This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Pixel2Phys: Distilling Governing Laws from Visual Dynamics

Ruikun Li[2*†], Jun Yao[1,3†], Yingfan Hua[1], Shixiang Tang[1,4], Biqing Qi[1],
Bin Liu[3], Wanli Ouyang[1,4], Yan Lu[1,4‡]
[1]Shanghai Artificial Intelligence Laboratory [2]Tsinghua University
[3]University of Science and Technology of China [4]The Chinese University of Hong Kong

Abstract

Discovering physical laws directly from high-dimensional visual data is a long-standing human pursuit but remains a formidable challenge for machines, representing a fundamental goal of scientific intelligence. This task is inherently difficult because physical knowledge is low-dimensional and structured, whereas raw video observations are high-dimensional and redundant, with most pixels carrying little or no physical meaning. Extracting concise, physically relevant variables from such noisy data remains a key obstacle. To address this, we propose Pixel2Phys, a collaborative multi-agent framework adaptable to any Multimodal Large Language Model (MLLM). It emulates human scientific reasoning by employing a structured workflow to extract formalized physical knowledge through iterative hypothesis generation, validation, and refinement. By repeatedly formulating, and refining candidate equations on high-dimensional data, it identifies the most concise representations that best capture the underlying physical evolution. This automated exploration mimics the iterative workflow of human scientists, enabling AI to reveal interpretable governing equations directly from raw observations. Across diverse simulated and real-world physics videos, Pixel2Phys discovers accurate, interpretable governing equations and maintaining stable long-term extrapolation where baselines rapidly diverge.

1. Introduction

Discovering physical laws from observational data lies at the core of scientific understanding and has historically driven major advances across physics, astronomy, and the natural sciences [1, 7, 44, 49]. As AI systems increasingly participate in scientific workflows [20, 32, 53], the ability to infer governing equations directly from visual observations becomes a critical step toward genuine AI for Science. Human scientists typically achieve such discoveries by distilling structured, low-dimensional physical variables from complex visual phenomena and formulating compact symbolic laws [9, 36]. However, this manual process is slow and labor-intensive, classic examples such as translating Tycho Brahe's astronomical measurements into Kepler's laws required decades of expert reasoning. Automating this capability, *visual equation discovery*, would therefore accelerate scientific progress.

However, visual equation discovery is extremely challenging because the true physical signals in a video are low-dimensional and concise, yet they are submerged within a large amount of visually irrelevant content such as textures, lighting variations, and background motion. These redundant components dominate the pixel space and obscure the compact physical structure that scientific laws depend on. Prior approaches have attempted to address this difficulty in several ways, but each faces fundamental limitations. (1) Supervised equation-prediction models [2, 27, 39, 53]: map structured physical data directly to symbolic equations. However, they cannot process raw video pixels and rely on pre-extracted, noise-free state variables, which are difficult to obtain from complex visual scenes. (2) Unsupervised latent-coding methods: first learn latent representations from videos using autoencoding or predictive objectives [5, 16–18, 28, 51], and then apply symbolic regression on these latents. However, because the latent spaces are determined by reconstruction or next-frame prediction optimization process rather than physical structure, they are often non-unique and easily entangle visually salient but physically irrelevant factors. (3) Recent Multimodal Large Language Models (MLLMs) combine visual understanding with strong symbolic reasoning, suggesting potential for equation discovery [3, 4, 30]. Yet directly prompting an MLLM mainly retrieves and recombines prior knowledge from its training corpus; without a structured workflow, it struggles to infer new physical variables or derive laws purely from raw visual data. These challenges highlight the need for a method that can reliably extract the physically meaningful structure hidden inside high-dimensional

*Work done during an internship at Shanghai AI Laboratory.
†Equal contribution.
‡Corresponding author (luyan@pjlab.org.cn).

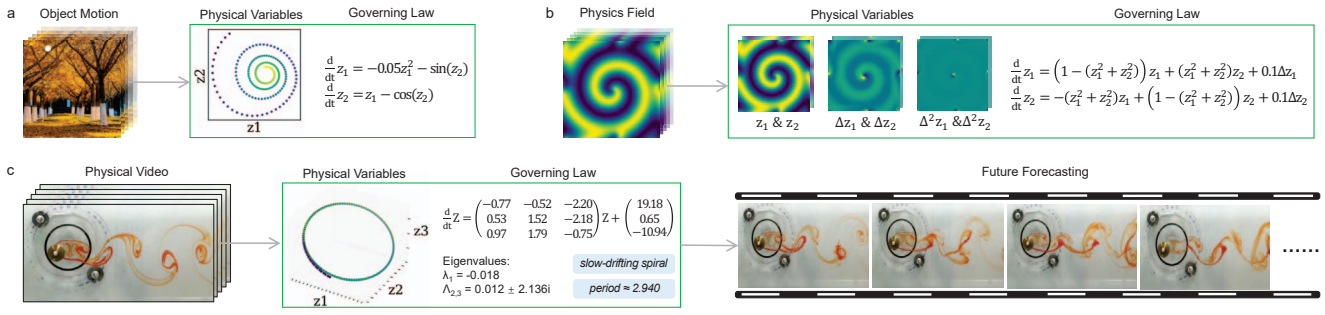


Figure 1. Distill governing laws from videos. (a) Trajectory dynamics of moving objects. (b) Spatiotemporal dynamics of time-varying physical fields. (c) Intrinsic dynamics of physical phenomena.

videos and support the discovery of new equations.

To address these challenges, we propose Pixel2Phys, a collaborative multi-agent framework that can be paired with any MLLM and organizes visual equation discovery into a structured, iterative scientific workflow (Figure 2a). At the center of the framework is the Plan Agent, which coordinates three specialized agents, the Variable Agent, the Equation Agent, and the Experiment Agent, and determines the refinement strategy across iterations. The Variable Agent extracts physical quantities from videos through several complementary defined tools that correspond to different classes of physical systems, enabling it to recover object-level motion, field-level evolution, and other low-dimensional structures that often underlie governing equations. The Equation Agent forms equation candidates by dynamically utilizing symbolic regression components, such as adjusting sparsity constraint strength or selecting different symbolic libraries. The Experiment Agent evaluates each candidate equation by several metrics, such as simulating its predicted dynamics, measuring reconstruction discrepancies, testing temporal extrapolation, and summarizing these results into a structured report for the Plan Agent. The Plan Agent integrates these reports and issues targeted instructions, such as enforcing stronger sparsity in the Equation Agent or requiring the Variable Agent to incorporate dynamical constraints, for a next refinement step. Through this iterative cycle of hypothesis formation, evaluation, and adjustment, Pixel2Phys progressively filters away visually irrelevant factors and converges toward variables and equations that best describe the true underlying dynamics. Crucially, each step in this workflow is mechanical and fully within the execution capabilities of modern MLLMs, allowing Pixel2Phys to combine the model's strong single-step reasoning with structured multi-step coordination, and thereby to discover new variables and governing equations beyond what is present in the model's training data. We evaluate our framework on three categories of physics videos with increasing difficulty. The results demonstrate that by distilling accurate and interpretable governing equations, our proposed framework improves extrapolation accuracy (RMSE) by 45.35% over baselines.

Our contributions can be summarized as follows:

- We propose a novel multi-agent framework for visual equation discovery, where an MLLM planner coordinates specialized agents to parse complex visual dynamics at multiple granularities.
- We design an iterative, co-optimization reasoning process that breaks the circular dependency between visual representation learning and law discovery.
- Extensive experiments on three challenging categories of physics videos demonstrate that our framework not only discovers physically interpretable governing equations but also improves long-term extrapolation accuracy by 45.35%.

2. Related Work

2.1. Inferring Physics from Video

Existing approaches generally tackle visual dynamics from two perspectives. The first category focuses on **implicit neural dynamics**, learning latent representations to perform future prediction. Modules like Neural ODEs [10, 18] and Koopman operators [5] are often integrated to model continuous evolution. Other works decompose video into PDE dynamics [16, 51] or disentangled representations [14, 52]. While effective for prediction, these methods encapsulate physical laws within black-box networks, lacking explicit interpretability. The second category, **physics-informed perception**, imposes strong inductive biases. Researchers have utilized inverse graphics [23] or enforced rigid body constraints [25] to infer specific properties like mass and friction. However, these methods rely heavily on *a priori* knowledge of the governing equations, limiting their ability to discover unknown laws from unfamiliar physical phenomena.

2.2. Visual Equation Discovery

Equation discovery from video is a nascent field that typically follows two paradigms: (1) **Pipeline-based approaches** [34, 54], which extract explicit trajectories of ob-

jects and then apply symbolic regression. These methods rely on pre-trained trackers and struggle with continuous fields (e.g., fluids) where objects are undefined. (2) **End-to-end approaches** [8, 46], which learn a latent coordinate space for equation regression. However, they face a critical **circular dependency**: extracting a good variable space requires knowledge of the dynamics, while finding the dynamics requires a clean variable space. Consequently, they often settle for sub-optimal solutions in complex visual scenarios. **Our work introduces a third paradigm:** a collaborative, multi-agent framework. By establishing a reasoning-driven feedback loop, we enable variable extraction and equation discovery to mutually refine each other, effectively breaking the circular dependency and extending discovery to diverse visual dynamics.

3. Problem Fomulation

In the task of inferring governing laws from high-dimensional data, the goal is to find a compact and accurate symbolic expression. We are given a sequence of high-dimensional visual observations (i.e., a video), $\{\mathbf{X}(t)\}_{t=1}^{N}$, where each frame $\mathbf{X}(t) \in \mathbb{R}^D$ (D is the pixel space, e.g., $C \times H \times W$). We assume this video observes an underlying, unknown, low-dimensional dynamical system evolving on an intrinsic manifold [8], $\mathcal{Z} \subseteq \mathbb{R}^d$, where $d \ll D$. Our goal is to simultaneously discover both the intrinsic physical variables $z(t) \in \mathcal{Z}$ and their symbolic governing law f, which defines the system's evolution, $\frac{dz}{dt} = f(z(t))$. This presents a challenging dual discovery problem, as both the coordinate system z and the function f are unknown. This problem can be decomposed into two coupled sub-problems:

- **Physical Variable Extraction.** Learning an encoder $\phi : \mathbb{R}^D \to \mathbb{R}^d$ that maps the high-dimensional visual observation $\mathbf{X}(t)$ to its corresponding intrinsic physical state $z(t) = \phi(\mathbf{X}(t))$. To ensure $z(t)$ is an informative representation, this component often includes a decoder $\psi : \mathbb{R}^d \to \mathbb{R}^D$, used to enforce a reconstruction constraint $\mathbf{X}(t) \approx \psi(z(t))$.
- **Governing Law Distillation.** Identifying an interpretable, symbolic expression for the dynamics function f from a library of candidate functions Θ (e.g., polynomials, trigonometric terms). This function must accurately model the evolution of the extracted variables $z(t)$.

The synergy between these components is the central challenge. The quality of the extracted variables $z = \phi(\mathbf{X})$ directly dictates the simplicity and accuracy of the distilled law f. Conversely, a simple and sparse law f provides a powerful signal to guide the variable extraction process ϕ, compelling it to filter out dynamically irrelevant visual components. Our framework is designed to solve this co-optimization problem, seeking a self-consistent pair of an intrinsic variable space $\mathcal{Z}$ and its symbolic law f that fit the observed data and generalize for long-term prediction.

4. Method

4.1. Pixel2Phys Method Overview

The Pixel2Phys framework mimics the collaborative workflow of a human scientific team to solve the dual discovery problem defined in Section 3, involving observing, hypothesizing, experimenting, and refining. As illustrated in Figure 2, Pixel2Phys consists of four agents with distinct roles: The Plan Agent acts as the team's central coordinator, responsible for setting goals, analyzing reports, diagnosing bottlenecks, and providing instructional prompts $\mathcal{I}_{\text{plan}}$ to other agents to guide the next iteration. The Variable Agent executes visual parsing tasks, i.e., parsing and extracting low-dimensional physical variables $z(t)$ from the high-dimensional observations $\mathbf{X}(t)$. The Equation Agent is responsible for distilling the symbolic governing equation f from $z(t)$. Finally, the Experiment Agent validates the quality of the current $(\mathcal{Z}, f)$ pair. In each iteration, the Variable, Equation, and Experiment Agent are required to return a report (denoted as $\mathcal{R}_{var}$, $\mathcal{R}_{equ}$, and $\mathcal{R}_{exp}$) to the Plan Agent, respectively, containing both quantitative performance metrics and qualitative visual diagnostics for the next decision-making step. These agents are instantiated as MLLMs and operate under a unified protocol where they process visual data $\mathbf{X}$ and textual prompts $\mathcal{P}$ to generate textual responses $\text{MLLM}(\mathbf{X}, \mathcal{P})$. Algorithm C.5 outlines the pseudocode of Pixel2Phys's execution process.

4.2. Plan Agent: Global Planning in Iterative Reasoning

Our framework replaces a simple pipeline with an iterative reasoning loop driven by the Plan Agent. This agent is the central coordinator that orchestrates the co-optimization of the variable space $\mathcal{Z}$ and the governing law f.

The reasoning loop commences with an initialization step ($k = 0$). The Plan Agent interprets the user's task and activates the initial iteration, yielding the candidate pair $(\mathcal{Z}_0, f_0)$ alongside the reports $\mathcal{R}^0_{\text{var}}$, $\mathcal{R}^0_{\text{equ}}$, and $\mathcal{R}^0_{\text{exp}}$ from each agent. In subsequent iterations ($k \geq 1$), the Plan Agent aggregates these reports to conduct a two-fold diagnosis. It first inspects the visualizations in $\mathcal{R}^{k-1}_{\text{exp}}$ to assess qualitative dynamical fidelity, then scrutinizes the specific quantitative metrics with $\mathcal{R}^{k-1}_{\text{var}}$ and $\mathcal{R}^{k-1}_{\text{equ}}$ to pinpoint the exact bottleneck. Based on this analysis, it formulates a instructional prompt $\mathcal{I}^k_{\text{plan}}$ to resolve the identified failure mode:

- **Variable Refinement.** When the diagnosis attributes failure to a poor $\mathcal{Z}$ (e.g., high reconstruction errors), the Plan Agent instructs the Variable Agent to re-extract $\mathcal{Z}$. Crucially, Plan Agent will provide the equation f_{k-1} to activate the physics-informed loss (Section 4.3).
- **Equation Refinement.** When $\mathcal{Z}$ is deemed high-quality

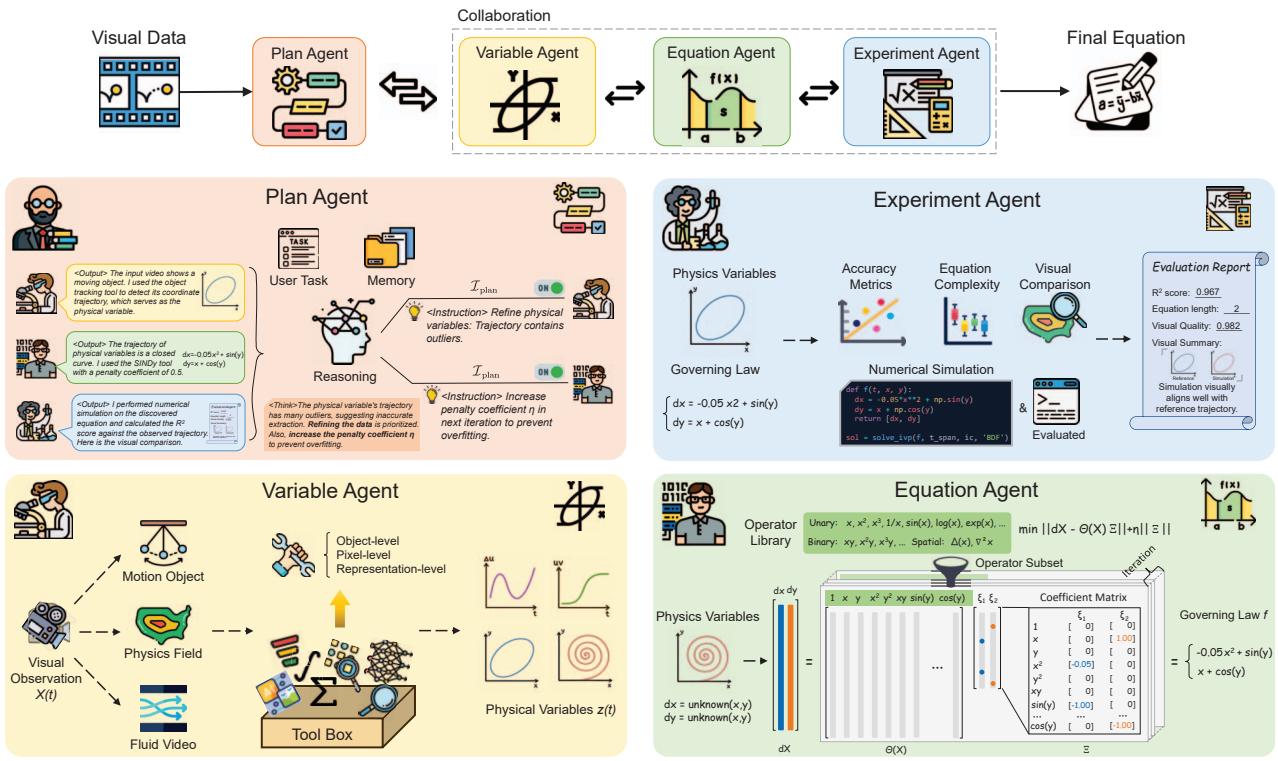


Figure 2. The multi-agent collaboration framework of Pixel2Phys.

but f_{k-1} is inaccurate or overly complex, the Plan Agent instructs the Equation Agent to modify its equation search by adjusting the configuration hyperparameters.

This iterative refinement loop continues until the Plan Agent determines that $\mathcal{R}^k_{\text{exp}}$ satisfies the success criteria (Appendix E.2). The final $(\mathcal{Z}_k, f_k)$ pair is then returned as the solution.

4.3. Variable Agent: Variable Extraction from Visual Data

Given a video sequence $\mathbf{X}(t)$, the Variable Agent extracts physical variables $z(t)$ by deploying specific parsing tools. We provide multi-granularity tools for flexible usage to accommodate physical information presented at object, pixel, and representation levels. The agent dynamically selects the corresponding tool based on both the video's visual properties and the explicit instructional prompt $\mathcal{I}_{\text{plan}}$ issued by the Plan Agent. Specifically, we provide object-level tools for videos of moving objects, pixel-level tools for physical fields, and autoencoder-based tools for complex scientific phenomena.

Object-level Tool. For moving objects, such as celestial revolution, the Variable Agent extracts the trajectory $z(t) = [x(t), y(t)]$. We employ visual foundation models for zero-shot object segmentation and tracking to avoid training on specific shapes of objects. Specifically, we utilize Segment Anything [26] to segment potential objects in every frame. The agent then filters out static targets, as detailed in Appendix C.1, and records the centroid coordinates of moving objects as $z(t)$.

Pixel-level Tool. For physical fields governed by PDEs (Figure 1b), video frames $\mathbf{X}(t)$ are treated as samples of a continuous field $u(\mathbf{x}, t)$ discretized at pixels $\mathbf{x}$ [7, 50]. Physical dynamics emerge from local spatial interactions. We equip the Variable Agent with the fixed convolutional kernels to compute spatial derivatives directly from the pixel grid, yielding operators such as the Laplacian Δ and biharmonic Δ^2. Since the governing equation is valid at any pixel, the agent randomly samples a sparse subset of pixels to efficiently collect the spatial operators as physical variables $z(t) = [u(t), \Delta u(t), \Delta^2 u(t), \dots]$.

Representation-level Tool. For complex scientific phenomena, underlying dynamics are often obscured by device noise and lighting fluctuations (Figure 1c). We adapt a novel physics-informed autoencoder to compress $\mathbf{X}(t)$ into a latent representation $z(t) = \phi(\mathbf{X}(t))$, ensuring the underlying variables remain as simple as possible. This tool operates in two modes depending on the availability of physical priors. First, when the instructional prompt

$\mathcal{I}_{\text{plan}}$ contains a governing equation f discovered in the previous iteration, the loss function comprises two components $\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_{\text{eq}}\mathcal{L}_{\text{eq}}$. The first term is the self-supervised reconstruction error ensuring z retains sufficient observational information. The second term represents the physics consistency loss with coefficient λ_{eq} defined as $\mathcal{L}_{\text{eq}} = \|\mathcal{F}(z) - f(z)\|^2$, where $\mathcal{F}(z)$ denotes the numerical derivative via central differences and $f(z)$ represents the symbolic derivative derived from equation f. This constraint forces the latent space to adhere to the governing equation f, guiding the encoder to focus on physical dynamics while filtering out textural details and noise. Second, in the absence of physical information, such as during the cold-start phase, the autoencoder performs only self-supervised reconstruction and functions as a standard autoencoder. This approach ensures the Variable Agent identifies a latent space $\mathcal{Z}$ that balances visual reconstruction with dynamic simplicity. Crucially, it enables the output from the downstream Equation Agent to iteratively refine the extraction of physical variables. Finally, the process concludes by returning the extracted variables $z(t)$ alongside a report $\mathcal{R}_{\text{var}}$. This report encapsulates the specific tools employed and their hyperparameter configurations, such as the model size of SAM and its mask size.

4.4. Equation Agent: Dynamic Symbolic Regression

Considering that most true equations $\frac{dz}{dt} = f(z(t))$ exhibit sparsity within a high-dimensional space of candidate functions [6, 8], the agent identifies the sparse active terms in this function space through a three-step process.

Data and Derivative Estimation. Given the discrete time series $z(t)$ organized into a state matrix Z, the agent first estimates the time derivative $\dot{Z}$ numerically. This step utilizes the central difference method [40] $\mathcal{F}(z)$ to ensure methodological consistency, thereby establishing the left-hand side (LHS) of the target equation.

Candidate Library Construction. Subsequently, the agent constructs a candidate library matrix $\Theta(Z)$, wherein each column represents a potential nonlinear function of the state $z(t)$ that constitutes the right-hand side (RHS) of the governing equation. The candidate library incorporates the following component terms

- Polynomial terms including $1, z, z^2, z_1z_2, \ldots$
- Transcendental terms $\sin(z), \cos(z), exp(z), \ldots$.

The equation agent iteratively optimizes the candidate library in the following process.

Sparse Regression and Law Distillation. The discovery task is formulated as solving the overdetermined linear system $\dot{Z} \approx \Theta(Z)\Xi$, where $\dot{Z}$ denotes the derivative matrix, $\Theta(Z)$ represents the candidate library, and Ξ is the unknown sparse coefficient matrix. The optimization objective is formulated as $\|\dot{Z} - \Theta(Z)\Xi\|_2^2 + \lambda_{sp}\|\Xi\|_1$, wherein the second term enforces the sparsity of active terms. To solve for Ξ, the agent employs the Sequential Thresholded Least-Squares (STLSQ) algorithm [6] (detailed in Appendix C.2). The sparsity threshold λ_{sp} is determined via the Plan Agent's instruction $\mathcal{I}_{\text{plan}}$, enabling active guidance over the parsimony of the discovered law. Ultimately, the non-zero entries in $\Xi = [\xi_1, \xi_2, \ldots, \xi_d]$ are reconstructed into symbolic equations f.

Finally, the process concludes by returning the discovered equation f and a report $\mathcal{R}_{\text{equ}}$, containing the candidate library and sparsity threshold λ_{sp} for the decision of the Plan Agent.

4.5. Experiment Agent: Equation Evaluation and Feedback

The Experiment Agent's role is to rigorously validate the quality and consistency of the $(\mathcal{Z}, f)$ pair discovered in the current iteration. It receives the variable time series $z(t)$ and the symbolic law f and generates a multi-dimensional evaluation report, $\mathcal{R}_{\text{exp}}$, for the Plan Agent. This validation protocol assesses three key aspects.

Equation Quality. To assess the governing law f, Experiment Agent computes two quantitative metrics: (1) R^2 score by comparing the numerical derivative $\mathcal{F}(z)$ with the symbolic derivative $f(z)$; and (2) Complexity, measured by the number of terms, L_0, of the coefficient matrix Ξ.

Variable Quality. The Experiment Agent generates phase portraits of the trajectory $z(t)$ to visually characterize the variable space. Concise governing laws typically manifest as structured chaotic attractors, whereas noisy or tangled trajectories suggest the presence of significant interference. This image is passed to the Plan Agent, which visually assesses whether the trajectory exhibits a clear, low-dimensional dynamical structure.

Extrapolation Fidelity. The agent performs a long-term numerical simulation $z_{\text{pred}}(t)$ by integrating the discovered law f from an initial condition $z(0)$. It then (1) computes the Root Mean Square Error (RMSE) between the predicted $z_{\text{pred}}(t)$ and the extracted variables from held-out future frames $z_{\text{gt}}(t)$ over an unseen time horizon; and (2) visualizes the plots of the predicted and ground-truth trajectories. A self-consistent pair $(\mathcal{Z}, f)$ yields reliable long-term predictions.

Finally, the Experiment Agent aggregates all these quantitative metrics (R^2, L_0, RMSE) and plotted figures into the structured report $\mathcal{R}_{\text{exp}}$ (see Appendix C.3 for structure),

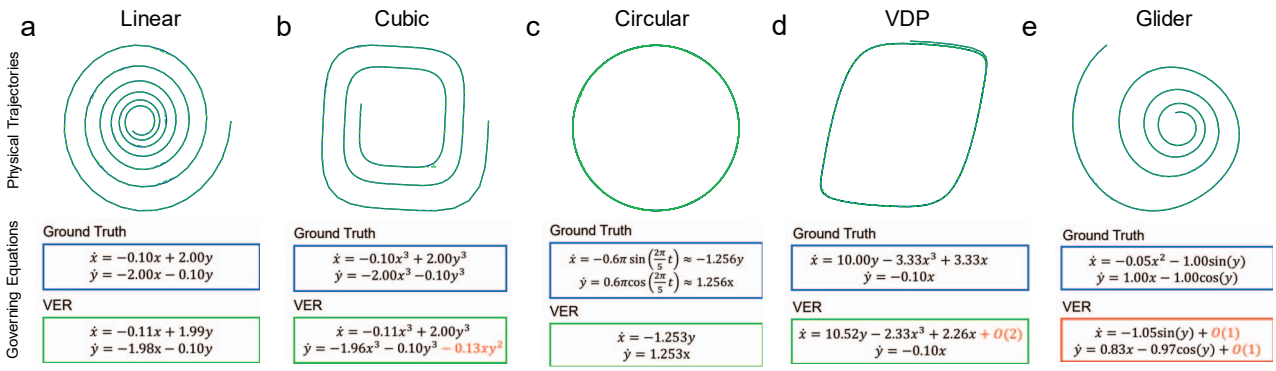


Figure 3. Reasoning results of motion objects: blue line is ground truth; green dashed line is trajectories inferred by PixelsPhys.

Table 1. Average performance of motion objects over 5 runs with varying seeds. Best results are in bold.

Case	Method	Terms Found	False Positives	R^2@1000
Linear	AE-SINDy	-	-	$0.0046_{\pm 0.0028}$
	Latent-ODE	-	-	$0.0154_{\pm 0.0081}$
	Coord-Equ	Yes	$1.100_{\pm 0.3600}$	$0.8647_{\pm 0.0554}$
	Pixel2Phys	Yes	$\mathbf{0}$	$\mathbf{0.9913_{\pm 0.0000}}$
Cubic	AE-SINDy	-	-	$0.0720_{\pm 0.0122}$
	Latent-ODE	-	-	$0.0039_{\pm 0.0013}$
	Coord-Equ	No	$3.400_{\pm 1.2800}$	$0.2632_{\pm 0.1928}$
	Pixel2Phys	Yes	$\mathbf{0.3900_{\pm 0.3620}}$	$\mathbf{0.9886_{\pm 0.0082}}$
Circular	AE-SINDy	-	-	$0.3647_{\pm 0.0716}$
	Latent-ODE	-	-	$0.0240_{\pm 0.0028}$
	Coord-Equ	Yes	$0.2000_{\pm 0.0040}$	$0.9903_{\pm 0.0057}$
	Pixel2Phys	Yes	$\mathbf{0}$	$\mathbf{1.0000_{\pm 0.0000}}$
VDP	AE-SINDy	-	-	$0.2483_{\pm 0.0182}$
	Latent-ODE	-	-	$0.0433_{\pm 0.0102}$
	Coord-Equ	Yes	$2.3100_{\pm 0.6590}$	$0.4920_{\pm 0.1302}$
	Pixel2Phys	Yes	$\mathbf{0.9900_{\pm 0.0030}}$	$\mathbf{0.9954_{\pm 0.0047}}$
Glider	AE-SINDy	-	-	$0.0310_{\pm 0.0030}$
	Latent-ODE	-	-	$0.0360_{\pm 0.0041}$
	Coord-Equ	No	$\mathbf{2.1800_{\pm 0.4900}}$	$0.9129_{\pm 0.0102}$
	Pixel2Phys	No	$3.0000_{\pm 0.0000}$	$\mathbf{0.9995_{\pm 0.0000}}$

which serves as the foundation for the Plan Agent's next reasoning step.

5. Experiments

We validate Pixel2Phys on three categories of interdisciplinary videos: object motion (object-level), physical fields (pixel-level), and real-world recordings of scientific phenomena (representation-level). We employ GPT-4o as LLM backbone by default. All configurations for Pixel2Phys across experiments are provided in Appendix E.2.

5.1. Discovery of Object-level Dynamics

Consistent with prior works [22, 34, 46], we utilize videos of object motion to assess the model's ability to accurately derive symbolic motion equations from high-dimensional observations.

5.1.1. Datasets

We validate on five dynamical systems (Figure 3): *Linear*, *Cubic*, *Circular*, *Van Der Pol (VDP)*, and *Glider*. These equations include various characteristics such as linear terms, nonlinear terms, and significant differences in time scales [15], which together form an evaluation benchmark across different levels of difficulty. Video generation details are provided in Appendix B.

5.1.2. Setups

We categorize the baselines into two classes based on interpretability. The first class encodes latent vectors and learns the implicit dynamics, including AE-SINDy [8] and Latent-ODE [10]. The second class explicitly predicts object coordinates to infer their governing equations, represented by Coord-Equ [34]. We train on the initial 200 steps and evaluate using the R^2 score on 1,000-step extrapolated coordinate trajectories. For AE-SINDy and Latent-ODE, we align their latent predictions to the ground-truth coordinates via Procrustes analysis before evaluation (Appendix D.1). Additionally, for methods capable of inferring symbolic expressions, we evaluate two equation-related metrics: 1) Terms Found: A binary score (Yes/No) indicating if all true terms are correctly identified. 2) False Positives: The number of incorrect terms included in the final equation, measuring parsimony.

5.1.3. Main Results

Table 1 and Figure 3 reveal three key insights. First, implicit methods (Latent-ODE and AE-SINDy) suffer from extrapolation collapse ($R^2 \approx 0$). This empirically validates the necessity of our multi-granularity design, demonstrating that generic representations fail to capture rigid-body physics without explicit object-level parsing. Second, Pixel2Phys demonstrates superior parsimony and robustness. Compared to Coord-Equ, our framework significantly reduces false positives and achieves near-perfect long-term prediction, with all discovered symbolic expressions de-

Table 2. Average RMSE and VPS (± std from 5 runs) of long-term prediction. Best results are in bold. The threshold of VPS is 0.5.

Method	Lambda-Omega		Brusselator		FitzHugh–Nagumo		Swift–Hohenberg	
	RMSE ↓	VPS@0.5 ↑	RMSE ↓	VPS@0.5 ↑	RMSE ↓	VPS@0.5 ↑	RMSE ↓	VPS@0.5 ↑
Black-box Models								
FNO [31]	$0.68_{\pm 0.04}$	$477.00_{\pm 20.82}$	$415.34_{\pm 81.70}$	$19.10_{\pm 3.96}$	$0.89_{\pm 0.07}$	$116.20_{\pm 6.81}$	$11.02_{\pm 4.09}$	$52.10_{\pm 9.04}$
UNO [37]	$0.42_{\pm 0.05}$	$764.80_{\pm 16.97}$	$423.64_{\pm 82.42}$	$27.30_{\pm 4.23}$	$67.41_{\pm 5.26}$	$104.00_{\pm 13.73}$	$0.48_{\pm 0.08}$	$90.80_{\pm 5.81}$
WNO [45]	$96.98_{\pm 6.35}$	$41.20_{\pm 6.10}$	$68.67_{\pm 7.80}$	$9.60_{\pm 3.67}$	$34.10_{\pm 3.09}$	$22.60_{\pm 5.24}$	$1.95_{\pm 0.39}$	$31.30_{\pm 8.93}$
Symbolic-regression Models								
PDE-Find [38]	$0.67_{\pm 0.00}$	$492.00_{\pm 0.00}$	$1.56_{\pm 0.00}$	$40.00_{\pm 0.00}$	$0.63_{\pm 0.00}$	$54.00_{\pm 0.00}$	$0.19_{\pm 0.00}$	$200.00_{\pm 0.00}$
SGA-PDE [11]	$0.92_{\pm 0.10}$	$126.60_{\pm 6.51}$	$0.14_{\pm 0.02}$	$1000.00_{\pm 0.00}$	NaN	NaN	NaN	NaN
LLM-PDE [13]	$0.55_{\pm 0.04}$	$438.40_{\pm 23.20}$	NaN	NaN	$0.62_{\pm 0.12}$	$55.90_{\pm 4.20}$	$0.69_{\pm 0.13}$	$34.30_{\pm 13.03}$
Pixel2Phys	$\mathbf{0.03}_{\pm \mathbf{0.00}}$	$\mathbf{1000.00}_{\pm \mathbf{0.00}}$	$\mathbf{0.12}_{\pm \mathbf{0.01}}$	$\mathbf{1000.00}_{\pm \mathbf{0.00}}$	$\mathbf{0.16}_{\pm \mathbf{0.01}}$	$\mathbf{1000.00}_{\pm \mathbf{0.00}}$	$\mathbf{0.18}_{\pm \mathbf{0.06}}$	$\mathbf{200.00}_{\pm \mathbf{0.00}}$

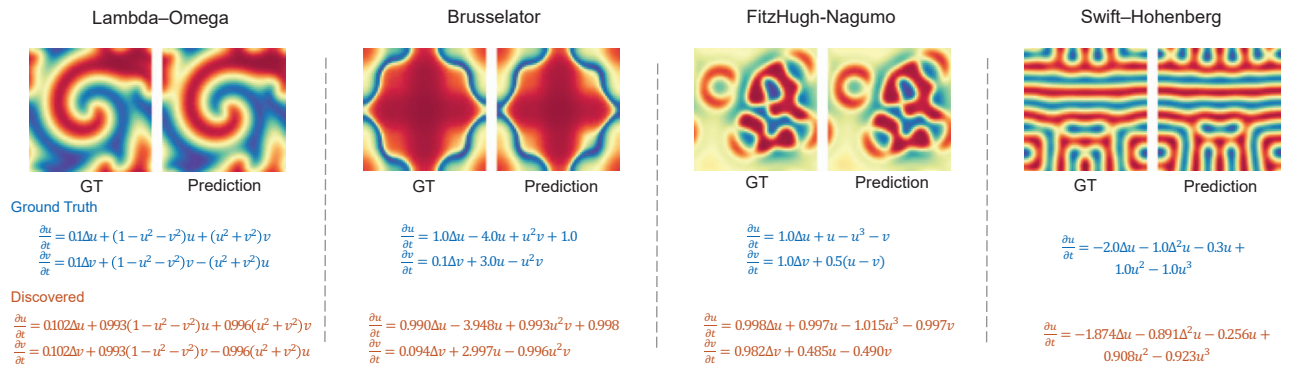


Figure 4. Reasoning results of physical fields: equations in blue is ground truth; equations in orange is inferred by PixelsPhys.

tailed in Appendix F.1. Finally, the *Glider* case highlights Pixel2Phys's practicality: although the exact symbolic matching is not achieved due to complex trigonometric terms, the near-perfect extrapolation ($R^2 = 0.9995$) and visual overlap in Figure 3e confirm that Pixel2Phys successfully captured the underlying physical attractor. Furthermore, Pixel2Phys transcends simulated settings, successfully recovering gravitational laws from real-world videos, despite complex backgrounds and noisy tracking (see Appendix F.4).

5.2. Discovery of Pixel-level Dynamics

In these scenarios, videos represent discrete grid samplings of time-varying fields driven by PDEs [8, 38]. The objective is to capture these pixel-level interactions to derive the underlying PDE equations.

5.2.1. Datasets

We conduct experiments on four representative reaction-diffusion equations: Lambda–Omega (LO) [8], Brusselator (Bruss) [33], FitzHugh–Nagumo (FHN) [47] and Swift–Hohenberg (SH) [41]. We solve the differential equations numerically as datasets. The details of the data generation are provided in Appendix B.

5.2.2. Setups

Baselines are categorized into black-box neural operators and symbolic regression methods. Models predict evolution over 1,000 steps (200 for SH system) from initial frames. We evaluate performance using root mean square error (RMSE) and valid prediction steps (VPS), defined as the duration where prediction error remains below a specific threshold (details in Appendix E).

5.2.3. Main Results

Table 2 reveals that black-box neural operators suffer from severe error accumulation in long-term rollout and result in extremely low valid prediction steps, confirming that implicit approximations fail to maintain dynamical stability without physical constraints. Existing symbolic baselines also show significant fragility as SGA-PDE and LLM-PDE frequently fail to converge or yield suboptimal fits marked as NaN. SGA tends to overfit due to the unconstrained search space of genetic algorithms while LLM-PDE lacks visual perception and biases towards over-simplified expressions that miss accurate terms as detailed in the full equation list in Appendix F.1. In contrast, Pixel2Phys consistently achieves the lowest RMSE and near-perfect stability across all datasets by integrating precise numerical tools within a reasoning loop. Figure 4 further demonstrates that our framework correctly identifies complex high-order

operators like the bi-harmonic term to capture the exact governing mechanism. Moreover, our framework scales to complex real-world PIV datasets, accurately recovering 2D Navier-Stokes components (detailed in Appendix F.4).

5.3. Discovery of Representation-level Dynamics

This category involves real-world scientific recordings, which suffer from low signal-to-noise ratios due to uncontrolled lighting and sensor noise. Consequently, effective physical components are embedded implicitly. The goal is to discover a compact representation space that filters visual noise to capture these implicit evolution mechanisms.

5.3.1. Datasets

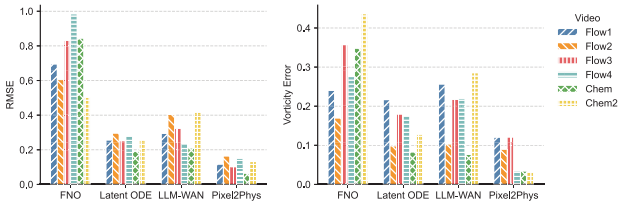
We collect six videos: four visualizing Kármán vortex streets (fluid dynamics) [35, 42] and two recording Belousov-Zhabotinsky reactions (chemical oscillators) [21]. Both represent canonical complex systems governed by low-dimensional attractors but manifested through high-dimensional, noisy visual patterns. All videos are cropped and converted to grayscale (details in Appendix B).

5.3.2. Setups

We benchmark against FNO, Latent-ODE, and the advanced video generation model Wan2.2 [48]. Given the limited sequence length (fewer than 300 frames), models are trained on the full sequence and evaluated on their ability to autoregressively reconstruct the entire video from the first frame. For Wan2.2, we freeze pretrained weights and use GPT-4o to generate descriptive text prompts for conditioning. In addition to RMSE, we also used vorticity error [24] to evaluate the accuracy of vorticity, whose formula is shown in Appendix D.2.

5.3.3. Main Results

Figure 6 presents a compelling comparison where the 14B-parameter Wan2.2 generates visually realistic textures, it fails to maintain dynamical consistency, evidenced by the spatial drift of vortices at $1.5s$ and $2.0s$ relative to the ground truth. This stems from the qualitative nature of generative prompts (see Appendix E.3), whereas Pixel2Phys distills quantitative governing laws (e.g., eigenvalues $\lambda_{2,3} \approx \pm 2.136i$ in Figure 1c) that dictate a strict oscillation period ($T \approx 0.294$) to ensure precise alignment. Visually, our predictions appear less textured, which is attributed to the selective filtering of the co-optimization mechanism. Irrelevant components like uneven lighting are discarded to extract pure dynamics on the intrinsic manifold. This physical fidelity is further corroborated by the quantitative results in Figure 5, where Pixel2Phys consistently achieves the lowest prediction error, confirming that the discovered parsimonious law successfully captures the dominant high-dimensional behavior despite the filtration of visual noise.



(a) RMSE (b) Vorticity Error

Figure 5. Comparison of prediction errors across all models on physical phenomenon videos.

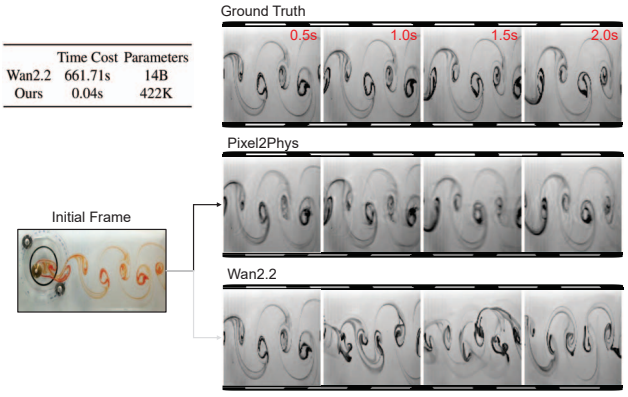


Figure 6. Prediction results of PixelsPhys and Wan2.2 for the Water Flow video.

5.4. Ablation Study and Robustness

To verify the necessity of co-optimization, we replace the Plan Agent with a static serial workflow. As detailed in Appendix F.2, the absence of the equation feedback loop results in a highly rugged variable space, failing to distill parsimonious laws. We further test robustness by substituting the LLM backbone with smaller-scale models. Results in Appendix F.3 show that Pixel2Phys maintains superior accuracy even with limited reasoning capacity, demonstrating that our collaborative agentic architecture effectively reduces the dependence on raw model scale. Finally, detailed case studies visualizing the step-by-step reasoning process for each video category are provided in Appendix F.5.

6. Conclusion

In this work, we present Pixel2Phys, a framework that automates the discovery of governing laws from visual dynamics. By coordinating specialized agents, our approach replaces static pipelines with an iterative co-optimization process. Crucially, it utilizes preliminary symbolic laws to reversely guide visual variable extraction, effectively resolving the coupling between variable extraction and law discovery. Experiments show that Pixel2Phys recovers parsimonious equations and achieves robust long-term prediction, marking a solid step towards interpretable visual modeling.

Acknowledgments

This work is supported by the Shanghai Artificial Intelligence Laboratory.

References

[1] Dimitrios Angelis, Filippos Sofos, and Theodoros E Karakasidis. Artificial intelligence in physical sciences: Symbolic regression trends and perspectives. *Archives of Computational Methods in Engineering*, 30(6):3845–3865, 2023. 1

[2] 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. 1

[3] Abeba Birhane, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. Science in the age of large language models. *Nature Reviews Physics*, 5(5):277–280, 2023. 1

[4] Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023. 1

[5] Oumayma Bounou, Jean Ponce, and Justin Carpentier. Online learning and control of complex dynamical systems from sensory input. *Advances in Neural Information Processing Systems*, 34:27852–27864, 2021. 1, 2

[6] Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15):3932–3937, 2016. 5, 7

[7] Qinglong Cao, Ding Wang, Xirui Li, Yuntian Chen, Chao Ma, and Xiaokang Yang. Teaching video diffusion model with latent physical phenomenon knowledge. *arXiv preprint arXiv:2411.11343*, 2024. 1, 4

[8] Kathleen Champion, Bethany Lusch, J Nathan Kutz, and Steven L Brunton. Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences*, 116(45):22445–22451, 2019. 3, 5, 6, 7

[9] Boyuan Chen, Kuang Huang, Sunand Raghupathi, Ishaan Chandratreya, Qiang Du, and Hod Lipson. Automated discovery of fundamental variables hidden in experimental data. *Nature Computational Science*, 2(7):433–442, 2022. 1

[10] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018. 2, 6

[11] Yuntian Chen, Yingtao Luo, Qiang Liu, Hao Xu, and Dongxiao Zhang. Symbolic genetic algorithm for discovering open-form partial differential equations (sga-pde). *Physical Review Research*, 4(2):023174, 2022. 7

[12] Jingwen Cheng, Ruikun Li, Huandong Wang, and Yong Li. Sparse diffusion autoencoder for test-time adapting prediction of complex systems. *arXiv preprint arXiv:2505.17459*, 2025. 4

[13] Mengge Du, Yuntian Chen, Zhongzheng Wang, Longfeng Nie, and Dongxiao Zhang. Large language models for automatic equation discovery of nonlinear dynamics. *Physics of Fluids*, 36(9), 2024. 7

[14] Stathi Fotiadis, Mario Lino Valencia, Shunlong Hu, Stef Garasto, Chris D Cantwell, and Anil Anthony Bharath. Disentangled generative models for robust prediction of system dynamics. In *International Conference on Machine Learning*, pages 10222–10248. PMLR, 2023. 2

[15] John Guckenheimer. Dynamics of the van der pol equation. *IEEE Transactions on Circuits and Systems*, 27(11): 983–989, 1980. 6

[16] Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11474–11484, 2020. 1, 2

[17] Irina Higgins, Peter Wirnsberger, Andrew Jaegle, and Aleksandar Botev. Symetric: Measuring the quality of learnt hamiltonian dynamics inferred from vision. *Advances in Neural Information Processing Systems*, 34:25591–25605, 2021.

[18] Florian Hofherr, Lukas Koestler, Florian Bernard, and Daniel Cremers. Neural implicit representations for physical parameter inference from a single video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2093–2103, 2023. 1, 2

[19] Peiyan Hu, Haodong Feng, Hongyuan Liu, Tongtong Yan, Wenhao Deng, Tianrun Gao, Rong Zheng, Haoren Zheng, Chenglei Yu, Chuanrui Wang, et al. Realpdebench: A benchmark for complex physical systems with real-world data. *arXiv preprint arXiv:2601.01829*, 2026. 23

[20] Yingfan Hua, Ruikun Li, Jun Yao, Guohang Zhuang, Shixiang Tang, Bin Liu, Wanli Ouyang, and Yan Lu. Finetuning large language model as an effective symbolic regressor. *arXiv preprint arXiv:2508.09897*, 2025. 1

[21] JL Hudson and JC Mankin. Chaos in the belousov–zhabotinskii reaction. *The Journal of Chemical Physics*, 74 (11):6171–6177, 1981. 8, 5

[22] Yayati Jadhav and Amir Barati Farimani. Dominant motion identification of multi-particle system using deep learning from video. *Neural Computing and Applications*, 34(20): 18183–18193, 2022. 6

[23] Miguel Jaques, Michael Burke, and Timothy Hospedales. Physics-as-inverse-graphics: Unsupervised physical parameter estimation from video. *arXiv preprint arXiv:1905.11169*, 2019. 2

[24] Jinhee Jeong and Fazle Hussain. On the identification of a vortex. *Journal of fluid mechanics*, 285:69–94, 1995. 8

[25] Rama Krishna Kandukuri, Jan Achterhold, Michael Moeller, and Joerg Stueckler. Physical representation learning and parameter identification from video using differentiable physics. *International Journal of Computer Vision*, 130(1): 3–16, 2022. 2

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 4

[27] William La Cava, Bogdan Burlacu, Marco Virgolin, Michael Kommenda, Patryk Orzechowski, Fabrício Olivetti de

França, Ying Jin, and Jason H Moore. Contemporary symbolic regression methods and their relative performance. *Advances in neural information processing systems*, 2021 (DB1):1, 2021. 1

[28] Ruikun Li, Huandong Wang, and Yong Li. Learning slow and fast system dynamics via automatic separation of time scales. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4380–4390, 2023. 1

[29] Ruikun Li, Jingwen Cheng, Huandong Wang, Qingmin Liao, and Yong Li. Predicting the dynamics of complex system via multiscale diffusion autoencoder. *arXiv preprint arXiv:2505.02450*, 2025. 4

[30] Ruikun Li, Yan Lu, Shixiang Tang, Biqing Qi, and Wanli Ouyang. Mllm-based discovery of intrinsic coordinates and governing equations from high-dimensional data. *arXiv preprint arXiv:2505.11940*, 2025. 1

[31] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020. 7

[32] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024. 1

[33] Ryan Lopez and Paul J Atzberger. Gd-vaes: Geometric dynamic variational autoencoders for learning nonlinear dynamics and dimension reductions. *arXiv preprint arXiv:2206.05183*, 2022. 7

[34] Lele Luan, Yang Liu, and Hao Sun. Distilling governing laws and source input for dynamical systems from videos. *arXiv preprint arXiv:2205.01314*, 2022. 2, 6

[35] Bernd R Noack, Konstantin Afanasiev, Marek Morzyński, Gilead Tadmor, and Frank Thiele. A hierarchy of low-dimensional models for the transient and post-transient cylinder wake. *Journal of Fluid Mechanics*, 497:335–363, 2003. 8, 5

[36] Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*, 2021. 1

[37] Md Ashiqur Rahman, Zachary E Ross, and Kamyar Azizzadenesheli. U-no: U-shaped neural operators. *arXiv preprint arXiv:2204.11127*, 2022. 7

[38] Samuel H Rudy, Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Data-driven discovery of partial differential equations. *Science advances*, 3(4):e1602614, 2017. 7

[39] Parshin Shojaee, Kazem Meidani, Amir Barati Farimani, and Chandan Reddy. Transformer-based planning for symbolic regression. *Advances in Neural Information Processing Systems*, 36:45907–45919, 2023. 1

[40] R Charles Swanson and Eli Turkel. On central-difference and upwind schemes. *Journal of computational physics*, 101 (2):292–306, 1992. 5

[41] Jack Swift and Pierre C Hohenberg. Hydrodynamic fluctuations at the convective instability. *Physical Review A*, 15(1): 319, 1977. 7, 4

[42] Kunihiko Taira, Steven L Brunton, Scott TM Dawson, Clarence W Rowley, Tim Colonius, Beverley J McKeon, Oliver T Schmidt, Stanislav Gordeyev, Vassilios Theofilis, and Lawrence S Ukeiley. Modal analysis of fluid flows: An overview. *AIAA journal*, 55(12):4013–4041, 2017. 8, 5

[43] Qwen Team. Qwen3 technical report, 2025. 22

[44] Wassim Tenachi, Rodrigo Ibata, and Foivos I Diakogiannis. Deep symbolic regression for physics guided by units constraints: toward the automated discovery of physical laws. *The Astrophysical Journal*, 959(2):99, 2023. 1

[45] Tapas Tripura and Souvik Chakraborty. Wavelet neural operator for solving parametric partial differential equations in computational mechanics problems. *Computer Methods in Applied Mechanics and Engineering*, 404:115783, 2023. 7

[46] Silviu-Marian Udrescu and Max Tegmark. Symbolic pregression: Discovering physical laws from distorted video. *Physical Review E*, 103(4):043307, 2021. 3, 6

[47] Pantelis R Vlachas, Georgios Arampatzis, Caroline Uhler, and Petros Koumoutsakos. Multiscale simulations of complex systems by learning their effective dynamics. *Nature Machine Intelligence*, 4(4):359–366, 2022. 7

[48] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 8

[49] Yiqun Wang, Nicholas Wagner, and James M Rondinelli. Symbolic regression in materials science. *MRS Communications*, 9(3):793–805, 2019. 1

[50] Hao Wu, Fan Xu, Chong Chen, Xian-Sheng Hua, Xiao Luo, and Haixin Wang. Pastnet: Introducing physical inductive biases for spatio-temporal video prediction. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 2917–2926, 2024. 4

[51] Xinheng Wu, Jie Lu, Zheng Yan, and Guangquan Zhang. Disentangling stochastic pde dynamics for unsupervised video prediction. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 1, 2

[52] Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, and Peter J Ramadge. Learning physics constrained dynamics using autoencoders. *Advances in Neural Information Processing Systems*, 35:17157–17172, 2022. 2

[53] Jie Ying, Haowei Lin, Chao Yue, Yajie Chen, Chao Xiao, Quanqi Shi, Yitao Liang, Shing-Tung Yau, Yuan Zhou, and Jianzhu Ma. A neural symbolic model for space physics. *Nature Machine Intelligence*, pages 1–16, 2025. 1

[54] Zitong Zhang, Yang Liu, and Hao Sun. Vision-based discovery of nonlinear dynamics for 3d moving target. *arXiv preprint arXiv:2404.17865*, 2024. 2