and a trajectory plot $I_{\text{trajectory}}$) and numerical measurements $\mathcal{D}$ (e.g., measurements of position, velocity, and acceleration over time $\{(t_i, x(t_i), v(t_i), a(t_i))\}_{i=1}^{T}$), the goal is to infer an interpretable governing law and output a symbolic hypothesis $\hat{\mathcal{F}}$:

$$\pi_\theta : \mathcal{E} \mapsto \hat{\mathcal{F}}, \tag{1}$$

where π_θ denotes the VIPER-R1 vision-language policy. The output $\hat{\mathcal{F}}$ is a symbolic equation (terms + coefficients) intended to be human-readable and predictive.

3.2 Motion Structure Induction (MSI)

MSI is a two-stage curriculum that trains VIPER-R1 to (i) form a causal rationale from multimodal observations and (ii) translate it into a precise symbolic law.

3.2.1 MSI-1: Joint Causal Reasoning and Symbolic Hypothesis

The initial stage mirrors a scientist's first encounter with a new phenomenon: concurrently observing, reasoning, and formulating a preliminary idea. Given $\mathcal{E}$, the model generates a C-CoT: C followed by a candidate law $\mathcal{F}$, serialized in a fixed template and treated as an initial symbolic ansatz. Let $\mathcal{D}_{\text{PS}}$ denote the training corpus, where each instance provides empirical evidence $\mathcal{E}$ paired with expert rationale C and the ground-truth law $\mathcal{F}^\star$, i.e., $(\mathcal{E}, C, \mathcal{F}^\star) \sim \mathcal{D}_{\text{PS}}$. We maximize the likelihood of the concatenated target $Y = (C, \mathcal{F}^\star)$ over $\mathcal{D}_{\text{PS}}$:

$$\mathcal{L}_{\text{MSI-1}} = -\mathbb{E}_{(\mathcal{E},Y)\sim\mathcal{D}_{\text{PS}}} \sum_{t=1}^{|Y|} \log \pi_\theta(y_t \mid \mathcal{E}, y_{<t}). \tag{2}$$

This joint objective is crucial: it compels the VIPER-R1 to ground its symbolic output in an explicit, physically-motivated reasoning process. The model must learn not just *what* the governing law is, but *why* it takes that form, based on visual cues within the evidence. We denote the resulting checkpoint by π_{MSI1}.

3.2.2 MSI-2: C-CoT-Guided Symbolic Formulation

The second stage of the curriculum refines VIPER-R1's ability to translate a well-formed physical rationale into a precise symbolic law. To focus learning on symbolic syntax and semantics, MSI-2 conditions on the ground-truth rationale C and trains the model to generate the ground-truth law $\mathcal{F}^\star$:

$$\mathcal{L}_{\text{MSI-2}} = -\mathbb{E}_{(\mathcal{E},C,\mathcal{F}^\star)\sim\mathcal{D}_{\text{PS}}} \sum_{t=1}^{|\mathcal{F}^\star|} \log \pi_\theta(f_t \mid \mathcal{E}, C, f_{<t}). \tag{3}$$

The loss is computed exclusively on the tokens of the symbolic law $\mathcal{F}^\star$. We denote the resulting supervised formulation branch by π_{MSI2}. MSI-2 is retained as a separate supervised branch for ablation and direct evaluation as an oracle-rationale formulation upper bound: it is analytically useful for isolating formulation quality under gold rationales, but it is not part of the deployed inference path. RGSC instead serves as the subsequent reinforcement-learning stage that calibrates the deployed hypothesis-generation policy.

The two-stage MSI curriculum is designed around two complementary considerations. First, it explicitly decouples the challenging cognitive step of causal reasoning from the syntactic precision required for symbolic formulation, improving both training stability and learning efficiency. Second, it enforces a hierarchical abstraction that moves from evidence to rationale and then to equation: the model first distills multimodal empirical evidence into a qualitative explanation via C-CoT, and only then translates this rationale into a well-formed symbolic law.

Together, MSI-1 and MSI-2 provide complementary supervised branches for later evaluation, with MSI-2 serving as an oracle-rationale analytical reference rather than as a deployable path. RGSC then calibrates the hypothesis-generation policy with structure-aware rewards.

3.3 Reward-Guided Symbolic Calibration (RGSC)

MSI yields a policy that produces *plausible* symbolic hypotheses, but it may still be insensitive to fine-grained structural correctness or may drift toward superficially reasonable (yet incorrect) laws. We therefore introduce **Reward-Guided Symbolic Calibration (RGSC)**, a reinforcement-learning stage that directly optimizes the hypothesis-generation policy with analytically computable, structure-aware rewards. Concretely, we adopt **Group Relative Policy Optimization (GRPO)** [28], which removes the need for a learned critic by using *within-group* relative rewards as a baseline. This is particularly suitable for symbolic law discovery, where rewards can be computed deterministically from the generated equation [28].

Group sampling of candidate laws. For each empirical evidence instance $\mathcal{E}$, we sample a group of G candidate governing equations $\{\hat{\mathcal{F}}_i\}_{i=1}^G$ from the current policy π_θ:

$$\hat{\mathcal{F}}_i \sim \pi_\theta(\cdot \mid \mathcal{E}), \qquad i \in \{1, \ldots, G\}. \tag{4}$$

Here π_θ is initialized from the MSI-1 model π_{MSI1}, and $\hat{\mathcal{F}}_i$ denotes the i-th sampled symbolic hypothesis for the same evidence $\mathcal{E}$.

Reward design. Each sampled hypothesis $\hat{\mathcal{F}}_i$ receives a scalar reward $R(\hat{\mathcal{F}}_i; \mathcal{E})$ that prioritizes (i) valid output protocol, (ii) parameter-agnostic structural correctness, and (iii) exact symbolic agreement when possible. Let $\mathcal{F}^\star$ denote the canonical ground-truth law. We use a weighted sum:

$$R(\hat{\mathcal{F}}_i; \mathcal{E}) = w_f R_{\text{format}}(\hat{\mathcal{F}}_i) + w_s R_{\text{struct}}(\hat{\mathcal{F}}_i, \mathcal{F}^\star) \quad + w_a R_{\text{acc}}(\hat{\mathcal{F}}_i, \mathcal{F}^\star). \tag{5}$$

(1) Format reward. This binary reward enforces strict adherence to the predefined `<think>...<answer>` protocol, ensuring the output is reliably parseable:

$$R_{\text{format}}(\hat{\mathcal{F}}_i) = \mathbb{I}\left[\hat{\mathcal{F}}_i \in \mathcal{P}\right], \tag{6}$$

where $\mathcal{P}$ denotes the set of outputs that match the `<think>...<answer>` protocol and admit deterministic extraction of the symbolic law.

(2) Parameter-agnostic structural reward. Our primary signal evaluates whether the model identifies the *correct term structure* independent of coefficient values or parameter names. We parse $\hat{\mathcal{F}}_i$ and $\mathcal{F}^\star$ into symbolic expressions, decompose each into an additive term set, and normalize each

term into a *skeleton* by replacing all non-core coefficients or parameters with signed units while preserving core variables (e.g., x, v, t) and operators. Denoting the resulting skeleton term sets by $\mathcal{T}(\hat{\mathcal{F}}_i)$ and $\mathcal{T}(\mathcal{F}^\star)$, we define

$$R_{\text{struct}}(\hat{\mathcal{F}}_i, \mathcal{F}^\star) = \frac{\left|\mathcal{T}(\hat{\mathcal{F}}_i) \cap \mathcal{T}(\mathcal{F}^\star)\right|}{\left|\mathcal{T}(\hat{\mathcal{F}}_i) \cup \mathcal{T}(\mathcal{F}^\star)\right|}. \tag{7}$$

This Jaccard score provides a dense signal in $[0, 1]$ that rewards *topological correctness* of the discovered law. This is the most critical component for our task, designed to assess the fundamental topological correctness of the posited physical law.

(3) Exact-match accuracy reward. To encourage full symbolic precision, we use a strict correctness signal:

$$R_{\text{acc}}(\hat{\mathcal{F}}_i, \mathcal{F}^\star) = \mathbb{I}\Big[\text{IsZero}\Big(\hat{\mathcal{F}}_i - \mathcal{F}^\star\Big)\Big], \tag{8}$$

where $\text{IsZero}(u)$ returns true if a CAS (e.g., SymPy) can verify $u \equiv 0$ after simplification.

Group-relative advantage (critic-free baseline). Given rewards $\{r_i\}_{i=1}^G$ for the group, GRPO computes *relative* advantages by normalizing within the group:

$$A_i = \frac{r_i - \mu}{\sigma + \epsilon},\ \mu = \frac{1}{G}\sum_{j=1}^{G} r_j,\ \sigma = \sqrt{\frac{1}{G}\sum_{j=1}^{G}(r_j - \mu)^2}. \tag{9}$$

For outcome supervision, we broadcast the same sequence-level advantage A_i to all tokens of the sampled hypothesis $\hat{\mathcal{F}}_i$, yielding a low-variance learning signal without training a value network.

GRPO objective with KL regularization. Let $\pi_{\theta_{\text{old}}}$ be the behavior policy used to sample $\hat{\mathcal{F}}_i$, and let π_{ref} be the frozen MSI-1 reference policy. GRPO updates π_θ by optimizing a PPO-style clipped surrogate using the group-relative advantage, while regularizing deviation from π_{ref} via KL:

$$\max_\theta\ \mathbb{E}_{\mathcal{E}}\left[\frac{1}{G}\sum_{i=1}^{G} A_i\ \min\Big(\rho_i(\theta),\ \tilde{\rho}_i(\theta)\Big) - \beta\,\text{KL}\Big(\pi_\theta(\cdot \mid \mathcal{E}) \,\|\, \pi_{\text{ref}}(\cdot \mid \mathcal{E})\Big)\right], \tag{10}$$

where $\rho_i(\theta) = \frac{\pi_\theta(\hat{\mathcal{F}}_i|\mathcal{E})}{\pi_{\theta_{\text{old}}}(\hat{\mathcal{F}}_i|\mathcal{E})}$, $\tilde{\rho}_i(\theta) = \text{clip}\big(\rho_i(\theta), 1-\tau, 1+\tau\big)$, τ is the clipping threshold, and β controls KL term. This objective anchors updates to the MSI-initialized reference policy π_{ref} via KL regularization, while promoting hypotheses that achieve higher within-group relative reward.

3.4 Agentic Refinement via Symbolic Residual Realignment (SR^2)

After RGSC, VIPER-R1 outputs a high-confidence candidate law $\hat{\mathcal{F}}_0$, which we interpret as an acceleration model

$$\hat{\mathcal{F}}_0: \quad a\ =\ \hat{f}_0(x, v, t). \tag{11}$$

While $\hat{f}_0$ typically captures the dominant *structure* of the dynamics, it can still exhibit systematic *numerical* mismatch due to imperfect coefficients, missing higher-order effects, or unmodeled forcing. In the final stage, we therefore optionally invoke an external symbolic regression (SR) solver [7] to correct the remaining discrepancy. We term this residual-refinement procedure **Symbolic Residual Realignment (SR^2)**, an external correction module that realigns a first-order hypothesis with empirical measurements.

SR^2 in three steps. Rather than rediscovering the full law from scratch, SR^2 constrains SR to model only the residual induced by $\hat{\mathcal{F}}_0$, which substantially reduces the search space and improves robustness.

Step 1: Residual construction. Given empirical evidence $\mathcal{E}$ with targets $a_{\text{GT}}(t)$, we evaluate $\hat{f}_0(x, v, t)$ on the observed states and form:

$$r(t)\ =\ a_{\text{GT}}(t)\ -\ \hat{f}_0\left(x(t), v(t), t\right). \tag{12}$$

Step 2: Symbolic regression on the residual. We fit a parsimonious symbolic correction $f_{\text{res}}(x, v, t)$ by running SR with inputs (x, v, t) and targets $r(t)$:

$$f_{\text{res}} \leftarrow \text{SR}\big((x, v, t),\, r(t)\big). \tag{13}$$

Step 3: Realigned law. We then compose the original hypothesis with the learned correction:

$$\hat{\mathcal{F}}_{\text{final}} :\ a = \hat{f}_{\text{final}}(x, v, t) = \hat{f}_0(x, v, t) + f_{\text{res}}(x, v, t). \tag{14}$$

The SR^2 procedure is summarized in Algorithm 2; additional implementation details are deferred to Appendix B.4.

3.5 The PhysSymbol Dataset

PhysSymbol contains 10,000 multimodal instances split into two parts. All training and quantitative evaluation in this paper use the *Mechanics* split, which covers low-dimensional systems over (x, v, t) with 2–5-term ODEs drawn from springs, damping, forcing, and simple control libraries. Each instance contains paired $x(t)$ and $v(x)$ plots, a motion table $M = \{(t_i, x_i, v_i, a_i)\}_{i=1}^{T}$, the canonical law $\mathcal{F}$, and a C-CoT trace that links visual motifs such as spirals, decay envelopes, or limit cycles to symbolic terms. Full construction details are deferred to Appendix D.

4 Experiments

4.1 Experimental Setup

Dataset. All main experiments use the PhysSymbol Mechanics split, a synthetic benchmark designed to reduce contamination risk and to match the paper's paired-plot protocol. Full construction details are deferred to Appendix D.

External benchmark scope. Among LLM-SRBench subsets, we evaluate `lsr_synth_phys_osc`, the subset that matches the paper's paired $x(t)$ and $v(x)$ evidence format.

Models and baselines. VIPER-R1-3B and VIPER-R1-7B are built on Qwen-VL-2.5 3B and 7B. We compare against PySR [7] and a broad set of strong VLM baselines: *GPT-5* [22], *GPT-5 mini* [22], *GPT-4o mini* [23], *GPT-4o* [24], *Grok 3* [38], *GPT-o3* [23], *Claude-4-Sonnet* [2], *Claude-3.7-Sonnet* [1], *Qwen-VL-Max* [3], *Qwen-VL-2.5-72B* [4], and *Gemini 2.5 Pro* [6].

Reporting and metrics. Unless noted otherwise, we report the *final checkpoint* of each completed training stage rather than selecting the best checkpoint on the test set. MSI2 is reported only as an oracle-rationale formulation upper bound, not as part of the deployed inference path. Our core metrics are S_{struct} (parameter-agnostic term-skeleton Jaccard), S_{acc} (exact symbolic-match rate), and Post-SR^2 MSE (end-to-end error after residual refinement). We use the same S_{acc} notation consistently in the modality and stage diagnostics alongside ROUGE-L. Further training and evaluation details are deferred to Appendix B.

4.2 Main Results and Analysis

Table 1 reports the full PhysSymbol comparison against eleven VLM baselines and PySR, and Figure 4 visualizes the same ranking. On this benchmark, VIPER-R1-7B delivers the best structural score ($S_{\text{struct}} = 0.812$) and the best learned-model Post-SR^2 MSE (0.032), while its S_{acc} value (0.487) is effectively tied with VIPER-R1-3B (0.488). VIPER-R1's core role is to produce a stronger first-order symbolic hypothesis: paired visual evidence plus structure-aware calibration improve symbolic structure induction in this focused low-dimensional regime, so S_{struct} and S_{acc} are the standalone evidence for the learned model, whereas Post-SR^2 MSE is reported separately as an end-to-end utility metric after constrained residual realignment. This pattern is informative rather than contradictory: scaling from 3B to 7B primarily improves topological structure induction rather than token-level exactness, suggesting that the additional capacity is spent on recovering the global symbolic skeleton once the formula family is already near the exact-match boundary. PySR remains numerically strong but structurally weak ($S_{\text{struct}} = 0.352$), reinforcing that low numeric error alone does not imply recovery of the correct governing law.

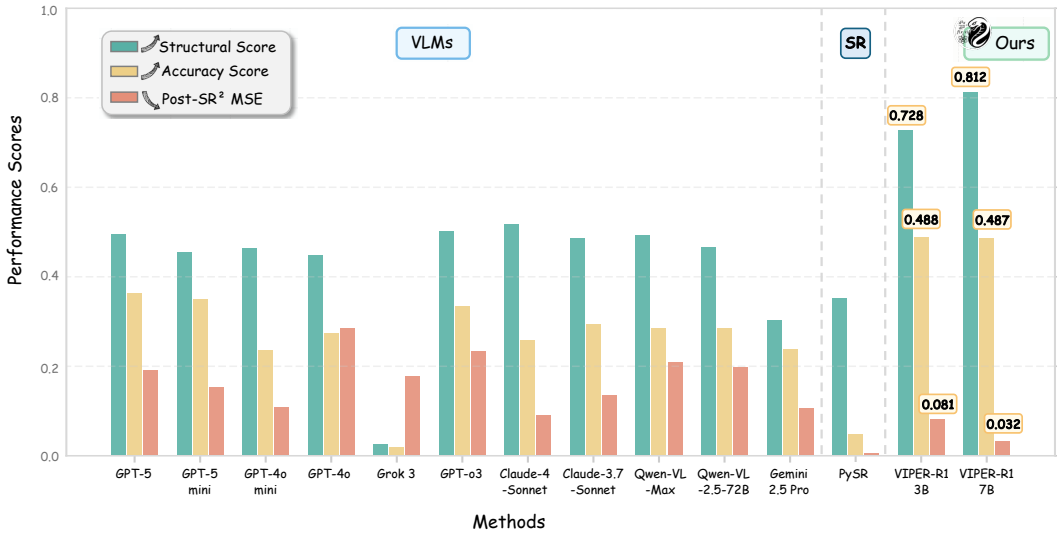


Figure 4: PhysSymbol comparison across the three main metrics.

Table 1: Full PhysSymbol Mechanics comparison. Point estimates are final-checkpoint results; the $\pm$ entries are archived descriptive variability summaries, not confidence intervals over independent retraining runs. Higher is better for S_{struct} and S_{acc}; lower is better for Post-SR^2 MSE.

Category	**Method**	$S_{\text{struct}} \uparrow$	$S_{\text{acc}} \uparrow$	**Post-SR^2 MSE** $\downarrow$
VLMs	GPT-5	0.494 ± 0.012	0.363 ± 0.015	0.192 ± 0.018
	GPT-5 mini	0.455 ± 0.010	0.350 ± 0.011	0.154 ± 0.013
	GPT-4o mini	0.463 ± 0.014	0.235 ± 0.009	0.109 ± 0.012
	GPT-4o	0.449 ± 0.011	0.274 ± 0.013	0.286 ± 0.025
	Grok 3	0.026 ± 0.005	0.019 ± 0.003	0.177 ± 0.020
	GPT-o3	0.502 ± 0.015	0.335 ± 0.014	0.234 ± 0.019
	Claude-4-Sonnet	0.518 ± 0.009	0.257 ± 0.010	0.091 ± 0.008
	Claude-3.7-Sonnet	0.485 ± 0.013	0.294 ± 0.012	0.136 ± 0.015
	Qwen-VL-Max	0.493 ± 0.016	0.285 ± 0.014	0.210 ± 0.022
	Qwen-VL-2.5-72B	0.466 ± 0.014	0.284 ± 0.015	0.198 ± 0.017
	Gemini 2.5 Pro	0.302 ± 0.018	0.237 ± 0.016	0.107 ± 0.011
SR-Only	PySR	0.352 ± 0.021	0.048 ± 0.025	$\mathbf{0.006 \pm 0.000}$
Ours	VIPER-R1-3B	0.728 ± 0.006	0.488 ± 0.007	0.081 ± 0.005
	VIPER-R1-7B	$\mathbf{0.812 \pm 0.004}$	$\mathbf{0.487 \pm 0.005}$	0.032 ± 0.003

VIPER-R1-3B also outperforms the larger general-purpose VLM baselines, which suggests that the gain comes from the staged training recipe rather than scale alone. The picture is therefore consistent across all three metrics: visual grounding improves symbolic topology recovery, and a stronger first-order symbolic hypothesis gives the downstream residual-refinement stage a materially better residual space to solve.

4.3 Focused Ablations and Transfer

Modality value on PhysSymbol. Table 2 (left) compares the final 7B RGSC checkpoint under four input settings. Removing visual evidence causes a large drop in symbolic quality: the full multimodal and image-only settings both remain above 0.80 on S_{struct}, whereas trajectory-only and PySR numeric-only fall to 0.3701 and 0.352. Notably, image-only attains the highest structural score (0.8593), indicating that the paired plots already encode most topology cues for the current prompt on this in-domain split. Full multimodal, however, achieves the highest exact-match accuracy (0.4870 vs. 0.4565), suggesting that the added trajectory table helps convert visually plausible structures into fully correct formulas. We therefore keep full multimodal as the paper's default deployment condition: on the matched official external subset, it also has higher point estimates than image-only on both S_{struct} and S_{acc}, although the bootstrap intervals overlap (Table 3).

Table 2: PhysSymbol diagnostic comparisons. Left: modality comparison on the Mechanics test split with the final 7B RGSC checkpoint, using pre-SR^2 formula-generation metrics. Right: stage ablation. MSI2 is shown only as an oracle-rationale formulation upper bound and is not part of the deployed inference path. Higher is better for S_{struct}, S_{acc}, and ROUGE-L.

(a) Modality Ablation

Variant	$S_{\text{struct}} \uparrow$	$S_{\text{acc}} \uparrow$	**ROUGE-L**$\uparrow$
Full multimodal	0.8120	0.4870	0.6184
Image-only	0.8593	0.4565	0.6185
Trajectory-only	0.3701	0.3166	0.2675
PySR numeric-only	0.352	0.048	0.2674

(b) Stage Ablation

Variant	$S_{\text{struct}} \uparrow$	$S_{\text{acc}} \uparrow$	**ROUGE-L**$\uparrow$
Base	0.2060	0.0861	0.1450
MSI1	0.8028	0.4220	0.5968
$\text{MSI2}_{(\text{cond.})}$	1.0000	0.9713	0.9834
RGSC	0.8120	0.4870	0.6184

Table 3: Focused external transfer on the official LLM-SRBench subset `lsr_synth_phys_osc`, which matches the paper's paired $x(t)$ and $v(x)$ evidence protocol. We report bootstrap 95% intervals for S_{struct} and S_{acc} over the 44 adapted examples. Higher is better for both metrics.

Variant	$S_{\text{struct}} \uparrow$	$S_{\text{acc}} \uparrow$
Full multimodal	0.4414 [0.3960, 0.4865]	0.3385 [0.3191, 0.3588]
Image-only	0.4311 [0.3785, 0.4848]	0.3201 [0.2938, 0.3511]
Trajectory-only	0.3471 [0.3013, 0.3948]	0.2632 [0.2321, 0.2969]
PySR numeric-only	0.4220 [0.3551, 0.4930]	0.3187 [0.2873, 0.3527]

Stage contribution ablation. Table 2 (right) separates the deployed RGSC path from the two supervised MSI branches. MSI1 provides the dominant unconditional gain, lifting S_{struct} from 0.2060 to 0.8028, and RGSC further improves it to 0.8120 under the same deployment-time input format. MSI2 is shown only as an oracle-rationale formulation upper bound, so its near-saturated scores are analytically useful for isolating formulation quality under gold rationales but are not part of the deployed inference path.

Focused external transfer. We next evaluate the final RGSC checkpoint, without retraining, on the official LLM-SRBench [30] subset `lsr_synth_phys_osc`, which matches the paper's paired $x(t)$ and $v(x)$ evidence protocol. VIPER-R1 transfers directly to this external setting, where the full multimodal condition remains the strongest symbolic point-estimate configuration, with the highest point estimates on S_{struct} and S_{acc} among the four input conditions (Table 3). Its bootstrap 95% intervals over the 44 adapted examples overlap with those of image-only and PySR, so PySR remains a strong classical baseline on this narrow oscillator subset. Noise robustness, fine-tuned VLM baselines, PySR search-budget sensitivity, and a small real-data probe are reported in Appendix C.

Qualitative example. Figure 5 shows a driven oscillator whose law contains a coupled term $-0.592\,xv$. In its rationale, VIPER-R1 attributes the sustained oscillation in $x(t)$ to restoring and driving forces, the decay in amplitude to linear damping, and the spiral shape of $v(x)$ to a nonlinear interaction, which it proposes as $-vx$. The pre-tool hypothesis contains exactly the four ground-truth term families, and every coefficient is within 1.1% of its true value. Appendix E reports further cases.

4.4 Discussion

The experiments support a consistent interpretation of VIPER-R1: its main advantage is not that it replaces numerical symbolic regression, but that it changes the *starting point* of symbolic search. Instead of asking a solver to recover the full governing law from numerical trajectories alone, VIPER-R1 first uses paired phase-space and trajectory evidence to form a stronger first-order symbolic hypothesis. This matters in low-dimensional dynamics because the paired plots expose qualitative structure, such as damping, restoring forces, forcing, coupling, and nonlinear deformation, that is often difficult to infer robustly from finite trajectory tables alone.

The modality results support the same interpretation. Image-only can peak on in-domain S_{struct} because paired plots already encode most coarse topology cues, while full multimodal remains the preferred deployment condition because the trajectory table helps resolve residual ambiguity when producing a fully specified law.

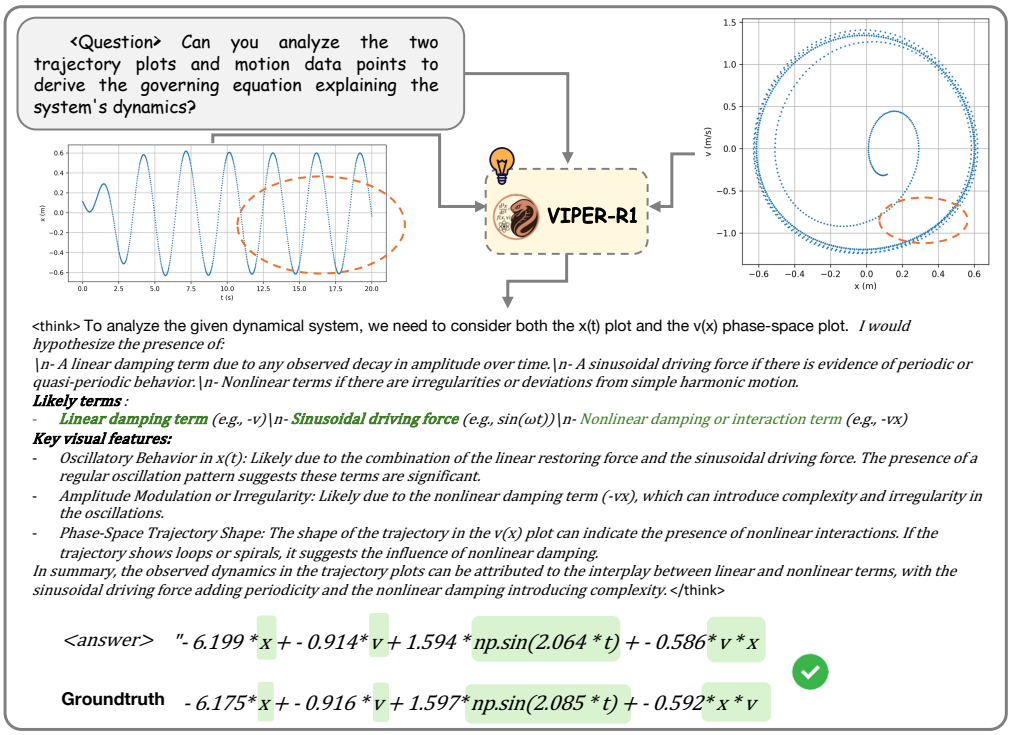


Figure 5: Qualitative example: driven oscillator with interaction damping. Top: inputs and VIPER-R1 rationale. Bottom: pre-tool hypothesis and ground truth. SR^2 diagnostics are in Figure 12.

This view clarifies the roles of MSI and RGSC. MSI provides the dominant gain by learning an evidence-to-rationale-to-equation mapping, converting visual and numerical observations into a physically grounded symbolic hypothesis. RGSC then serves as a targeted calibration stage: rather than relearning this mapping, it sharpens the deployed policy toward term-topology correctness under the same test-time interface. Thus, the smaller MSI→RGSC gap relative to Base→MSI is expected, reflecting a staged design in which supervised learning builds the visual-symbolic prior and reinforcement learning calibrates the distribution of plausible hypotheses. Finally, the metric breakdown highlights VIPER-R1's core advantage before tool use. S_{struct} and S_{acc} evaluate the learned pre-tool symbolic hypothesis, whereas Post-SR^2 MSE measures how effectively that hypothesis seeds constrained residual realignment. PySR's low MSE shows that flexible numerical fitting can reduce error without reliably recovering term topology; VIPER-R1 instead provides a visually grounded candidate law that makes downstream correction better posed and more interpretable. This conclusion is further supported by the external-transfer results (Table 3), the fine-tuned baselines and PySR budget analyses in Appendix C, and the boundary-of-evidence discussion in Appendix G.

5 Conclusion

We introduced **VIPER-R1**, a multimodal framework for visual formula discovery from low-dimensional dynamical systems using paired trajectory and phase-space evidence. Across PhysSymbol, paired visual evidence plus structure-aware calibration improve symbolic structure induction, yielding stronger first-order hypotheses for downstream refinement than the base model and trajectory-only formulations. These gains translate into the strongest learned-model structural performance in our main comparison and provide a better starting point for residual realignment. More broadly, the paper shows that the main advantage over classical SR is not universal backend dominance, but better first-order hypothesis formation before downstream numerical refinement. In this regime, visual grounding and structure-aware calibration play complementary roles: one constrains the plausible law family, and the other sharpens the deployed symbolic hypothesis. Beyond the in-domain benchmark, VIPER-R1 also transfers without retraining to the matched official external subset, suggesting that the learned visual-symbolic inductive bias extends beyond a single dataset family.

Acknowledgments and Disclosure of Funding

This work is supported by Shanghai Artificial Intelligence Laboratory and the 2026 New Generation Artificial Intelligence-National Science and Technology Major Project No. 3004.

References

[1] Anthropic. Claude 3.7 sonnet and claude code. `https://www.anthropic.com/news/claude-3-7-sonnet`, 2025.

[2] Anthropic. Introducing claude 4. `https://www.anthropic.com/news/claude-4`, 2025.

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[5] Luca Biggio, Tommaso Bendinelli, Alexander Neitz, Aurelien Lucchi, and Giambattista Parascandolo. Neural symbolic regression that scales. In *International Conference on Machine Learning*, pages 936–945. Pmlr, 2021.

[6] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilaï Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith Sheshan, Raia Hadsell, Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan Horgan, Alvin Abdagic, Lior Belenki, James Allingham, Anima Singh, Theo Guidroz, Srivatsan Srinivasan, Herman Schmit, Kristen Chiafullo, Andre Elisseeff, Nilpa Jha, Prateek Kolhar, Leonard Berrada, Frank Ding, Xiance Si, Shrestha Basu Mallick, Franz Och, Sofia Erell, Eric Ni, Tejasi Latkar, Sherry Yang, Petar Sirkovic, Ziqiang Feng, Robert Leland, Rachel Hornung, Gang Wu, Charles Blundell, Hamidreza Alvari, Po-Sen Huang, Cathy Yip, Sanja Deur, Li Liu, Gabriela Surita, Pablo Duque, Dima Damen, Johnson Jia, Arthur Guez, Markus Mircea, Animesh Sinha, Alberto Magni, Paweł Stradomski, Tal Marian, Vlado Galić, Wenhu Chen, Hisham Husain, Achintya Singhal, Dominik Grewe, François-Xavier Aubet, Shuang Song, Lorenzo Blanco, Leland Rechis, Lewis Ho, Rich Munoz, Kelvin Zheng, Jessica Hamrick, Kevin Mather, Hagai Taitelbaum, Eliza Rutherford, Yun Lei, Kuangyuan Chen, Anand Shukla, Erica Moreira, Eric Doi, Berivan Isik, Nir Shabat, Dominika Rogozińska, Kashyap Kolipaka, Jason Chang, Eugen Vušak, Srinivasan Venkatachary, Shadi Noghabi, Tarun Bharti, Younghoon Jun, Aleksandr Zaks, Simon Green, Jeshwanth Challagundla, William Wong, Muqthar Mohammad, Dean Hirsch, Yong Cheng, Iftekhar Naim, Lev Proleev, Damien Vincent, Aayush Singh, Maxim Krikun, Dilip Krishnan, Zoubin Ghahramani, Aviel Atias, Rajeev Aggarwal, Christo Kirov, Dimitrios Vytiniotis, Christy Koh, Alexandra Chronopoulou, Pawan Dogra, Vlad-Doru Ion, Gladys Tyen, Jason Lee, Felix Weissenberger, Trevor Strohman, Ashwin Balakrishna, Jack Rae, Marko Velic, Raoul de Liedekerke, Oded Elyada, Wentao Yuan, Canoee Liu, Lior Shani, Sergey Kishchenko, Bea Alessio, Yandong Li, Richard Song, Sam Kwei, Orion Jankowski, Aneesh Pappu, Youhei Namiki, Yenai Ma, Nilesh Tripuraneni, Colin Cherry, Marissa Ikonomidis, Yu-Cheng Ling, Colin Ji, Beka Westberg, Auriel Wright, Da Yu, David Parkinson, Swaroop Ramaswamy, Jerome Connor, Soheil Hassas Yeganeh, Snchit Grover, George Kenwright, Lubo Litchev, Chris Apps, Alex Tomala, Felix Halim, Alex Castro-Ros, Zefei Li, Anudhyan Boral, Pauline Sho, Michal Yarom, Eric Malmi, David Klinghoffer, Rebecca Lin, Alan Ansell, Pradeep Kumar S, Shubin Zhao, Siqi Zuo, Adam Santoro, Heng-Tze Cheng, Solomon Demmessie, Yuchi Liu, and Nicole Brichtova. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv e-prints*, page arXiv:2507.06261, July 2025.

[7] Miles Cranmer. Interpretable machine learning for science with PySR and symbolicregression.jl, May 2023.

[8] Arya Grayeli, Atharva Sehgal, Omar Costilla-Reyes, Miles Cranmer, and Swarat Chaudhuri. Symbolic regression with a learned concept library. *Advances in Neural Information Processing Systems*, 37:44678–44709, 2024.

[9] Sam Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian Neural Networks. *arXiv e-prints*, page arXiv:1906.01563, June 2019.

[10] Ming Hu, Chenglong Ma, Wei Li, Wanghan Xu, Jiamin Wu, Jucheng Hu, Tianbin Li, Guohang Zhuang, Jiaqi Liu, Yingzhou Lu, et al. A survey of scientific large language models: From data foundations to agent frontiers. *arXiv preprint arXiv:2508.21148*, 2025.

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

[12] John R Koza. Genetic programming as a means for programming computers by natural selection. *Statistics and computing*, 4:87–112, 1994.

[13] William La Cava, Patryk Orzechowski, Bogdan Burlacu, Fabrício Olivetti de França, Marco Virgolin, Ying Jin, Michael Kommenda, and Jason H Moore. Contemporary symbolic regression methods and their relative performance. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[14] Ruikun Li, Yan Lu, Shixiang Tang, Biqing Qi, and Wanli Ouyang. Mllm-based discovery of intrinsic coordinates and governing equations from high-dimensional data, May 2025.

[15] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

[16] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, December 2022.

[17] Pingchuan Ma, Tsun-Hsuan Wang, Minghao Guo, Zhiqing Sun, Joshua B. Tenenbaum, Daniela Rus, Chuang Gan, and Wojciech Matusik. LLM and simulation as bilevel optimizers: A new paradigm to advance physical scientific discovery. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 33940–33962, 21–27 Jul 2024.

[18] 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.

[19] Matteo Merler, Katsiaryna Haitsiukevich, Nicola Dainese, and Pekka Marttinen. In-context symbolic regression: leveraging language models for function discovery. *arXiv preprint arXiv:2404.19094*, 2024.

[20] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024.

[21] T Nathan Mundhenk, Mikel Landajuela, Ruben Glatt, Claudio P Santiago, Daniel M Faissol, and Brenden K Petersen. Symbolic regression via deep reinforcement learning enhanced genetic programming seeding. In *Advances in Neural Information Processing Systems*, pages 24912–24923, 2021.

[22] OpenAI. Introducing gpt-5. `https://openai.com/index/introducing-gpt-5/`, 2025.

[23] OpenAI. Introducing openai o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/, 2025.

[24] OpenAI:, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Ramesh, et al. GPT-4o System Card. *arXiv e-prints*, October 2024.

[25] Brenden K Petersen, Mikel Landajuela, T Nathan Mundhenk, Claudio P Santiago, Soo K Kim, and Joanne T Kim. Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. *arXiv preprint arXiv:1912.04871*, 2019.

[26] Chandan K. Reddy and Parshin Shojaee. Towards scientific discovery with generative ai: Progress, opportunities, and challenges. In *AAAI Conference on Artificial Intelligence*, 2025.

[27] Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.

[28] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[29] 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.

[30] Parshin Shojaee, Ngoc-Hieu Nguyen, Kazem Meidani, Amir Barati Farimani, Khoa D Doan, and Chandan K. Reddy. LLM-SRBench: A new benchmark for scientific equation discovery with large language models. In *Forty-second International Conference on Machine Learning*, 2025.

[31] Steven H Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Studies in Nonlinearity. Westview Press, 2001.

[32] Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025.

[33] Fangzheng Sun, Yang Liu, Jian-Xun Wang, and Hao Sun. Symbolic physics learner: Discovering governing equations via monte carlo tree search. In *The Eleventh International Conference on Learning Representations*, 2023.

[34] Silviu-Marian Udrescu and Max Tegmark. Ai feynman: A physics-inspired method for symbolic regression, April 2020.

[35] Marco Virgolin and Solon P. Pissis. Symbolic regression is NP-hard. *Transactions on Machine Learning Research*, 2022.

[36] Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

[37] Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862. Association for Computational Linguistics, 2024.

[38] xAI. Grok 3 beta — the age of reasoning agents. https://x.ai/news/grok-3, 2025.

[39] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[40] Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. A comprehensive survey of scientific large language models and their applications in scientific discovery. *arXiv preprint arXiv:2406.10833*, 2024.

Table 4: MSI supervised fine-tuning hyperparameters.

Hyperparameter	Setting
Optimizer	AdamW
Learning rate	2×10^{-5}
Warmup steps	300
Batch size	64
Gradient clipping	1.0
Weight decay	0.1
Max sequence length	4096 tokens
Epochs	3
Loss weights	$L_{\text{eq}} : L_{\text{cot}} : L_{\text{vis}} = 1.0 : 0.3 : 0.1$

A Appendix

This supplementary material provides additional details on the proposed method and experimental results that could not be included in the main manuscript due to page limitations. Specifically, this appendix is organized as follows.

- Sec. B describes the training setup, model configurations, and additional evaluation details, offering comprehensive experimental specifications.
- Sec. C presents supplementary experiments, including noise robustness, comparisons with fine-tuned VLM baselines, PySR search-budget sensitivity, and a small real-data transfer probe.
- Sec. D provides more details about PhysSymbol and discusses how we collected, filtered, and reconstructed a high-quality dataset.
- Sec. E includes more visualization cases.
- Sec. F gives the impact statement, and Sec. G summarizes the main limitations and boundary-of-evidence considerations.

B Details of Training and Evaluation

B.1 Training Settings and Details

This subsection provides complete implementation details for the three core components of VIPER-R1: MSI, RGSC, and Symbolic Residual Realignment.

Main-table uncertainty reporting. Under the `strict_comparable` evaluation used for local checkpoints, generation is run deterministically, while closed-source APIs are queried through the same prompt-and-scoring pipeline with provider-side variability still possible. Accordingly, the paper's learned-model claims are based on the final-checkpoint point estimates, while Table 1 in the main text reports the full comparison with companion $\pm$ entries reported only as descriptive variability summaries from the archived comparison records; they should not be interpreted as confidence intervals over independent retraining runs.

MSI Training Configuration. MSI fine-tunes the base VLM using supervised learning to induce visually grounded symbolic structures. Training employs the AdamW optimizer with a learning rate of 2×10^{-5}, cosine annealing, and a warmup duration of the first 300 optimization steps. A batch size of 64 is used throughout, with gradient clipping at 1.0 and weight decay set to 0.1. The model processes sequences up to 4096 tokens and is trained for three epochs over PhysSymbol. The total loss combines three components: an equation-token loss that teaches the symbolic structure, a causal CoT generation loss that preserves interpretable reasoning, and a visual-alignment loss that encourages consistency between plot features and symbolic predictions. These losses are mixed with weights 1.0, 0.3, and 0.1, respectively.

Reward-Guided Symbolic Calibration (RGSC). RGSC refines the hypothesis policy via GRPO. For each evidence $\mathcal{E}$ in a batch, we sample a group of G candidate laws $\{\hat{\mathcal{F}}_i\}_{i=1}^{G}$ from $\pi_\theta(\cdot \mid \mathcal{E})$.

Table 5: RGSC reinforcement learning hyperparameters.

Hyperparameter	Setting
Rollout length	4
Samples per instance	6
Batch size	32 (192 hypotheses)
Learning rate	5×10^{-6}
Optimizer	AdamW
KL coefficient β	0.02
Advantage normalization	Within each instance group ($G = 6$)
Reward weights	$w_s = 5,\ w_f = 1,\ w_a = 1$

Table 6: Training Computational Cost Analysis

Component	Hardware	Duration	GPU-hours
MSI	8 × A800	4 hours	32
RGSC	8 × A800	42 hours	336
Total	-	-	**368**

Each $\hat{\mathcal{F}}_i$ is scored by a weighted reward

$$R(\hat{\mathcal{F}}_i) = w_f R_{\text{format}}(\hat{\mathcal{F}}_i) + w_s R_{\text{struct}}(\hat{\mathcal{F}}_i, \mathcal{F}^\star) + w_a R_{\text{acc}}(\hat{\mathcal{F}}_i, \mathcal{F}^\star), \tag{15}$$

where R_{struct} is the parameter-agnostic term-skeleton Jaccard similarity (see Sec. B.3). We normalize rewards within each group to obtain relative advantages and update π_θ with a KL regularizer toward the MSI reference policy π_{ref}.

SR^2 (Symbolic Residual Realignment). Let the post-RGSC hypothesis be $\hat{\mathcal{F}}_0 : a = \hat{f}_0(x, v, t)$. SR^2 fits the residual acceleration

$$r(t) = a_{\text{GT}}(t) - \hat{f}_0(x(t), v(t), t), \tag{16}$$

via symbolic regression, yielding a correction $f_{\text{res}}(x, v, t)$ and the final law $\hat{\mathcal{F}}_{\text{final}} : a = \hat{f}_0(x, v, t) + f_{\text{res}}(x, v, t)$.

Training Compute. All training runs were performed on NVIDIA A800 (80GB) GPUs. As shown in Table 6, MSI requires only a short supervised phase, while RGSC, our reinforcement-learning based symbolic calibration stage—constitutes the majority of the compute due to multi-sample rollout generation and reward evaluation. Despite this, the overall training cost remains modest compared to typical large-model finetuning workloads, as VIPER-R1 focuses on targeted structural refinement rather than end-to-end large-scale retraining.

B.2 Training and Evaluation Prompts

The behavior and reasoning process of VIPER-R1 are carefully guided by a series of structured prompts tailored to each stage of our training and evaluation pipeline. These prompts define the model's role, the available evidence, and the expected output format. This structured approach is crucial for decoupling complex tasks, progressively building the model's capability for scientific discovery, and keeping the evaluation protocol parsable and reproducible. Below, we detail the main prompts used across training and deployment.

B.2.1 Prompt for MSI step 1

In the initial MSI step, as shown in Figure 6, the goal is to teach the model to perform end-to-end reasoning, connecting raw visual phenomena directly to a final governing equation. The prompt instructs the model to act as a scientific assistant, verbalize its step-by-step analysis, and provide a conclusive answer in a structured format.

System Prompt

```
You are a helpful scientific assistant.  Given trajectory images and motion data from
a physical system, reason step-by-step to explain the observed behavior, then output
the governing equation.  Wrap your reasoning process in <think> </think> and your final
equation in <answer> </answer>.
```

Figure 6: System prompt for the first SFT stage (MSI).

B.2.2 Prompt for MSI step 2

In the second step of MSI, as shown in Figure 7, we decouple the task: the model is provided with the pre-computed reasoning chain (C-CoT) and is tasked only with translating this analysis into a precise symbolic equation. This prompt focuses the model's training on the final, crucial step of symbolic formulation.

System Prompt

```
You are a helpful scientific assistant.  Given the reasoning steps for a physical system
and its trajectory images, output the corresponding governing equation.  The reasoning is
provided in <think> </think> tags, and your answer should be placed in <answer> </answer>
tags.
```

Figure 7: System prompt for the second SFT stage (CoT-Aware).

B.2.3 Prompt for RGSC

During the reinforcement learning phase, as shown in Figure 8, the prompt is refined to encourage more abstract and generalized symbolic reasoning. It explicitly asks the model to use symbolic placeholders for unknown parameters, which is essential for discovering general physical laws rather than fitting to specific numerical instances. This prompt guides the generation of hypotheses that are then evaluated by our reward function.

B.2.4 Prompt for Local Evaluation

For the paper's learned-model rows, local checkpoint evaluation under `strict_comparable` uses the user-side prompt shown in Figure 9. This deployment-time prompt reuses the canonical benchmark instruction, appends a strict answer-only suffix, and passes the paired phase-space and trajectory images alongside the text in the multimodal message. The `image-only` and `trajectory-only` ablations are derived from the same base template by removing the motion-data block or the plot references, rather than by introducing separate prompt families. OpenAI-compatible API evaluations may additionally add an answer-only system prompt for formatting enforcement, but that wrapper is API-specific and is not the canonical local checkpoint path used for the paper's learned-model rows.

B.3 Detailed Reward Function Formulation for RGSC

The total reward signal $R(\hat{\mathcal{F}}_i)$ used during the Reward-Guided Symbolic Calibration (RGSC) stage is a weighted sum of three distinct components. Each component is designed to evaluate a specific aspect of the generated Symbolic Ansatz S_i, allowing for a balanced and effective policy optimization. The composite reward is defined as:

$$R(\hat{\mathcal{F}}_i) = w_f R_{\text{format}}(\hat{\mathcal{F}}_i) + w_s R_{\text{struct}}(\hat{\mathcal{F}}_i, \mathcal{F}^\star) + w_a R_{\text{acc}}(\hat{\mathcal{F}}_i, \mathcal{F}^\star), \tag{17}$$

where $\mathcal{F}^\star$ is the Canonical Governing Equation, and w_f, w_s, w_a are hyperparameters that weight the contribution of each reward component. Below, we detail the formulation of each component.

Format Reward (R_{format}). The primary purpose of this reward is to enforce a consistent and parsable output structure, which is crucial for both interpretability and automated evaluation. We use regular expressions to verify that the model's output strictly adheres to our predefined template,

System Prompt

```
The user provides visual and trajectory data of a physical phenomenon. The Assistant's
task is to act as a physicist. First, think step-by-step about the underlying physical
principles in <think> tags. Then, derive and state the final governing equation in
<answer> tags. The equation should use symbolic placeholders for unknown parameters (e.g.,
k, c, F) and standard variables for the system (x, v, t).
```

Figure 8: System prompt for the RGSC stage.

Evaluation Prompt

```
Please analyze the following two trajectory plots and the associated motion data points.
Your task is to derive the governing equation that best explains the dynamics of the
system.

Then, output the final derived equation, and enclose your answer within <answer> and
</answer> tags.

Format your response as follows: <answer>[The final equation here]</answer>

Here is the data: t,x,v,a
[instance-specific numeric table follows here]

FINAL: Reply with exactly ONE line: <answer>...</answer>. Do NOT output any other text.
```

Figure 9: Representative evaluation-time user prompt for local checkpoint inference under `strict_comparable`. In the full multimodal setting, the paired phase-space and trajectory images are attached alongside this text prompt.

which requires a reasoning process enclosed within `<think>...</think>` tags followed by a final symbolic formula within `<answer>...</answer>` tags. This is a binary reward:

$$R_{\text{format}}(\hat{\mathcal{F}}_i) = \begin{cases} 1 & \text{if format is correct} \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

Parameter-Agnostic Structural Reward (R_{struct}). This is the core reward signal in RGSC, designed to measure the *topological* correctness of a candidate law while remaining invariant to coefficient values and parameter names. Given empirical evidence $\mathcal{E}$, the policy samples candidate equations $\{\hat{\mathcal{F}}_i\}_{i=1}^{G}$. For a particular candidate $\hat{\mathcal{F}}_i$ and the ground-truth law $\mathcal{F}_{\text{GT}}$, we compute a structural similarity score that rewards recovering the correct set of basis terms (e.g., restoring and damping components) irrespective of whether coefficients are written as (k, c) or instantiated numerically.

Concretely, we (i) parse the `<answer>` equation from $\hat{\mathcal{F}}_i$ and $\mathcal{F}^\star$ into symbolic expressions and decompose each into a set of constituent terms (additive terms if the expression is a sum; otherwise the expression itself), and (ii) normalize each term into a *structural skeleton* by replacing all numeric constants and symbolic parameters with a signed unit (± 1), while preserving the physical variables (x, v, t) and the mathematical operators. Let $\mathcal{T}(\hat{\mathcal{F}}_i)$ and $\mathcal{T}(\mathcal{F}^\star)$ denote the resulting sets of skeletonized terms for the candidate and ground truth, respectively. The reward is the Jaccard similarity:

$$R_{\text{struct}}(\hat{\mathcal{F}}_i, \mathcal{F}^\star) = \frac{\left|\mathcal{T}(\hat{\mathcal{F}}_i) \cap \mathcal{T}(\mathcal{F}^\star)\right|}{\left|\mathcal{T}(\hat{\mathcal{F}}_i) \cup \mathcal{T}(\mathcal{F}^\star)\right|} \in [0, 1]. \tag{19}$$

Exact-Match Accuracy Reward (R_{acc}). This component provides a strict correctness check, awarding a binary reward only when the candidate equation is mathematically identical to the ground-truth law. We parse the `<answer>` equation from $\hat{\mathcal{F}}_i$ and $\mathcal{F}^\star$ into symbolic expressions and verify

equivalence via `sympy`:

$$R_{\text{acc}}(\hat{\mathcal{F}}_i, \mathcal{F}^\star) = \begin{cases} 1 & \text{if } f_{\text{sc}}(\hat{\mathcal{F}}_i, \mathcal{F}^\star) = 0, \\ 0 & \text{otherwise,} \end{cases} \tag{20}$$

where $f_{\text{sc}}(\hat{\mathcal{F}}_i, \mathcal{F}^\star) = \texttt{sympy.simplify}\Big(\hat{\mathcal{F}}_i - \mathcal{F}^\star\Big)$.

By combining R_{format}, R_{struct}, and R_{acc}, RGSC shapes an optimization landscape that prioritizes recovering the correct physical structure (typically via a dominant weight w_s), while still enforcing a consistent output format and encouraging exact symbolic agreement when attainable.

B.4 Algorithm Pseudocode

To provide a comprehensive and reproducible overview of our methodology, we present the detailed pseudocode for our framework's two primary components: the end-to-end training process and the inference-time refinement procedure.

Algorithm 1 outlines the complete training framework for the VIPER-R1. This algorithm details the two primary phases through which the model is forged: first, the supervised Motion Structure Induction curriculum, which teaches the model to form hypotheses from visual data; and second, the subsequent reinforcement learning phase of Reward-Guided Symbolic Calibration, which purifies the model's symbolic generation policy.

Following this, Algorithm 2 describes the inference procedure. This algorithm formalizes the Agentic Refinement via Symbolic Residual Realignment process, wherein the fully trained VIPER-R1 generates an initial hypothesis and then proactively invokes an external symbolic regression tool to produce a final, empirically-realigned physical law.

B.5 Evaluation Metrics

To provide a holistic assessment of our framework's performance, we employ a suite of distinct metrics, each designed to capture a different facet of success, from high-level structural correctness to final empirical accuracy.

S_{struct} **(Structural Score).** This is our primary metric for evaluating the core capability of the VIPER-R1: its ability to generate a topologically correct symbolic hypothesis. This score measures the structural similarity between the generated formula and the canonical equation, intentionally ignoring numerical coefficients to focus purely on the underlying physical structure.

S_{acc} **(Exact-Match Rate).** To measure the exactness of the generated formulas, we use a strict symbolic accuracy score. This metric evaluates whether the generated formula is mathematically identical to the canonical equation. It serves as a challenging measure of the model's ultimate precision. In the modality and stage diagnostics in the main text, we report the same exact-match metric directly as S_{acc} for consistency with the main comparison tables.

Post-SR^2 Mean Squared Error (MSE). This metric evaluates the end-to-end performance of the entire VIPER-R1 framework by measuring how well the final, refined formula fits the observed data. It quantifies the empirical accuracy after the **SR^2** stage has been completed.

Calculation: Let S_{final} be the final symbolic law produced by our framework. This expression is converted into a callable function $a_{\text{final}}(x, v, t)$. The MSE is then computed over the N data points in the test set's trajectory data:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} \left(a_{\text{GT}}(t_i) - a_{\text{final}}(x_i, v_i, t_i)\right)^2, \tag{21}$$

where $a_{\text{GT}}(t_i)$ is the ground-truth acceleration at time t_i. A lower MSE indicates a better fit to the observed physical reality and thus a more successful discovery.

Algorithm 1 VIPER-R1 Training: Motion Structure Induction (MSI) and Reward-Guided Symbolic Calibration (RGSC)

Inputs: PhysSymbol corpus $\mathcal{D}_{\mathrm{PS}}$; pretrained VIPER-R1 policy parameters θ_0; MSI steps $N_{\text{MSI-1}}$, $N_{\text{MSI-2}}$;
RGSC steps N_{RGSC}; group size G; reward weights w_f, w_s, w_a; GRPO hyperparameters β, τ, ϵ.
Outputs: Calibrated VIPER-R1 policy π_θ.

Initialize VIPER-R1 policy $\pi_\theta \leftarrow \mathrm{InitializeModel}(\theta_0)$
Stage 1: Motion Structure Induction (MSI)
MSI-1: Joint Causal Reasoning and Symbolic Hypothesis
for $i = 1$ to $N_{\text{MSI-1}}$ **do**
  Sample $(\mathcal{E}, Y) \sim \mathcal{D}_{\mathrm{PS}}$ with $Y = (C, \mathcal{F})$
  Update θ by minimizing $\mathcal{L}_{\text{MSI-1}}(\theta)$ (Eq. (2))
end for
Save MSI-1 checkpoint $\pi_{\mathrm{MSI1}} \leftarrow \pi_\theta$
MSI-2: C-CoT-Guided Symbolic Formulation
for $i = 1$ to $N_{\text{MSI-2}}$ **do**
  Sample $(\mathcal{E}, C, \mathcal{F}) \sim \mathcal{D}_{\mathrm{PS}}$
  Update θ by minimizing $\mathcal{L}_{\text{MSI-2}}(\theta)$ (Eq. (3))
end for
Save MSI-2 supervised branch $\pi_{\mathrm{MSI2}} \leftarrow \pi_\theta$ (kept only for ablation / oracle upper-bound analysis; not used in the deployed inference path)
Restore the branch used for RGSC training: $\pi_\theta \leftarrow \pi_{\mathrm{MSI1}}$
Freeze MSI-1 reference policy $\pi_{\mathrm{ref}} \leftarrow \pi_{\mathrm{MSI1}}$
Stage 2: Reward-Guided Symbolic Calibration (RGSC)
for $k = 1$ to N_{RGSC} **do**
  Sample empirical evidence instance $(\mathcal{E}, \mathcal{F}^\star) \sim \mathcal{D}_{\mathrm{PS}}$
  Set behavior policy $\pi_{\theta_{\mathrm{old}}} \leftarrow \pi_\theta$
  for $i = 1$ to G **do**
    Sample candidate law $\hat{\mathcal{F}}_i \sim \pi_{\theta_{\mathrm{old}}}(\cdot \mid \mathcal{E})$ (Eq. (4))
    Compute reward
      $r_i \leftarrow w_f R_{\mathrm{format}}(\hat{\mathcal{F}}_i) + w_s R_{\mathrm{struct}}(\hat{\mathcal{F}}_i, \mathcal{F}^\star) + w_a R_{\mathrm{acc}}(\hat{\mathcal{F}}_i, \mathcal{F}^\star)$ (Eq. (5))
  end for
  Compute group-relative advantages $A_i \leftarrow \mathrm{Normalize}(r_1, \ldots, r_G)$ (Eq. (9))
  Update θ by maximizing the GRPO objective with KL regularization to π_{ref}:
    $\max_\theta \ \frac{1}{G}\sum_{i=1}^{G} A_i \min\Big(\rho_i(\theta), \tilde{\rho}_i(\theta)\Big) - \beta\,\mathrm{KL}\Big(\pi_\theta(\cdot \mid \mathcal{E}) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid \mathcal{E})\Big)$ (Eq. (10))
    where $\rho_i(\theta) = \frac{\pi_\theta(\hat{\mathcal{F}}_i|\mathcal{E})}{\pi_{\theta_{\mathrm{old}}}(\hat{\mathcal{F}}_i|\mathcal{E})}$ and $\tilde{\rho}_i(\theta) = \mathrm{clip}(\rho_i(\theta), 1-\tau, 1+\tau)$.
end for
return π_θ

C Extended Experiments

C.1 Noise Robustness Study

As a synthetic-noise stress test, we injected Gaussian noise into the input trajectories at varying levels of σ. As shown in Table 7, VIPER-R1 retains high structural understanding even when the input is substantially corrupted, whereas baseline VLMs degrade rapidly.

C.2 Comparison against Fine-tuned VLMs.

To ensure a fair comparison, we fine-tuned several strong baselines on the PhysSymbol training set. This includes open-source models (Qwen-VL-Max, InternVL3-8B) and the closed-source model (GPT-4o via fine-tuning API).

As shown in Table 8, fine-tuning narrows the gap to our method but does not match VIPER-R1's structural performance; in particular, GPT-4o fine-tuning is roughly tied on exact-match accuracy while remaining clearly worse on S_{struct}. This pattern is more consistent with our intended claim: SFT

Algorithm 2 Agentic Refinement via Symbolic Residual Realignment (SR^2)

Inputs: Calibrated VIPER-R1 policy π_θ (after RGSC); empirical evidence $\mathcal{E} = \{\mathcal{V}, \mathcal{D}\}$; symbolic regression solver $\mathrm{SR}(\cdot)$.

Outputs: Realigned governing law $\hat{\mathcal{F}}_{\text{final}}$.

Step 0: VIPER-R1 hypothesis (acceleration model).
Sample / decode a high-confidence candidate law $\hat{\mathcal{F}}_0 \sim \pi_\theta(\cdot \mid \mathcal{E})$
Interpret $\hat{\mathcal{F}}_0$ as $a = \hat{f}_0(x, v, t)$ (Eq. (11))
Step 1: Residual construction.
Extract supervision targets $a_{\mathrm{GT}}(t)$ and states $(x(t), v(t), t)$ from $\mathcal{D}$
Evaluate $\hat{f}_0$ on observed states and form residual
$r(t) \leftarrow a_{\mathrm{GT}}(t) - \hat{f}_0(x(t), v(t), t)$ (Eq. (12))
Step 2: Symbolic regression on the residual.
Fit a parsimonious correction term:
$f_{\text{res}} \leftarrow \mathrm{SR}\big((x, v, t),\, r(t)\big)$ (Eq. (13))
Step 3: Realigned law composition.
Compose the final law:
$\hat{\mathcal{F}}_{\text{final}} :\ a = \hat{f}_{\text{final}}(x, v, t) = \hat{f}_0(x, v, t) + f_{\text{res}}(x, v, t)$ (Eq. (14))
return $\hat{\mathcal{F}}_{\text{final}}$

Table 7: Noise robustness on the PhysSymbol Mechanics test split under additive Gaussian perturbations to the input trajectories. We report only S_{struct} at each noise level σ. Higher is better.

	Level of σ		
Model	**0.05**	**0.10**	**0.20**
GPT-5	0.31	0.26	0.14
Qwen-VL-Max	0.34	0.29	0.18
Gemini-2.5-pro	0.30	0.27	0.15
VIPER-R1-7B	**0.78**	**0.71**	**0.63**

alone can improve symbolic formulation, but RGSC still improves the structure of the base hypothesis distribution, while the external SR^2 module provides an additional end-to-end numerical-refinement step.

C.3 Supplementary Note on Stage Ablations

The authoritative stage-ablation evidence for the current paper is the main-text stage contribution analysis under the same reporting protocol used throughout the paper. We do not repeat the older 3B/7B appendix ablation table here because those archived rows mix model-scale comparison with a different evaluation setup and are not directly comparable to the current main-text numbers. Final model-size comparison remains in Table 1, while the current stage contribution evidence is summarized in Table 2.

C.4 PySR Search-Budget Sensitivity

We include one additional sensitivity check to quantify how much a classical symbolic-regression baseline can move as the search budget changes. This experiment uses the PhysSymbol Mechanics test split and the same evaluation protocol as the main paper, but it evaluates the numeric-only PySR baseline under three budgets: *small* ($\texttt{iterations} = 3$, $\texttt{maxsize} = 8$), *paper-default* $(10, 15)$, and *large* $(40, 20)$. All other settings are fixed. For this numeric-only baseline, the Post-SR^2 MSE column reports the final report-split MSE of the fitted formula under the same end-to-end scoring protocol; we do not stack an additional second refinement pass on top of PySR.

Table 9 shows that search budget affects symbolic recovery and numerical fit differently. Structural scores increase from 0.3384 to 0.352 and 0.4423 across the small, default, and large budgets, respectively. Accuracy is non-monotonic (0.1955, 0.048, and 0.3153), as is report-split MSE ($1.25 \times$

Table 8: Comparison against fine-tuned VLM baselines on the PhysSymbol Mechanics test split. SFT denotes supervised fine-tuning with our MSI curriculum, and VIPER-R1 reports the full MSI+RGSC pipeline under the same evaluation protocol. Higher is better for S_{struct} and S_{acc}.

Model	**Training Method**	$S_{\text{struct}} \uparrow$	$S_{\text{acc}} \uparrow$
Qwen-VL-Max	SFT	0.554	0.399
InternVL3-8B	SFT	0.367	0.244
GPT-4o	SFT	0.621	0.488
VIPER-R1-7B	**MSI + RGSC**	**0.812**	**0.487**

Table 9: PySR search-budget sensitivity on the PhysSymbol Mechanics test split. Only `iterations` and `maxsize` vary across rows; all other settings remain fixed under the same evaluation protocol used in the main paper. Higher is better for S_{struct} and S_{acc}; lower is better for Post-SR^2 MSE and runtime.

Budget	**Iter.**	**Maxsize**	$\mathbf{S}_{\text{struct}} \uparrow$	$\mathbf{S}_{\text{acc}} \uparrow$	**Post-SR^2 MSE**$\downarrow$
Small	3	8	0.3384	0.1955	1.25×10^7
Paper-default	10	15	0.352	0.048	$\mathbf{0.006}$
Large	40	20	0.4423	0.3153	1.46×10^{-1}

Budget	**Avg runtime (s/example)**$\downarrow$	**Success rate**$\uparrow$
Small	0.297	1.000
Paper-default	0.489	1.000
Large	1.511	1.000

10^7, 0.006, and 1.46×10^{-1}). The default configuration achieves the lowest MSE, whereas the large configuration achieves the highest symbolic scores. Runtime increases from 0.297 to 0.489 and 1.511 seconds per example. The default row reports the same PySR result as Tables 1 and 2. These runs therefore do not support a uniform improvement across all metrics as search budget increases. They illustrate why symbolic recovery and final numerical fit are evaluated separately.

C.5 Appendix-Only Qualitative Transfer Probe on Real Experimental Data

Setup: Hamiltonian NN "Real Pendulum". We include one appendix-only sim-to-real probe on the *Real Pendulum* dataset used in the Hamiltonian Neural Networks (HNN) benchmark [9]. The trajectories are sourced from real measurements of a physical pendulum originally collected by Schmidt & Lipson [27]. We use this experiment only as a small-scale qualitative sanity check beyond the synthetic PhysSymbol generator, not as a benchmark-grade external evaluation.

Protocol. We reconstruct the same visualization modalities as in PhysSymbol by rendering phase portraits and time-series curves from the raw $(\theta, \dot{\theta})$ measurements, then pass each reconstructed instance to the same inference-and-scoring pipeline used in the main paper. Because we do not establish a standardized large-sample benchmark protocol for this appendix probe, the resulting numbers should be read as indicative rather than as benchmark-grade estimates.

Qualitative findings. Despite the real-world noise sources, VIPER-R1 consistently recovers the canonical nonlinear damped pendulum structure

$$\ddot{\theta} = -k \sin\theta - c\dot{\theta},$$

while strong general-purpose VLM baselines (e.g., GPT-5, Qwen-VL-Max, Gemini 2.5 Pro) frequently (i) omit the damping term $-c\dot{\theta}$ or (ii) incorrectly linearize the restoring force (e.g., replacing $\sin\theta$ by θ), indicating weaker sensitivity to the underlying *topological* term structure under distribution shift.

Table 10: Appendix-only qualitative transfer probe on the HNN *Real Pendulum* dataset. We reconstruct the same time-series and phase-portrait evidence protocol used in PhysSymbol and report indicative S_{struct}, S_{acc}, and Post-SR^2 MSE values. These numbers are included as a small-scale sanity check rather than a benchmark-grade external estimate. Higher is better for S_{struct} and S_{acc}; lower is better for Post-SR^2 MSE.

Model	$S_{\mathrm{struct}} \uparrow$	$S_{\mathrm{acc}} \uparrow$	**Post-SR^2 MSE** $\downarrow$
GPT-5	0.24	0.11	0.082
Qwen-VL-Max	0.28	0.13	0.074
Gemini 2.5 Pro	0.26	0.12	0.079
VIPER-R1	**0.67**	**0.41**	**0.031**

Results and Discussion. Table 10 suggests that VIPER-R1 recovers the damped nonlinear pendulum structure more reliably than the generic VLM baselines even after the synthetic-to-real shift. We use this result only as qualitative support that the learned visual-to-symbolic cues do not collapse immediately outside the synthetic generator; it is not part of the paper's main benchmark evidence.

D PhysSymbol Dataset Details

D.1 Dataset Overview and Motivation

To train and evaluate our proposed VIPER-R1 framework in the setting studied here, while also providing an auxiliary multimodal resource, we constructed **PhysSymbol**, a large-scale synthetic two-part dataset centered on low-dimensional dynamics in the present paper and accompanied by a separate research release.

PhysSymbol comprises **10,000 instances** organized into two complementary parts:

- **Part 1 (Mechanics Split):** 5,000 instances focused on classical mechanics and nonlinear dynamics, used for training and evaluating VIPER-R1 in this paper
- **Part 2 (Auxiliary Research Release):** 5,000 field-based instances provided as a reusable resource, but not used in this paper's quantitative evaluation

Each instance contains a complete multimodal representation of a physical system. Part 1 provides (1) visualizations, (2) numerical measurements, (3) ground-truth governing equations, and (4) expert-level causal reasoning annotations. Part 2 provides field-based visual diagnostics, numerical field data, and governing expressions as an auxiliary release only. In this paper, we use only the low-dimensional mechanics portion for training and quantitative evaluation.

D.2 Part 1: Classical Mechanics and Dynamics (Training Data)

D.2.1 Physics Term Library and Formula Generation

The foundation of PhysSymbol Part 1 lies in a carefully designed physics term library that encompasses the fundamental mechanisms commonly encountered in classical mechanics and nonlinear dynamics. Our term library includes 11 distinct categories of physical phenomena:

Linear and Nonlinear Restoring Forces:

- Linear elasticity: $-kx$ with $k \in [0.1, 10]$
- Cubic nonlinearity: $-\beta x^3$ with $\beta \in [0.01, 5]$
- Quintic nonlinearity: $-\delta x^5$ with $\delta \in [0.001, 1]$

Velocity-Dependent Damping:

- Linear damping: $-cv$ with $c \in [0.01, 2]$
- Cubic velocity damping: $-\alpha v^3$ with $\alpha \in [0.01, 5]$

- Quintic velocity damping: $-\eta v^5$ with $\eta \in [0.001, 1]$

External and Coupling Forces:

- Temporal periodic forcing: $F\sin(\omega t)$ with $F \in [0.1, 5]$, $\omega \in [0.5, 5]$
- Spatial periodic forcing: $F\sin(\omega x)$ with parameters in similar ranges
- Position-velocity coupling: $-\gamma x v$ with $\gamma \in [0.01, 5]$

Specialized Nonlinear Terms:

- Trigonometric nonlinearity: $-x\cos(x)$, $-x\sin(x)$ (parameter-free)
- Stochastic perturbations: $\sigma\mathcal{N}(0, 1)$ with $\sigma \in [0.01, 0.5]$

The formula generation process employs a structured combinatorial approach. Each governing equation is constructed by sampling 2-5 terms from the library, with a mandatory linear restoring force to ensure physical stability. Parameters are sampled uniformly from their respective ranges, and the resulting symbolic expression is converted into an executable function for numerical integration.

D.2.2 High-Fidelity Trajectory Simulation

For each generated governing equation of the form $\ddot{x} = f(x, \dot{x}, t)$, we perform high-resolution numerical integration using the adaptive Runge-Kutta method implemented in `scipy.integrate.solve_ivp`. The simulation protocol follows these specifications:

Temporal Parameters:

- Integration duration: $T = 20$ time units
- Sampling resolution: $N = 1000$ uniformly spaced points
- Time step: $\Delta t = 0.02$ (adaptive refinement as needed)

Initial Conditions: Initial position x_0 and velocity v_0 are independently sampled from uniform distributions over $[-1, 1]$ to ensure diversity in trajectory patterns while maintaining numerical stability.

Data Output: Each simulation yields a trajectory dataset $\{(t_i, x_i, v_i, a_i)\}_{i=1}^{N}$ containing temporal evolution of position, velocity, and acceleration. This data is exported as CSV files with full numerical precision for downstream analysis.

D.2.3 Dual Visualization Strategy for Part 1

Part 1 employs a specialized visualization approach tailored for dynamical systems analysis, generating complementary visualizations that capture different aspects of system dynamics:

Phase-Space Portraits (v vs x): These plots encode the kinematic structure and stability properties of the dynamical system. Phase portraits reveal crucial qualitative features such as:

- Closed orbits indicating conservative dynamics
- Spiral trajectories suggesting damped oscillations
- Multiple attractors or limit cycles in nonlinear systems
- Geometric signatures of different restoring force types

Temporal Trajectories (x vs t): These plots emphasize the time-domain behavior and temporal patterns:

- Oscillation frequencies and amplitude modulation
- Exponential growth or decay envelopes
- Periodic forcing signatures and resonance effects
- Transient dynamics and approach to steady states

Both visualizations are rendered as high-resolution images with consistent styling, axis labeling, and grid structures to ensure visual uniformity across the dataset.

D.2.4 Expert-Level Reasoning Annotation for Part 1

A key challenge in automated scientific discovery lies in bridging the gap between raw visual observation and symbolic reasoning. To address this, the PhysSymbol Part 1 corpus incorporates detailed Causal Chain-of-Thought (C-CoT) annotations that provide structured intermediate supervision between visual evidence and symbolic laws. These annotations are generated through a structured "Oracle-Annotation" pipeline using GPT-4o. Unlike the inference stage where the model must discover the law from scratch, during this data generation phase, the annotator is provided with the ground-truth symbolic equation and trajectory data to ensure the generated reasoning is factually correct and physically rigorous.

The annotation generation follows a specific three-stage protocol, enforced via system prompts:

- **Stage 1: Phenomenological Observation (Visual).** The prompt first requires the model to describe the qualitative features of the provided plots without jumping to equations. For the Phase-Space Portrait (v vs x), it must identify geometric features such as closed orbits (conservation), inward spirals (damping), or limit cycles (non-linearity). For the Temporal Trajectory (x vs t), it identifies amplitude modulation, frequency stability, or stochastic irregularities.
- **Stage 2: Physical Mechanism Inference (Physical).** Based solely on the visual observations in Stage 1, the model is asked to infer the underlying physical mechanisms. For example, observing an "exponential decay envelope" in the time domain must be explicitly linked to a "linear damping mechanism"; detecting "frequency dependence on amplitude" must be linked to "nonlinear restoring forces."
- **Stage 3: Symbolic Mapping and Verification (Symbolic).** Finally, the model translates these physical mechanisms into specific mathematical terms (e.g., mapping "linear damping" to the term $-c \cdot v$). Since GPT-4o has access to the ground truth in this generation phase, it performs a self-verification step to ensure the derived logic chain perfectly matches the canonical equation provided in the dataset.

This structured approach ensures that the C-CoT traces are not just generic descriptions, but causal bridges that teach the student model (VIPER-R1) how to derive specific symbolic terms from specific visual cues.

D.3 Part 2: Auxiliary Field-Based Research Release

Part 2 is an auxiliary release of 5,000 field-based examples spanning selected electromagnetism, thermodynamics, fluid mechanics, quantum-inspired toy systems, and general PDE-style operators. It is not used for training or quantitative evaluation in this paper; we include it only as a reusable resource for future multimodal field-reasoning studies.

Each Part 2 instance is generated on a 256×256 spatial grid and paired with eight diagnostic views: a scalar field, a vector field, $\partial\phi/\partial x$, $\partial\phi/\partial y$, gradient magnitude, divergence, curl, and Laplacian. The corresponding field data and governing expression are stored alongside these diagnostics. Parameters are sampled from physically meaningful ranges for each template, and color scaling is kept consistent within each field family.

D.4 Dataset Assembly and Multi-Format Generation

The final dataset assembly process integrates all components into a unified multimodal format suitable for different training stages and research applications:

Data Instance Structure for Part 1: Each complete instance follows the tuple format:

$$\begin{pmatrix} \text{Images:} & I_{\text{phase}}, I_{\text{trajectory}} \\ \text{Trajectory Data:} & M \\ \text{Ground-Truth Formula:} & \mathcal{F}^\star \\ \text{C-CoT Reasoning:} & C \end{pmatrix}$$

Data Instance Structure for Part 2: Each complete instance follows the format:

$$\begin{pmatrix} \text{Images:} & \{I_{\text{scalar}}, I_{\text{vector}}, I_{\nabla x}, I_{\nabla y}, \\ & I_{|\nabla|}, I_{\text{div}}, I_{\text{curl}}, I_{\nabla^2}\}, \\ \text{Field Data:} & F, \\ \text{Equation:} & \mathcal{F}, \\ \text{Parameters:} & P \end{pmatrix}$$

Multi-Stage Training Variants: To support our three-stage training pipeline, Part 1 generates three dataset variants:

1. **Stage 1 (MSI-Joint):** Full format requiring both reasoning generation and formula prediction
2. **Stage 2 (MSI-Guided):** C-CoT provided as input, only formula prediction required
3. **Stage 3 (RGSC):** Streamlined format for reinforcement learning with structural rewards

D.5 Dataset Statistics and Quality Assurance

The complete **PhysSymbol** corpus consists of 10,000 multimodal instances distributed as follows:

Part 1 Statistics (Training Data):

- 5,000 complete instances with trajectory-based visualizations
- Formula complexity: 2-5 terms per equation (average 3.2 terms)
- Coverage: 11 distinct physical mechanism types

Part 2 Statistics (Research Data):

- 5,000 complete instances with field/operator visualizations
- Coverage: 15 equation types across 5 physics disciplines
- Total files: 85,000 (8 visualizations + 8 data files + 1 info file per instance)
- Purpose: Auxiliary research release (not used in our training/testing)

To guarantee reliability and usability, we apply comprehensive quality control measures across both parts. Each generated equation undergoes numerical stability verification to avoid degenerate or divergent solutions. Representative samples are visually inspected to confirm clarity and readability of plots. For Part 1, the accompanying C-CoT rationales are validated through automated keyword analysis to ensure coherence with underlying formulas. Finally, file integrity checks are conducted across all multimodal components, ensuring the dataset is complete, consistent, and ready for large-scale experimentation.

The two-part design separates the current paper's training-and-evaluation regime (Part 1) from an auxiliary research resource (Part 2). In the present work, our quantitative claims are limited to the low-dimensional mechanics setting, while Part 2 is released only as a reusable dataset asset.

E Qualitative Cases

To complement the aggregate evidence in the main paper, we curate four qualitative case studies: three positive cases that illustrate where the visual-to-symbolic pipeline works as intended, and one near-miss that exposes a real failure mode. We keep the presentation deliberately concrete: each case shows the raw reasoning panel, acceleration-level diagnostics, and system-level replay plots.

Appendix roadmap: what each qualitative case is meant to show

Case	System pattern	Takeaway
Positive 1	Noisy non-linear damping	Visual evidence exposes linear restoring, cubic damping, and stochasticity simultaneously; SR^2 mainly sharpens the coefficients of an already correct symbolic skeleton.
Positive 2	Mixed restoring, forcing, and noise	The model recovers a multi-term law with quintic correction and periodic forcing, and the refined replay remains stable at the system level.
Positive 3	Driven oscillator with interaction damping	Coupled terms and driven dynamics are identified with near-exact acceleration replay, illustrating the strongest end-to-end behavior of the pipeline.
Failure	Periodic forcing with a missing non-linear interaction	The model captures the dominant oscillatory terms but misses a key interaction term, so long-horizon drift remains even after SR^2 refinement.

E.1 Positive Case 1: Noisy Non-linear Damping

This case supports the visual-grounding story behind the modality ablation in the main text. The paired $x(t)$ and $v(x)$ plots expose three distinct cues at once: oscillatory structure from the restoring force, phase-space distortion from the cubic damping term, and irregular fluctuations from stochastic forcing. VIPER-R1 recovers exactly that symbolic decomposition, which gives SR^2 a structurally correct starting point.

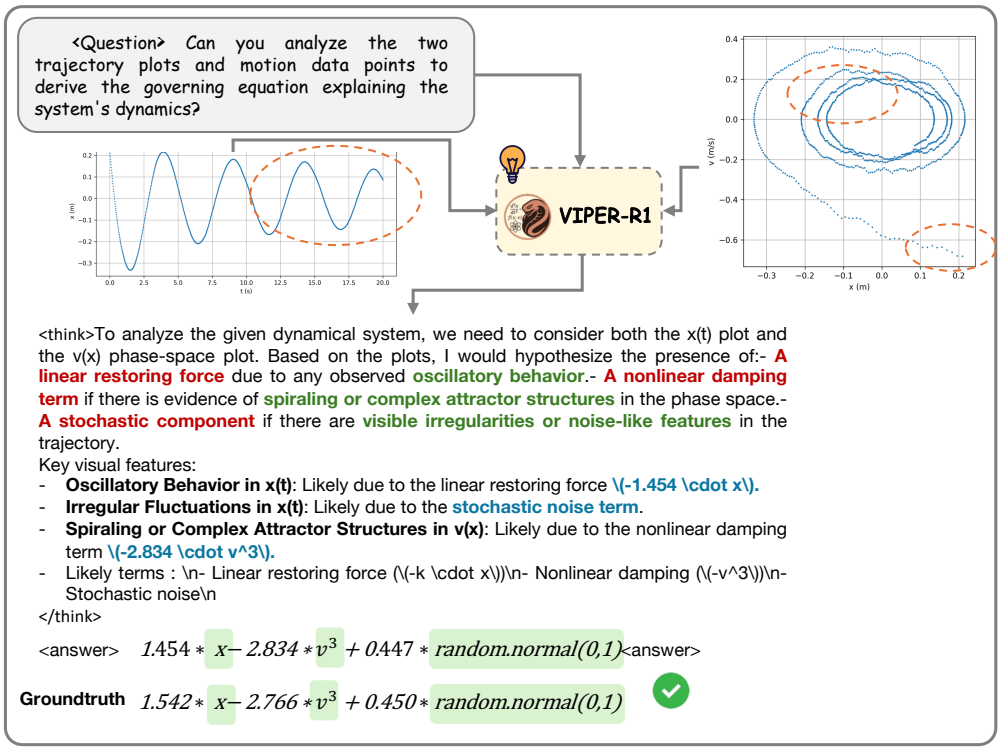
Because the ground truth itself contains stochastic forcing, we do not expect perfect trajectory overlap even after refinement. The relevant qualitative success here is that the model identifies the right term inventory and lands close enough in parameter space for the refined replay to remain physically plausible.

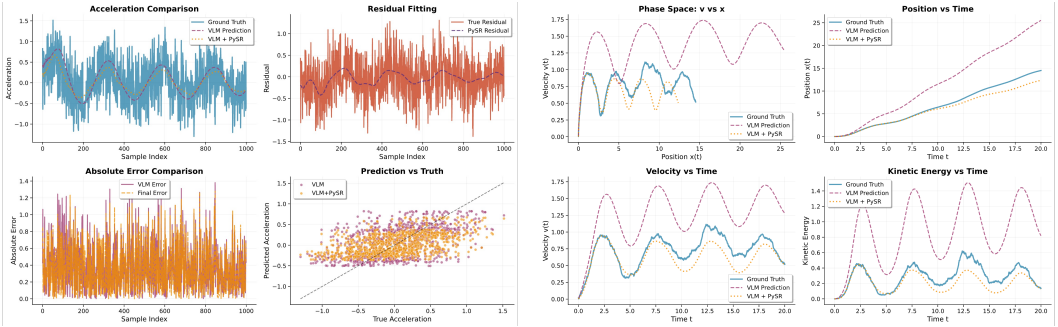
E.2 Additional Cases

The next three figures are organized as two more positive cases followed by one failure. Together they show where the pipeline is strongest and where it still breaks: multi-term in-domain recovery remains robust, while structurally incomplete hypotheses can still limit the downstream replay even when SR^2 reduces pointwise acceleration error.

The model's internal reasoning, extracted from its output, is as follows:

> <think> To analyze the given dynamical system, we consider both the $x(t)$ plot and the $v(x)$ phase-space plot. **Key visual features:**
>
> - **Oscillatory Behavior in x(t):** Suggests a linear restoring force ($-k \cdot x$).
> - **Distorted, Spiraling Attractor in v(x):** Indicates a non-linear damping term, likely dependent on a higher power of velocity ($-c \cdot v^3$).
> - **Irregular, High-Frequency Fluctuations in x(t):** Points to a stochastic noise term.
>
> **Conclusion:** The system likely combines a linear restoring force, non-linear damping ($-v^3$), and stochastic noise. </think>

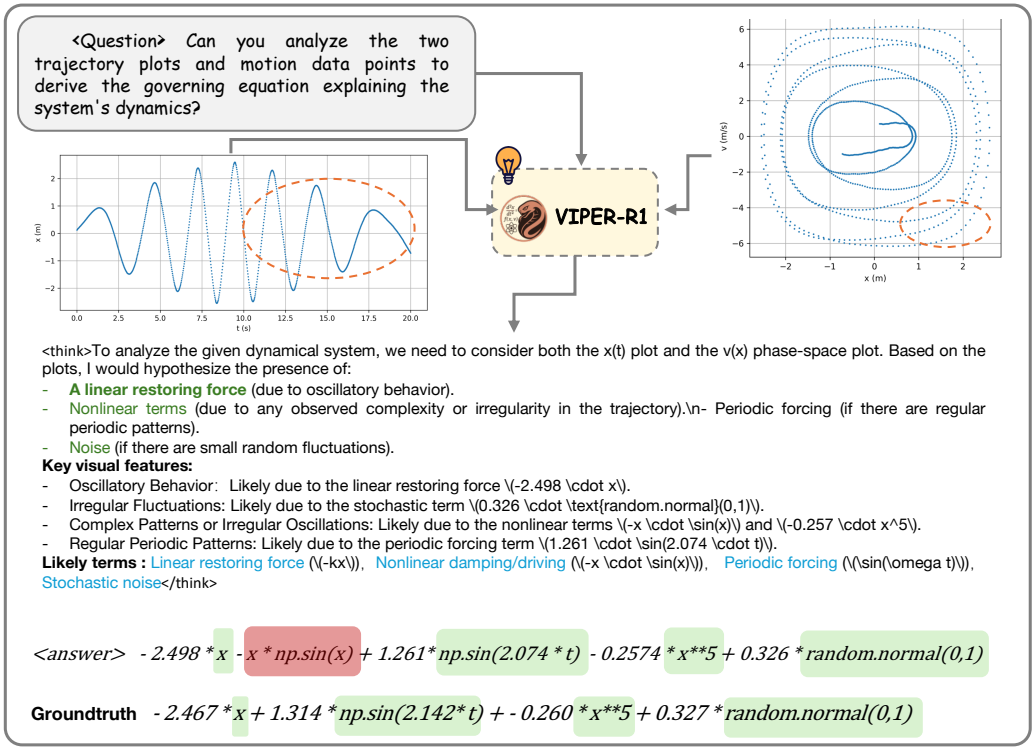


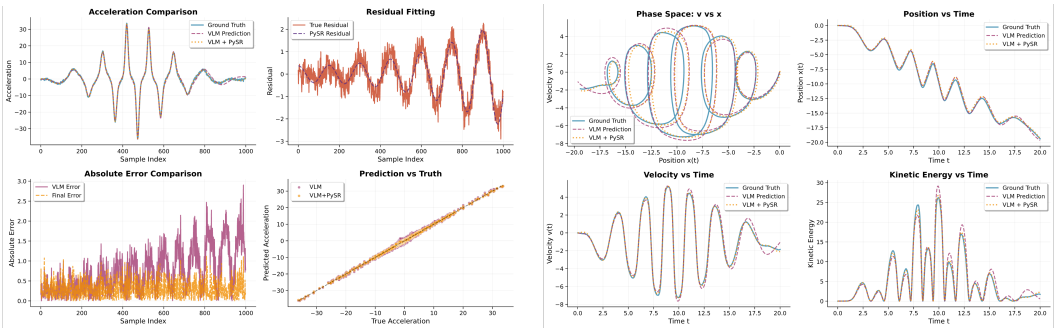


(a) Acceleration signal evaluation before (VLM-only) and after (VLM+SR^2).

(b) System-level trajectory comparison showing improved long-horizon consistency.

Figure 10: Positive Case 1: noisy non-linear damping with stochastic forcing. **Top:** VIPER-R1 identifies the linear restoring term, the cubic damping term, and the noise component directly from the paired $x(t)$ and $v(x)$ evidence. **Bottom:** SR^2 mainly sharpens the coefficient estimates; because the target system is noise-driven, the refined replay improves the dynamics without collapsing to a perfectly identical long-horizon trace.



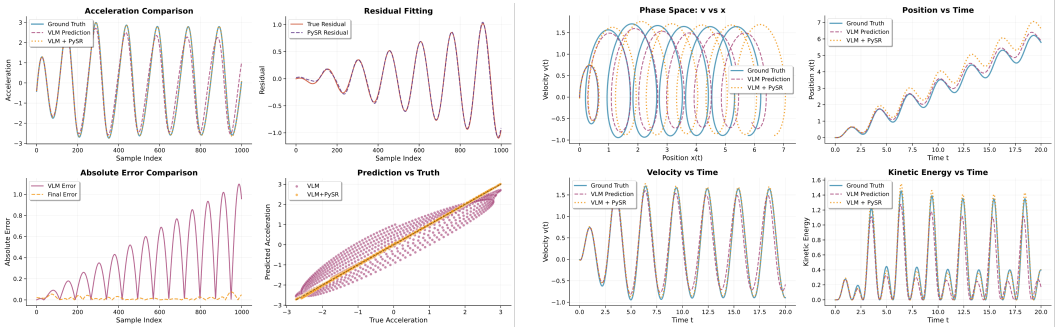


(a) Acceleration signal evaluation. (b) System-level trajectory comparison.

Figure 11: Positive Case 2: mixed restoring, forcing, and noise terms. **Top:** VIPER-R1 recovers a four-term law that includes linear restoring dynamics, a higher-order correction, periodic forcing, and stochastic noise. **Bottom:** both acceleration-level and system-level diagnostics remain close to the ground truth after SR^2 refinement.

E.3 Positive Case 2: Mixed Restoring, Forcing, and Noise Terms

This example is the cleanest illustration that the model can recover a genuinely multi-term symbolic law rather than a single dominant oscillator template. The predicted equation contains the same four ingredients as the ground truth: linear restoring force, higher-order correction, periodic forcing, and a stochastic component. After SR^2 refinement, the acceleration traces and long-horizon replay remain tightly aligned with the target system.



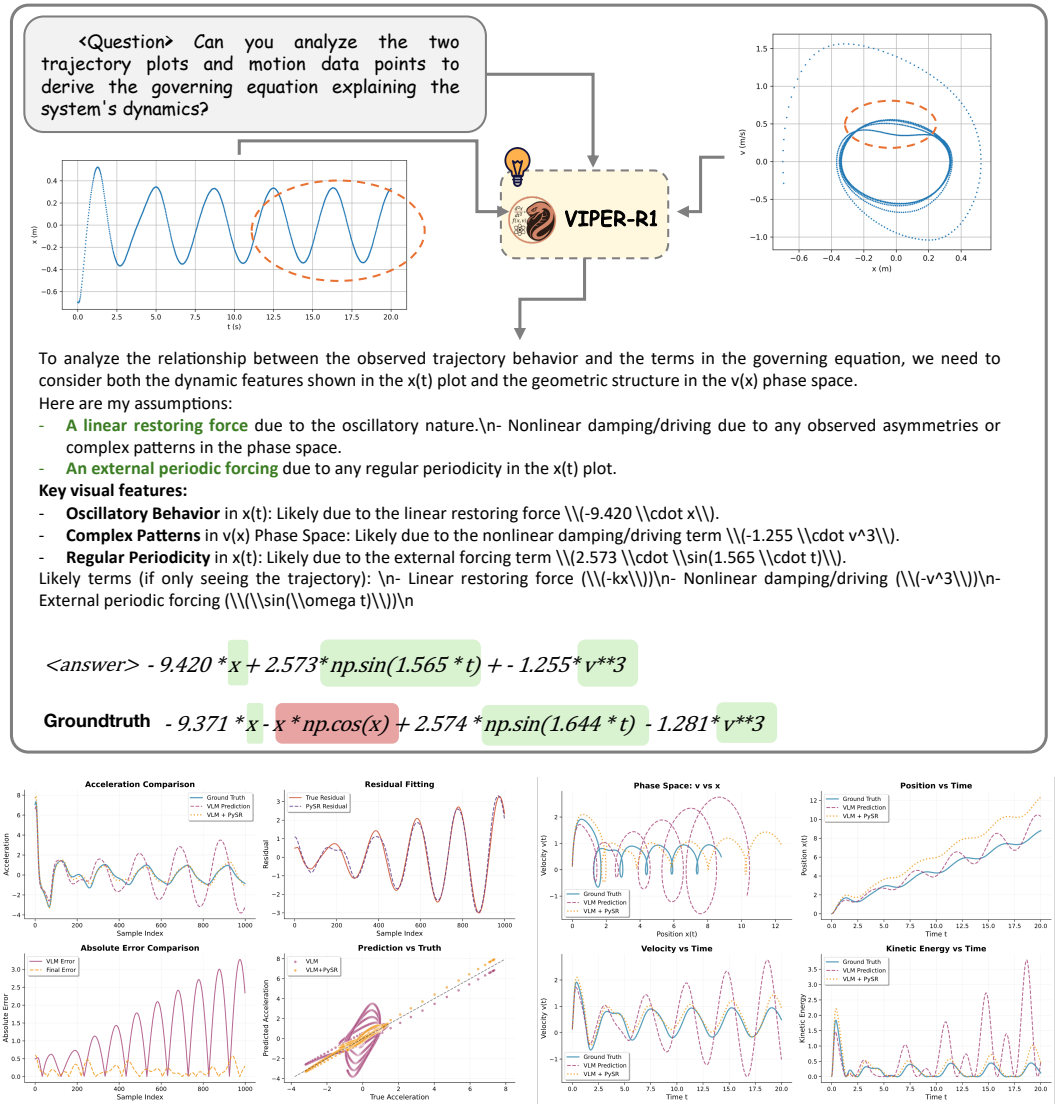
(a) Acceleration signal evaluation.

(b) System-level trajectory comparison.

Figure 12: Positive Case 3: driven oscillator with interaction damping (reasoning trace in Figure 5). After SR^2, the refined acceleration and trajectory diagnostics remain tightly matched to the ground truth.

E.4 Positive Case 3: Driven Oscillator with Interaction Damping

This is the strongest end-to-end replay case in the appendix. Beyond the standard restoring and forcing terms, the target law contains a coupled interaction term between position and velocity. VIPER-R1 identifies that interaction correctly, and the final refined replay is nearly indistinguishable from the ground truth at the acceleration level. The reasoning trace for this case appears in the main text (Figure 5).



(a) Acceleration signal evaluation. (b) System-level trajectory comparison.

Figure 13: Failure case: periodic forcing with a missing non-linear interaction. **Top:** VIPER-R1 captures the dominant linear, periodic, and cubic terms but omits the $-x\cos(x)$ interaction present in the ground truth. **Bottom:** SR^2 improves the local acceleration fit, but the structurally incomplete ansatz still produces noticeable long-horizon drift.

E.5 Failure: Missing a Non-linear Interaction under Periodic Forcing

This figure illustrates the failure mode we do not want to hide. The model captures the dominant oscillatory ingredients, including the linear restoring term, the periodic drive, and the cubic velocity term, but it misses the additional non-linear interaction $-x\cos(x)$ present in the ground truth. SR^2 still reduces the immediate acceleration error, yet the structurally incomplete ansatz leaves visible long-horizon drift in the replayed trajectories. This is exactly why we describe Post-SR^2 MSE as a tool-augmented end-to-end metric rather than proof that the original symbolic hypothesis is fully correct.

F Impact Statement

This paper studies vision-grounded formula discovery for low-dimensional dynamical systems from rendered trajectories and phase portraits. Potential positive impacts include more interpretable tools for scientific education, hypothesis generation, and physics-inspired model analysis. The main risk is over-trust: users may mistake a plausible symbolic fit, especially after external SR^2 refinement, for a validated scientific law. This risk is amplified when the method is deployed beyond the evidence protocol studied here. Accordingly, the system should be used as a tool for interpretable hypothesis generation and analysis, not as an automatic substitute for scientific validation.

G Limitations

Synthetic evidence and sim-to-real gap. PhysSymbol is synthetic by construction, with standardized simulation and rendering pipelines. Although this enables controlled evaluation of evidence $\rightarrow$ rationale $\rightarrow$ equation induction, it may not capture real experimental artifacts (e.g., sensor noise/drift, occlusion, calibration errors, uncontrolled backgrounds). Broader validation on diverse real experimental datasets (and especially real videos) remains future work.

Restricted equation families and physics coverage. Although PhysSymbol includes a PDE-style research split (Part 2), the quantitative evaluation in this paper focuses on curated ODE systems with small term counts from Part 1, plus a limited real-pendulum transfer study. This does not yet cover many practically important regimes, such as strongly chaotic or stiff dynamics, hybrid/discontinuous systems (contacts/events), partial observability, or high-dimensional coupled systems. Expanding the operator library, domain coverage, and external evaluation suite may change both visual cues and calibration difficulty.

Dependence on structured supervision. MSI benefits from expert-style C-CoT rationales paired with ground-truth laws, but this introduces curation and annotation overhead. Scaling to new scientific domains may require additional expert effort to maintain rationale quality and consistency.

Reward/metric brittleness under equivalent forms. RGSC uses protocol, structural-skeleton similarity, and exact-match signals that can be brittle under algebraically equivalent but syntactically different laws (e.g., re-parameterizations, constant folding, alternative normal forms), or when multiple valid formulations exist. More robust reward definitions that respect equivalence classes (e.g., invariances or conserved quantities) are an important direction.

Tool dependence and inference overhead. SR^2 improves numerical fidelity by invoking an external symbolic regression engine, but this adds system complexity and sensitivity to SR hyperparameters, and can be computationally expensive in noisy or difficult regimes. Failures in the tool stage may dominate end-to-end performance. Accordingly, Post-SR^2 MSE should be interpreted as tool-augmented end-to-end performance rather than proof that the raw VLM hypothesis is universally correct.

Training cost. VIPER-R1 requires multi-stage training and group-based sampling for GRPO, which is more expensive than pure supervised fine-tuning. Improving sample efficiency and reducing RL cost are important for practical scaling.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the scope of the work as vision-grounded formula discovery for low-dimensional dynamical systems, and the claims are aligned with the experiments on the PhysSymbol Mechanics split and the matched external subset. The Discussion and Limitations sections further clarify that the method improves first-order symbolic hypothesis formation under the paired-evidence regime, rather than claiming universal dominance over all classical SR settings.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated Limitations section in Appendix G, discussing the synthetic-data setting, restricted equation families, dependence on C-CoT supervision, reward/metric brittleness, SR^2 tool dependence, inference overhead, and training cost.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present formal theoretical results such as theorems or lemmas requiring assumptions and proofs. The mathematical expressions are used to define objectives, rewards, and evaluation metrics.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the method, objectives, reward functions, evaluation metrics, prompts, hyperparameters, compute, and dataset construction in the main text and appendices. Algorithm pseudocode and detailed MSI, RGSC and SR^2 configurations are provided to support reproduction.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code is released at `https://github.com/Jiaaqiliu/VIPER-R1` and the trained checkpoint at `https://huggingface.co/JiaaqiLiu/VIPER-R1-RL`. Dataset construction, splits, prompts, and hyperparameters are documented in Appendices B and D.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Training and test settings are specified in the main experiments section and Appendix B, including dataset splits, hyperparameters, optimizer, reward weights, prompts, inference protocol, and evaluation metrics.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The paper clearly distinguishes deterministic final-checkpoint point estimates from archived descriptive variability summaries and states that the archived ± entries are not confidence intervals over independent retraining runs. Bootstrap 95% intervals are reported for the external-transfer comparison where applicable.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix B reports the hardware and training cost for MSI and RGSC, including A800 GPU type, number of GPUs, wall-clock duration, and GPU-hours.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses physical systems and public or cited model or API baselines, involves no human subjects or sensitive personal data, and the authors have reviewed and complied with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix F discusses positive impacts, including interpretable tools for scientific education and hypothesis generation, as well as negative risks such as over-trust in plausible symbolic fits. It recommends using VIPER-R1 as a hypothesis-generation and analysis tool rather than a substitute for scientific validation.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk generative model, scraped internet dataset, or safety-sensitive asset. The released materials are for physics formula discovery on synthetic/controlled scientific data, and the main misuse risk is over-trust, which is discussed in the Impact Statement and Limitations.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the main existing assets and baselines used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces PhysSymbol and provides documentation of its construction, splits, modalities, formula generation process, visualizations, C-CoT traces, and limitations in Appendix D. The supplementary material is intended to provide code and data access and sample test data for verification.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing experiments or research with human subjects. The dataset is synthetic/simulated and consists of rendered plots, numerical trajectories/fields, symbolic laws, and generated or curated reasoning traces.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects, collection of personal data, or participant-facing experiments. Therefore IRB approval or equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper describes the use of an LLM-based oracle/annotation pipeline for generating expert-style C-CoT traces in PhysSymbol; this is part of the dataset construction rather than an undisclosed writing aid.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.