PhyMo: A Physical-Field Modality for Multimodal AI4Physics

Anonymous authors
Paper under double-blind review

Abstract

Multimodal learning is emerging as a powerful paradigm for AI for Physics (AI4Physics), where predicting physical systems requires the joint interpretation of heterogeneous observations, measurements, and domain knowledge. However, existing approaches typically represent physical quantities and governing equations as generic numerical or textual tokens, overlooking the physical constraints that determine their spatiotemporal interactions. To address this limitation, we introduce the **physical-field modality** and propose **PhyMo**, a physics-grounded multimodal framework that organizes heterogeneous measurements through PDE-associated operators. PhyMo follows a three-stage learning procedure: the physical-field encoder is first pretrained through field reconstruction under PDE residual supervision, its representations are subsequently aligned with visual embeddings in a shared latent space, and the fused multimodal representations are finally processed by corresponding downstream prediction heads. Experiments on five datasets spanning diverse physical environments show that PhyMo achieves state-of-the-art performance, compared to the strongest baseline on each dataset, demonstrating the superiority of PhyMo on multimodal representation learning in AI4Physics.

1 Introduction

Artificial intelligence (AI), and multimodal learning in particular, has emerged as a powerful paradigm for addressing complex scientific and societal challenges that require the integration of heterogeneous information sources (Yang et al. (2026)). This capability is especially valuable for AI for Physics (AI4Physics), where understanding and predicting physical-world systems often requires jointly modeling multimodal observations, measurements, and domain-specific physical knowledge (Abbas et al. (2026); Yang et al. (2026)). Recent advances in multimodal foundation models have demonstrated the potential of combining heterogeneous modalities, such as images, text, and numerical measurements, to address increasingly complex AI4Physics problems (Huang et al. (2025); Tang et al. (2026); Zou et al. (2026)). A common paradigm is to encode different sources of scientific information into a shared representation space, enabling a unified model to reason over observational data and domain knowledge. Thus, visual observations can provide rich spatial information, while textual representations can conveniently encode physical laws, mathematical formulas, scientific parameters, and numerical measurements (Zhu et al. (2026); Negrini et al. (2025)). Consequently, recent AI4Physics systems implicitly treat physical equations, parameters, and scientific data as textual or tokenized sequences, thereby enabling existing multimodal architectures to process heterogeneous scientific information.

However, directly representing physical information as text or generic tokens overlooks an essential property of physical data: physical quantities are not merely information symbols, but variables governed by physical laws. Treating them as ordinary text tokens may preserve their semantic identity while discarding the mathematical constraints that determine how they interact (Shen et al. (2025)). Taking the scenario of solar photovoltaic (PV) power forecasting as example, the data of wind speed is not simply an independent numerical feature; its spatial and temporal evolution is associated with transport and advection processes that can be described by advection-type partial differential equations (PDEs). Similarly, the propagation and spatial variation of solar irradiance are closely related to physical transport

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and diffusion processes, which can be characterized through diffusion-type PDEs. If wind speed, solar irradiance, and other physical measurements are directly serialized as text tokens and fed into a multimodal model, the model is not explicitly informed that these quantities should satisfy the corresponding physical relationships. As a result, the model may learn statistical correlations from the observed data while failing to adequately exploit the governing physical constraints (Nie et al. (2023)).

Motivated by this limitation, we introduce a new modality: **physical-field modality**, and the corresponding model **PhyMo**, which explicitly represents physical measurements according to the governing structures of physical systems rather than treating them as generic tokens. The key observation is that a broad class of physical laws can be formulated in terms of partial differential equations, providing a unified mathematical language for describing the evolution and interaction of physical fields. Based on this observation, we organize physical measurements into a set of PDEs that explicitly characterize their temporal and spatial dynamics, and introduce physics-based constraints through PDE residual losses during model training. The resulting physical-field modality enables PhyMo to learn representations that are simultaneously grounded in observational information and constrained by the underlying physical dynamics. Importantly, our formulation separates the representation of physical fields from the visual modality while retaining a unified multimodal learning framework, making the physical encoder naturally extensible to different physical systems and governing equations. This design therefore provides not only a principled mechanism for injecting physical knowledge into multimodal models, but also a scalable pathway toward incorporating diverse physical fields and PDEs into foundation-model-style architectures. By jointly exploiting visual observations and physics-grounded representations, our approach aims to learn predictive models that are both statistically accurate and physically consistent.

Our Contributions. (1) *New Physical-Field Modality.* To the best of our knowledge, this work is the first to formulate physical fields as a dedicated modality for multimodal AI4Physics, rather than treating physical measurements as generic numerical or textual inputs. By organizing heterogeneous physical quantities through PDE-associated operators, our formulation provides a unified representation that explicitly preserves the spatiotemporal structure and governing relationships of physical systems. (2) *Physics-Grounded Multimodal Framework.* We propose PhyMo, a three-stage framework consisting of physical-field encoder pretraining, multimodal alignment, and downstream prediction-head training. The physical-field encoder is pretrained through field reconstruction with PDE residual supervision to capture both physical states and governing dynamics. The learned physical embeddings are then aligned with visual representations in a shared latent space, and their fused representations are finally used by prediction heads for downstream prediction. (3) *Theoretical and Empirical Effectiveness.* We provide rigorous theoretical analyses showing the advantage of complementary multimodal information, physics-constrained representation learning, and PDE-structured physical inputs over conventional numerical or text representations. Extensive experiments further demonstrate that the proposed PhyMo consistently enhances multimodal representation quality and improves downstream forecasting performance, validating its effectiveness as a general representation paradigm for AI4Physics.

2 Related Work

Artificial intelligence has demonstrated strong potential for modeling complex physical-world systems. Pangu-Weather (Bi et al. (2023)) employs a 3D Earth-specific Transformer and hierarchical temporal aggregation to model global atmospheric evolution across multiple pressure levels. GraphCast (Lam et al. (2023)) represents the global atmosphere as a multi-scale graph and performs autoregressive message passing for medium-range weather forecasting. FuXi (Chen et al. (2023a)) adopts a cascaded forecasting architecture with specialized models for different lead-time ranges to mitigate error accumulation in long-range prediction. ClimaX (Nguyen et al. (2023)) introduces a Transformer-based foundation model with variable-specific tokenization and aggregation for transferable weather and climate modeling. NeuralGCM (Kochkov et al. (2024)) combines a differentiable atmospheric dynamical core with neural parameterizations of unresolved physical processes for weather and climate simulation. GenCast (Price et al. (2025)) formulates global probabilistic weather forecasting as conditional diffusion over atmospheric states to generate ensemble trajectories.

Aurora (Bodnar et al. (2025)) employs Perceiver-based encoders and decoders with a 3D Swin Transformer to learn transferable representations across diverse Earth-system forecasting tasks. Despite their strong predictive capability, these approaches predominantly operate on a single structured physical-data representation and do not explicitly exploit complementary modalities, limiting their ability to learn richer cross-modal representations of physical systems.

More recently, multimodal learning has been introduced to physical-science problems by jointly modeling complementary observations and scientific knowledge. AstroCLIP (Parker et al. (2024)) independently pretrains Transformer-based encoders for galaxy images and optical spectra, and subsequently aligns the two modalities through contrastive learning to obtain a shared representation for redshift estimation, physical-property prediction, and morphology classification. Maven (Zhang et al. (2024)) uses modality-specific Transformer encoders and contrastive learning to align supernova photometry and spectroscopy within a shared representation space. MultiMat (Moro et al. (2025)) jointly aligns crystal structures, density of states, charge-density fields, and textual descriptions through self-supervised multimodal contrastive learning for material-property prediction and discovery. AION-1 (Parker et al. (2026)) unifies heterogeneous astronomical observations, including multiband images, spectra, and scalar measurements, through modality-specific tokenization followed by Transformer-based multimodal masked modeling for a broad range of downstream astrophysical tasks. RadarQA (He et al. (2026)) adapts multimodal large language models to weather-forecast analysis by jointly processing radar observations, forecast sequences, physical attributes, and natural-language assessment reports. LLaMA-Vision for Neutrino Classification (Sagar et al. (2026)) fine-tunes LLaMA 3.2 Vision with QLoRA to jointly process detector pixel maps and physics-informed textual prompts, enabling multimodal classification and explanation of neutrino interaction events in high-energy physics. While these methods demonstrate the benefit of multimodal fusion, physical information is generally represented as observational arrays, symbolic or textual descriptions. However, the governing physical field itself is not explicitly formulated as a dedicated modality whose representation learning is constrained by its physical laws. Consequently, cross-modal representations are primarily learned from statistical correspondence between modalities rather than explicit governing dynamics, motivating our PDE-structured physical-field modality with physics-constrained representation learning.

3 Methodology

3.1 Motivation

As discussed in Section 1, existing multimodal paradigms for AI4Physics predominantly integrate scientific information through conventional modalities, such as images and text. Although such formulations provide a convenient interface for incorporating heterogeneous data, they often treat physical measurements and equations as generic numerical or textual tokens, without explicitly preserving the physical constraints governing their spatiotemporal evolution. This motivates us to introduce a dedicated **physical-field modality**. Ideally, such a modality should satisfy two fundamental requirements. First, it should provide a general and convenient representation for heterogeneous physical measurements, allowing different physical quantities to be systematically mapped into a common representation space. Second, the learned embeddings should be explicitly grounded in the governing physical laws, such that the representation captures not only statistical correlations among observations but also the underlying physical dynamics. In other words, we seek a modality-specific representation $\mathbf{z}_{\text{phy}}$ whose encoder

$$f_{\text{phy}} : \mathcal{X}_{\text{phy}} \to \mathcal{Z}_{\text{phy}} \tag{1}$$

maps heterogeneous physical observations $\mathcal{X}_{\text{phy}}$ into a latent space $\mathcal{Z}_{\text{phy}}$ while preserving the physical constraints associated with the corresponding fields.

In machine learning, a modality is not only characterized by the raw form of its data, but by the joint design of its data organization, model architecture, training paradigm, and learning objective (Liang et al. (2024); Xu et al. (2023); Zong et al. (2024)). For example, language models organize text as discrete tokens, process them with Transformer-based

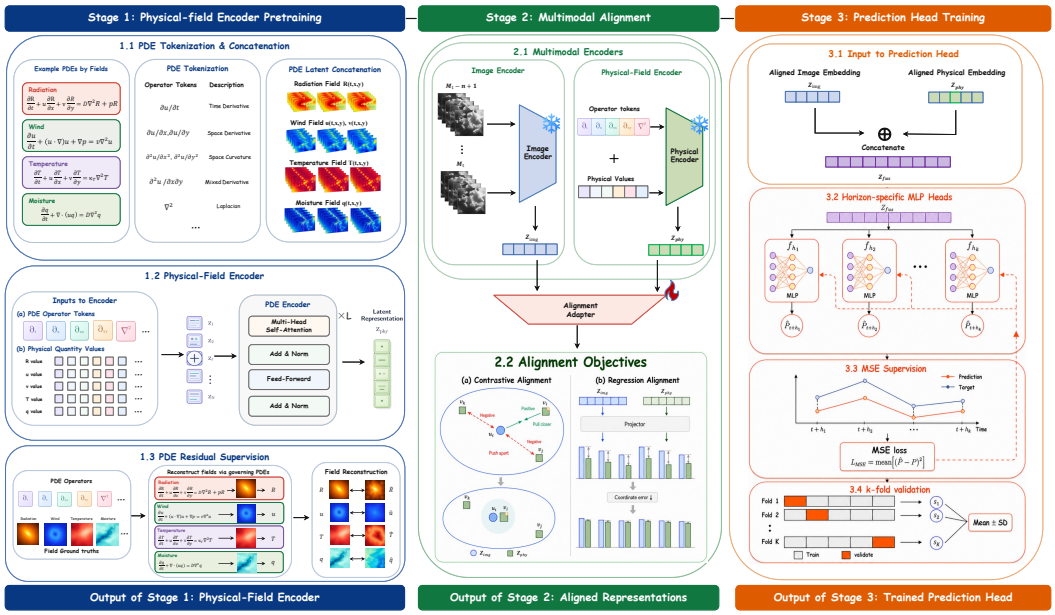


Figure 1: Overview of the proposed PhyMo. The framework consists of three stages: **Stage 1**, physical-field encoder pretraining with PDE-based tokenization and residual supervision; **Stage 2**, multimodal alignment between visual and physical-field representations; and **Stage 3**, prediction head training for downstream tasks.

architectures, and learn representations through token-prediction objectives such as next-token prediction. Note that physical measurements are fundamentally governed by physical laws, which can be expressed as partial differential equations (PDEs). Thus, PDEs provide a natural abstraction for organizing heterogeneous physical data. Accordingly, instead of treating physical measurements, such as temperature, radiation, wind and humidity, as isolated scalar inputs, we organize them as corresponding PDEs and enforce physical field reconstruction with PDE residual losses. Reconstruction preserves information about the physical state, while PDE-based supervision imposes physics-aware inductive bias on the learned representation. Under this design, physical quantities and PDE operators define the data structure, the encoder maps them into a latent space, field reconstruction defines the learning paradigm, and PDE residual loss provides the physics-grounded objective. The resulting embeddings therefore capture not only the observed physical data, but also the governing structure of the underlying physical system.

3.2 The Framework of PhyMo

Our framework of PhyMo is designed to integrate visual observations with the proposed physical-field modality for multimodal forecasting. As illustrated in Fig. 1, the overall training procedure consists of three stages: **(1) Physical-field Encoder Pretraining**, **(2) Multimodal Alignment**, and **(3) Prediction Head Training**. The first stage learns physics-aware representations from structured physical fields, the second stage aligns physical and visual representations into a compatible latent space, and the third stage performs target prediction based on the aligned multimodal embeddings.

Stage 1: Physical-field Encoder Pretraining. Given multiple physical fields, including radiation, wind, temperature, and moisture, we describe their spatiotemporal dynamics through the corresponding governing PDEs. Each physical field is represented by two complementary components: *PDE operator tokens*, which characterize the differential structure of the governing equations, and *physical quantity values*, which provide the numerical state of the physical system.

Specifically, the PDEs are decomposed into a collection of operators, such as temporal derivatives, spatial derivatives, mixed derivatives, and Laplacian operators. These oper-

ator tokens are combined with the corresponding physical measurements and fed into a Transformer-style physical-field encoder. For a physical input $\mathbf{x}_{\text{phy}}$, the encoder produces a latent representation

$$\mathbf{z}_{\text{phy}} = E_{\text{phy}}(\mathbf{x}_{\text{phy}}), \tag{2}$$

where E_{phy} denotes the physical-field encoder.

To ensure that the learned representation preserves the underlying physical structure, we train the encoder through physical-field reconstruction under PDE supervision. The latent representation is used to reconstruct the corresponding physical fields, while the reconstructed fields are constrained by their governing PDEs. Let $\hat{\mathbf{X}}_{\text{phy}}$ denote the reconstructed fields and $\mathcal{F}_m(\cdot)$ denote the governing PDE for the m-th physical field. The PDE residual supervision can be generally written as

$$\mathcal{L}_{\text{PDE}} = \sum_{m=1}^{M} \left\| \mathcal{F}_m \left(\hat{\mathbf{X}}_{\text{phy}}^{(m)} \right) \right\|_2^2. \tag{3}$$

Stage 2: Multimodal Alignment. After pretraining the physical-field encoder, we align its representation space with the visual representation space. Given a sequence of historical sky images $\mathcal{I}_t = \{I_{t-n+1}, \ldots, I_t\}$, an image encoder E_{img} (e.g., pretrained GNNs or ViTs) extracts the corresponding visual embedding $\mathbf{z}_{\text{img}} = E_{\text{img}}(\mathcal{I}_t)$. In parallel, the pretrained physical-field encoder produces $\mathbf{z}_{\text{phy}} = E_{\text{phy}}(\mathbf{x}_{\text{phy}})$. Since the two encoders are trained from fundamentally different modalities, their latent spaces are not necessarily directly compatible. We therefore introduce an alignment adapter to project the visual and physical representations into a shared latent space. Denoting the projected representations as $\tilde{\mathbf{z}}_{\text{img}}$ and $\tilde{\mathbf{z}}_{\text{phy}}$, the alignment stage optimizes two complementary objectives.

First, a contrastive objective encourages representations from matched image–physical-field pairs to remain close while separating mismatched pairs:

$$\mathcal{L}_{\text{con}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp\left(\text{sim}(\tilde{\mathbf{z}}_{\text{img}}^i, \tilde{\mathbf{z}}_{\text{phy}}^i)/\tau\right)}{\sum_{j=1}^{B} \exp\left(\text{sim}(\tilde{\mathbf{z}}_{\text{img}}^i, \tilde{\mathbf{z}}_{\text{phy}}^j)/\tau\right)}, \tag{4}$$

where B denotes the batch size, $\text{sim}(\cdot,\cdot)$ is a similarity function, and τ is the temperature parameter. Second, a regression-based alignment objective further reduces the discrepancy between paired representations:

$$\mathcal{L}_{\text{reg}} = \frac{1}{B} \sum_{i=1}^{B} \left\| \tilde{\mathbf{z}}_{\text{img}}^i - \tilde{\mathbf{z}}_{\text{phy}}^i \right\|_2^2. \tag{5}$$

The overall multimodal alignment objective is therefore

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{con}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}, \tag{6}$$

where λ_{reg} balances the two objectives.

Stage 3: Prediction Head Training. In the final stage, we use the aligned visual and physical representations for downstream tasks. The two modality embeddings are concatenated as $\mathbf{z}_{\text{fus}} = \tilde{\mathbf{z}}_{\text{img}} \oplus \tilde{\mathbf{z}}_{\text{phy}}$, where $\oplus$ denotes feature concatenation. Rather than using a single prediction function for all forecasting horizons, we employ a set of horizon-specific MLP heads $\{f_{h_1}, f_{h_2}, \ldots, f_{h_K}\}$, where each f_{h_k} is responsible for prediction at a specific future horizon h_k. The prediction at horizon h_k is given by $\hat{P}_{t+h_k} = f_{h_k}(\mathbf{z}_{\text{fus}})$. Then, the prediction heads are optimized using mean squared error between the predicted and ground-truth:

$$\mathcal{L}_{\text{pred}} = \frac{1}{K} \sum_{k=1}^{K} \left(\hat{P}_{t+h_k} - P_{t+h_k} \right)^2. \tag{7}$$

Furthermore, we perform k-fold validation during prediction-head training to evaluate model stability and select the prediction model. The validation results across different folds are aggregated to obtain the final performance estimate. Through this three-stage training procedure, the proposed framework progressively learns physics-informed representations and establishes a semantically consistent cross-modal alignment between physical fields and visual observations.

3.3 Theoretical Analysis

We provide a theoretical analysis from three complementary perspectives to characterize the role of multimodal fusion, physics-constrained representation learning, and PDE-structured physical information. Detailed assumptions and proofs are deferred to Appendix A.

We first establish that jointly exploiting visual and physical information is theoretically preferable to relying on either modality alone. Let $\mathbf{Z}_{\text{img}}$ and $\mathbf{Z}_{\text{phy}}$ denote the image and physical-field representations, respectively, and let Y denote the target.

Theorem 3.1 (Multimodal Bayes-Risk Dominance)**.** *Let*

$$\mathcal{R}^*(\mathbf{Z}) = \inf_f \mathbb{E}\left[(Y - f(\mathbf{Z}))^2\right]$$

denote the minimum achievable prediction risk given representation $\mathbf{Z}$*. Then*

$$\mathcal{R}^*\left(\mathbf{Z}_{\text{img}}, \mathbf{Z}_{\text{phy}}\right) \leq \min\left\{\mathcal{R}^*\left(\mathbf{Z}_{\text{img}}\right), \mathcal{R}^*\left(\mathbf{Z}_{\text{phy}}\right)\right\}. \tag{8}$$

Theorem 3.1 establishes an information-theoretic motivation: incorporating complementary physical information cannot increase the Bayes-optimal risk relative to either image-only or physics-only prediction, and can strictly reduce it when the two modalities contain complementary predictive information.

We next analyze whether the PDE supervision in Stage 1 can improve the quality of the learned physical representation. Let $\mathbf{X}^*_{\text{phy}}$ denote the underlying physical fields and define the physically admissible set as $\mathcal{S} = \left\{\mathbf{X} : \mathcal{F}_m\left(\mathbf{X}^{(m)}\right) = 0,\ \forall m = 1, \ldots, M\right\}$. Stage 1 optimizes the reconstruction objective together with $\mathcal{L}_{\text{PDE}} = \sum_{m=1}^{M} \left\| \mathcal{F}_m\left(\hat{\mathbf{X}}^{(m)}_{\text{phy}}\right)\right\|_2^2$.

Theorem 3.2 (Reconstruction-Risk Advantage of Physics-Constrained Representations)**.** *Assume that the governing PDEs are correctly specified such that* $\mathbf{X}^*_{\text{phy}} \in \mathcal{S}$*, and that* $\mathcal{S}$ *is a nonempty closed convex set. For an arbitrary unconstrained reconstruction* $\tilde{\mathbf{X}}_{\text{phy}}$*, let the physics-constrained reconstruction be* $\hat{\mathbf{X}}^*_{\text{phy}} = \Pi_{\mathcal{S}}\left(\tilde{\mathbf{X}}_{\text{phy}}\right)$*, where* $\Pi_{\mathcal{S}}$ *denotes the Euclidean projection onto* $\mathcal{S}$*. Then*

$$\mathbb{E}\left[\mathcal{L}_{\text{rec}}\left(\mathbf{Z}^*_{\text{phy}}\right)\right] \leq \mathbb{E}\left[\mathcal{L}_{\text{rec}}\left(\mathbf{Z}\right)\right], \tag{9}$$

for the corresponding unconstrained representation $\mathbf{Z}$*.*

Finally, we distinguish the proposed physical-field modality from directly feeding physical measurements as ordinary numerical features.

Theorem 3.3 (Theoretical Advantage of the Physical-Field Modality over Numerical-only Inputs)**.** *Let* $\mathbf{X}$ *denote the observed physical quantities and let* $\mathcal{D}(\mathbf{X})$ *collect the PDE-associated information used by the proposed modality in Stage 1. Then*

$$\mathcal{R}^*\left(\mathbf{X}, \mathcal{D}(\mathbf{X})\right) \leq \mathcal{R}^*\left(\mathbf{X}\right). \tag{10}$$

Theorem 3.3 provides a theoretical distinction between the proposed physical-field modality and conventional numerical inputs. For example, two physical systems may exhibit identical instantaneous measurements while following different local dynamics. Thus, state-only numerical representations cannot distinguish such cases, whereas PDE-structured representations can encode their distinct evolution. This result therefore motivates organizing physical information according to its governing structure rather than treating it as an unordered collection of scalar features.

4 Experiments

In this section, we present comprehensive experiments evaluating the proposed PhyMo framework under diverse datasets, including SKIPP'D (Nie et al. (2023)), Folsom (Pedro et al. (2019)), NREL (Hammond & Korgel (2026)), MeteoNet (Larvor et al. (2020)) and Boreas (Burnett et al. (2023)). The objective of these experiments is to address the following research questions: **Q1**: How does the proposed PhyMo framework perform compared to other AI4Physics baselines? **Q2**: To what extent do the proposed components of PhyMo

contribute to its multimodal representation capabilities on physical-world datasets? **Q3**: How sensitive is the performance of the proposed model to variations in its hyperparameters? **Q4**: What insights can be obtained from the prediction visualization of the proposed model? Due to page limit, we put part of the illustration in experiments in Appendix B.

4.1 Experiment Setup

Paradigms of Downstream Tasks. The proposed PhyMo is designed as a general multimodal representation-learning framework for AI4Physics rather than a task-specific forecasting model. The learned representations can be transferred to different downstream tasks through lightweight task-specific heads. However, due to the page limit, we have to instantiate regression as the primary downstream evaluation paradigm and consider representative tasks in the experiments, such as photovoltaic power and solar irradiance forecasting. Moreover, we conduct 5 experiments over varying random seeds and obtain the mean R^2 $\pm$ sample standard deviation, aiming for the statistically-robust evaluation.

Datasets. We evaluate our framework on five physical-world datasets containing images and complementary physical observations across diverse physical scenarios: (1) power prediction: SKIPP'D (Nie et al. (2023)); (2) radiance prediction: Folsom (Pedro et al. (2019)) and NREL (Hammond & Korgel (2026)); (3) weather forecasting: MeteoNet (Larvor et al. (2020)); (4) autonomous-driving speed prediction: Boreas (Burnett et al. (2023)). Due to page limit, we have to put the detailed introduction about datasets in Appendix B.1.

Baselines. As illustrated above, we choose regression as the paradigm of downstream tasks. Thus, we compare the performance of the proposed PhyMo against 10 baselines, including: (1) classical models: LASSO (Ranstam & Cook (2018)), RF (Pavlov (2000)) and TabM (Gorishniy et al. (2025)); (2) time-series models: LSTM (Hochreiter & Schmidhuber (1997)), TSMixer (Chen et al. (2023b)) and xLSTM (Beck et al. (2024)); and (3) the SOTA multimodal models: BiMamba (Zhang et al. (2025)), AstroCLIP (Parker et al. (2024)), Maven (Zhang et al. (2024)) and AION-1 (Parker et al. (2026)). The detailed introduction about the selected baselines can be found in Appendix B.2.

Evaluation Metrics. We evaluate forecasting performance using three widely adopted regression metrics: root mean squared error (RMSE), coefficient of determination (R^2), and mean absolute error (MAE). RMSE measures the overall prediction deviation while assigning larger penalties to large errors, MAE quantifies the average absolute discrepancy between predictions and ground-truth values, and R^2 evaluates the proportion of target variance explained by the model, with higher values indicating better predictive performance. For readability, we report R^2 in percentage form, i.e., $100 \times R^2$, in all experimental tables. The detailed illustration can be found in Appendix B.3.

Environments. For reproducibility, we report the hardware and software configurations used in our experiments. All experiments were conducted on a server equipped with an Intel(R) Xeon(R) Gold 6240 CPU @ 2.60GHz and a high-performance GPU with 80GB memory.

4.2 Performance Comparison

Table 1 shows that PhyMo achieves SOTA performance on SKIPP'D, Folsom, NREL, and MeteoNet, while ranking second on Boreas with performance close to the best-performing method. This consistent competitiveness across solar forecasting, meteorological prediction, and autonomous-driving scenarios demonstrates that PhyMo is not tailored to a specific task or data configuration. Instead, it provides a general framework for integrating visual observations with heterogeneous physical measurements. The results of Table 1 illustrate that the advantages of PhyMo are particularly evident on SKIPP'D, NREL and MeteoNet, which contain multiple interacting physical variables and spatially structured observations. This supports our central motivation illustrated in Section 3.1: organizing heterogeneous measurements as the physical-field modality and grounding their representations through field reconstruction and PDE residual supervision can capture the governing dependencies that are difficult to learn. Due to the facts that prior work only treat physical measurements as numerical text and subsequently ignore the governing laws behind them, they can only achieve sub-optimal performance compared to the proposed PhyMo.

Table 1: Experimental comparison results across datasets.

Dataset	LASSO	RF	LSTM	TabM	TSMixer	xLSTM	BiMamba	AstroCLIP	Maven	AION-1	**PhyMo (ours)**
SKIPP'D	$82.61_{\pm 0.01}$	$83.44_{\pm 0.03}$	$83.65_{\pm 0.48}$	$83.82_{\pm 0.26}$	$83.69_{\pm 0.42}$	$83.65_{\pm 0.30}$	$83.13_{\pm 0.56}$	$82.62_{\pm 0.31}$	$75.86_{\pm 1.68}$	$77.94_{\pm 1.55}$	$\mathbf{84.80_{\pm 0.14}}$
Folsom	$94.44_{\pm 0.03}$	$95.27_{\pm 0.05}$	$95.17_{\pm 0.03}$	$95.21_{\pm 0.02}$	$95.18_{\pm 0.02}$	$95.17_{\pm 0.02}$	$95.18_{\pm 0.03}$	$94.63_{\pm 0.03}$	$92.19_{\pm 0.09}$	$94.44_{\pm 0.03}$	$\mathbf{95.35_{\pm 0.03}}$
NREL	$73.40_{\pm 0.01}$	$74.98_{\pm 0.01}$	$77.37_{\pm 0.03}$	$76.23_{\pm 0.02}$	$75.50_{\pm 0.04}$	$77.33_{\pm 0.08}$	$77.33_{\pm 0.06}$	$77.38_{\pm 0.03}$	$72.67_{\pm 0.50}$	$73.72_{\pm 0.21}$	$\mathbf{79.31_{\pm 0.26}}$
MeteoNet	$96.38_{\pm 0.06}$	$93.38_{\pm 0.06}$	$97.65_{\pm 0.06}$	$97.61_{\pm 0.04}$	$97.64_{\pm 0.01}$	$97.78_{\pm 0.01}$	$97.76_{\pm 0.03}$	$97.45_{\pm 0.03}$	$91.23_{\pm 0.58}$	$93.81_{\pm 0.22}$	$\mathbf{98.36_{\pm 0.07}}$
Boreas	$96.08_{\pm 2.01}$	$13.11_{\pm 5.23}$	$92.77_{\pm 5.98}$	$\mathbf{96.95_{\pm 2.82}}$	$96.48_{\pm 1.85}$	$96.07_{\pm 2.23}$	$96.24_{\pm 1.25}$	$38.81_{\pm 11.55}$	$29.01_{\pm 2.71}$	$60.59_{\pm 6.99}$	$96.84_{\pm 1.28}$

Table 2: Ablation analysis of the proposed model on SKIPP'D dataset.

Experiment	Setting	Configuration	R^2
Modality	Both	Both modalities	$\mathbf{84.80_{\pm 0.14}}$
	w/o field	Field modality removed	$82.90_{\pm 0.21}$
	w/o image	Image modality removed	$83.45_{\pm 0.14}$
Fusion	EF	Early fusion	$\mathbf{84.80_{\pm 0.14}}$
	LF	Late fusion	$83.94_{\pm 0.13}$

4.3 Ablation Analysis

In order to answer **Q2**, we further conduct the ablation analysis to the proposed PhyMo, as shown in Table 2. As shown in Table 2, jointly incorporating the image and physical-field modalities yields an R^2 of 84.80, outperforming both unimodal variants. Removing the physical-field modality decreases performance to 82.90, while removing the image modality results in 83.45. These consistent degradations demonstrate that the two modalities provide complementary information: images capture the observable appearance, whereas the physical-field modality represents the underlying complex physical dynamics of real world scenarios. The larger degradation caused by removing the physical field further indicates that its physics-grounded representation contributes information that cannot be fully recovered from visual observations alone. The fusion ablation further shows that early fusion achieves 84.80, compared with 83.94 for late fusion. This result suggests that representation-level interaction is important for exploiting cross-modal dependencies. Early fusion (i.e., EF in Table 2) allows visual patterns to be jointly interpreted with their corresponding physical states before prediction, thereby facilitating a unified representation of observable phenomena and governing dynamics. By contrast, late fusion (i.e., LF in Table 2) combines modality-specific predictions only after the relevant representations have been learned independently, limiting fine-grained cross-modal interaction.

4.4 Hyperparameter Sensitivity Analysis

In order to answer **Q3**, we also conduct the hyperparameter sensitivity analysis to the proposed model, where the image span means the number of images fed into the model during the training and the reconstruction horizon means the horizons reconstructed by the model in Stage 1, as shown in Table 3. It shows that PhyMo is generally robust to the examined hyperparameters, while favoring moderate configurations. For the cross-modal embedding size, the performance remains within a narrow range across all settings, as well as the image span. For the reconstruction horizon, performance improves as the reconstruction set expands from K1 to K3 and remains comparable under K4 and K6, demonstrating that multi-horizon reconstruction provides richer temporal supervision but exhibits diminishing returns once the principal temporal scales are covered, which may be contributed to the fact that PV forecasting in SKIPP'D is more prone to the short-period histories instead of the long-term ones. Overall, PhyMo is not highly sensitive to small hyperparameter variations, although compact cross-modal representations, recent visual context, and a moderately diverse reconstruction horizon offer the most effective balance.

4.5 Prediction Visualization

In order to answer **Q4**, we visualize the prediction results of PhyMo on SKIPP'D, as shown in Fig. 2. The visualization reveals a clear dependence of forecasting behavior on weather-induced variability. Under sunny conditions, PhyMo closely tracks the ground-truth curves throughout the day, accurately recovering both the smooth diurnal evolution and the

Table 3: Sensitivity analysis of the proposed model on the SKIPP'D dataset.

(a) Cross-modal embedding size

EMB	Rep. width	H30	H60	H120	Mean
64	192	$87.23_{\pm 0.33}$	$83.20_{\pm 0.43}$	$77.13_{\pm 0.82}$	$84.71_{\pm 0.33}$
96	288	$87.20_{\pm 0.17}$	$83.16_{\pm 0.36}$	$77.35_{\pm 0.24}$	$\mathbf{84.80_{\pm 0.14}}$
160	480	$86.88_{\pm 0.38}$	$83.05_{\pm 0.29}$	$76.63_{\pm 0.47}$	$84.49_{\pm 0.18}$
192	576	$91.51_{\pm 0.15}$	$83.36_{\pm 0.55}$	$76.26_{\pm 0.54}$	$84.67_{\pm 0.26}$
256	768	$87.13_{\pm 0.19}$	$83.27_{\pm 0.29}$	$76.66_{\pm 0.56}$	$84.68_{\pm 0.28}$

(b) Image span

Setting	Span (frames)	H30	H60	H120	Mean
C5	6	$87.17_{\pm 0.23}$	$83.47_{\pm 0.52}$	$77.14_{\pm 0.51}$	$\mathbf{84.80_{\pm 0.14}}$
C15	16	$87.26_{\pm 0.29}$	$83.13_{\pm 0.52}$	$76.98_{\pm 1.09}$	$84.72_{\pm 0.21}$
C30	31	$87.22_{\pm 0.33}$	$82.98_{\pm 0.40}$	$77.08_{\pm 0.97}$	$84.72_{\pm 0.44}$
C45	46	$87.03_{\pm 0.43}$	$82.53_{\pm 0.82}$	$76.16_{\pm 0.74}$	$84.29_{\pm 0.40}$
C60	61	$86.85_{\pm 0.25}$	$82.58_{\pm 0.26}$	$76.73_{\pm 0.99}$	$84.34_{\pm 0.23}$

(c) Reconstruction horizon

Setting	Reconstruction set (min)	H30	H60	H120	Mean
K1	$\{5\}$	$87.08_{\pm 0.39}$	$83.17_{\pm 0.51}$	$76.50_{\pm 0.59}$	$83.52_{\pm 0.37}$
K2	$\{5, 15\}$	$86.95_{\pm 0.30}$	$82.83_{\pm 0.84}$	$76.87_{\pm 0.82}$	$83.85_{\pm 0.51}$
K3	$\{5, 15, 30\}$	$87.25_{\pm 0.25}$	$83.17_{\pm 0.38}$	$76.70_{\pm 0.77}$	$\mathbf{84.80_{\pm 0.14}}$
K4	$\{5, 15, 20, 30\}$	$87.21_{\pm 0.19}$	$83.08_{\pm 0.52}$	$76.26_{\pm 0.64}$	$84.75_{\pm 0.26}$
K6	$\{5, 10, 15, 20, 25, 30\}$	$86.88_{\pm 0.50}$	$83.03_{\pm 0.64}$	$76.78_{\pm 0.70}$	$84.78_{\pm 0.47}$

timing and magnitude of peak PV output, as reflected by consistently low RMSE and MAE values. Under cloudy conditions, the model still captures the overall generation envelope and major ramping events, including changes in the onset, peak, and decay of PV output. However, the errors increase as the ground truth becomes more volatile, particularly in periods characterized by abrupt, short-lived fluctuations.

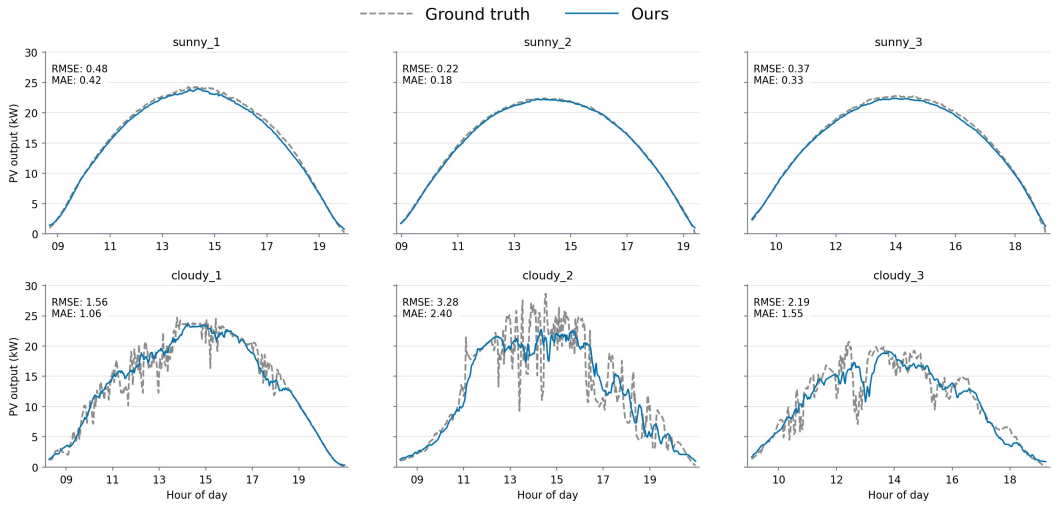


Figure 2: The prediction visualization of the proposed PhyMo on SKIPP'D.

5 Conclusion

This work introduced the **physical-field modality** and **PhyMo**, a physics-grounded multimodal framework for representing physical measurements according to their governing dynamics rather than as generic numerical or textual tokens. PhyMo organizes heterogeneous physical quantities through PDE-associated operators and adopts a three-stage learning procedure comprising physical-field encoder pretraining, cross-modal alignment, and downstream prediction. Our theoretical analyses demonstrate the benefits of complementary multimodal information, physics-constrained learning, and PDE-structured inputs. Across five datasets covering diverse physical environments, PhyMo achieves SOTA performance, compared to the strongest baseline on each dataset. These results establish physical fields as a principled modality for multimodal AI4Physics and provide a scalable framework for incorporating additional physical quantities, governing equations, and scientific domains.

References

Samar Abbas, Waqas Amin, Zhu Yuyu, Alireza Soleimani, Anna Pinnarelli, Pasquale Vizza, and Mehrdad Heidari. Physics-informed multimodal large language models for intelligent energy question answering. Results in Engineering, pp. 110718, 2026.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. Advances in Neural Information Processing Systems, 37:107547–107603, 2024.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. Nature, 619(7970): 533–538, 2023.

Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A Weyn, Haiyu Dong, et al. A foundation model for the earth system. Nature, 641(8065):1180–1187, 2025.

Keenan Burnett, David J Yoon, Yuchen Wu, Andrew Z Li, Haowei Zhang, Shichen Lu, Jingxing Qian, Wei-Kang Tseng, Andrew Lambert, Keith YK Leung, et al. Boreas: A multi-season autonomous driving dataset. The International Journal of Robotics Research, 42(1-2):33–42, 2023.

Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast. npj climate and atmospheric science, 6(1):190, 2023a.

Si-An Chen, Chun-Liang Li, Nate Yoder, Sercan O Arik, and Tomas Pfister. Tsmixer: An all-mlp architecture for time series forecasting. arXiv preprint arXiv:2303.06053, 2023b.

Yury Gorishniy, Akim Kotelnikov, and Artem Babenko. Tabm: Advancing tabular deep learning with parameter-efficient ensembling. In International Conference on Learning Representations, volume 2025, pp. 77899–77935, 2025.

Joshua E. Hammond and Brian A. Korgel. Nrel srrl minute-resolution sky imagery dataset. Hugging Face Datasets, 2026. NREL Solar Radiation Research Laboratory (SRRL) Sky Imagery Dataset.

Xuming He, Zhiyuan You, Junchao Gong, Couhua Liu, Xiaoyu Yue, Peiqin Zhuang, Wenlong Zhang, and Lei Bai. Radarqa: Multi-modal quality analysis of weather radar forecasts. Advances in Neural Information Processing Systems, 38:131737–131767, 2026.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9 (8):1735–1780, 1997.

Hong Huang, Ben Zhang, Jingzheng Ren, Qinglin Chen, Bingjian Zhang, Xiantai Zhou, Chang He, and Haoshui Yu. A multimodal foundation model of multi-physics for chemical reactor design. Chemical Engineering Science, pp. 122998, 2025.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, et al. Neural general circulation models for weather and climate. Nature, 632(8027):1060–1066, 2024.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. Science, 382(6677):1416–1421, 2023.

Gwennaëlle Larvor, Léa Berthomier, Vincent Chabot, Brice Le Pape, Bruno Pradel, and Lior Perez. Meteonet, an open reference weather dataset by meteo-france. 2020. URL https://www. kaggle. com/datasets/katerpillar/meteonet, 2020.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. ACM computing surveys, 56(10):1–42, 2024.

Viggo Moro, Charlotte Loh, Rumen Dangovski, Ali Ghorashi, Andrew Ma, Zhuo Chen, Samuel Kim, Peter Y Lu, Thomas Christensen, and Marin Soljačić. Multimodal foundation models for material property prediction and discovery. Newton, 1(1), 2025.

Elisa Negrini, Yuxuan Liu, Liu Yang, Stanley J Osher, and Hayden Schaeffer. A multimodal pde foundation model for prediction and scientific text descriptions. arXiv preprint arXiv:2502.06026, 2025.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. arXiv preprint arXiv:2301.10343, 2023.

Yuhao Nie, Xiatong Li, Andea Scott, Yuchi Sun, Vignesh Venugopal, and Adam Brandt. Skipp'd: A sky images and photovoltaic power generation dataset for short-term solar forecasting. Solar Energy, 255:171–179, 2023.

Liam Parker, Francois Lanusse, Siavash Golkar, Leopoldo Sarra, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Geraud Krawezik, Michael McCabe, Rudy Morel, et al. Astroclip: a cross-modal foundation model for galaxies. Monthly Notices of the Royal Astronomical Society, 531(4):4990–5011, 2024.

Liam Parker, Francois Lanusse, Jeff Shen, Ollie Liu, Tom Hehir, Leopoldo Sarra, Lucas Meyer, Micah Bowles, Sebastian Wagner-Carena, Helen Qu, et al. Aion-1: Omnimodal foundation model for astronomical sciences. Advances in Neural Information Processing Systems, 38:95386–95428, 2026.

Yu L Pavlov. Random forests. Vsp, 2000.

Hugo TC Pedro, David P Larson, and Carlos FM Coimbra. A comprehensive dataset for the accelerated development and benchmarking of solar forecasting methods. Journal of Renewable and Sustainable Energy, 11(3), 2019.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Probabilistic weather forecasting with machine learning. Nature, 637(8044):84–90, 2025.

Jonas Ranstam and Jonathan A Cook. Lasso regression. Journal of British Surgery, 105 (10):1348–1348, 2018.

Dikshant Sagar, Kaiwen Yu, Alejandro Yankelevich, Jianming Bian, and Pierre Baldi. Adapting vision-language models for neutrino event classification in high-energy physics. Communications Physics, 2026.

Yiqing Shen, Hao Ding, Lalithkumar Seenivasan, Tianmin Shu, and Mathias Unberath. Position: Foundation models need digital twin representations. arXiv preprint arXiv:2505.03798, 2025.

Yingheng Tang, Wenbin Xu, Jie Cao, Weilu Gao, Steven Farrell, Benjamin Erichson, Michael W Mahoney, Andy Nonaka, and Zhi Jackie Yao. A multimodal large language model for materials science. Nature Machine Intelligence, pp. 1–14, 2026.

Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. IEEE transactions on pattern analysis and machine intelligence, 45(10):12113–12132, 2023.

Xiaomeng Yang, Zhiyu Tan, Xiaohui Zhong, Mengping Yang, Qiusheng Huang, Lei Chen, Libo Wu, and Hao Li. A unified multimodal understanding and generation model for cross-disciplinary scientific research. arXiv preprint arXiv:2601.01363, 2026.

Gemma Zhang, Thomas Helfer, Alexander T Gagliano, Siddharth Mishra-Sharma, and V Ashley Villar. Maven: a multimodal foundation model for supernova science. Machine Learning: Science and Technology, 5(4):045069, 2024.

Qingyong Zhang, Zhuo Zeng, Quan Zhou, Xinzhi Ding, Zhengcheng Dong, Meng Tian, Dinghao Liu, Kai Mei, and Minglong Li. Multimodal bidirectional mamba framework for ultra-short-term photovoltaic power forecasting using sky images. Renewable Energy, pp. 124361, 2025.

Xiao Xiang Zhu, Zhitong Xiong, Yi Wang, Adam J Stewart, Konrad Heidler, Yuanyuan Wang, Zhenghang Yuan, Thomas Dujardin, Qingsong Xu, and Yilei Shi. On the foundations of earth foundation models. Communications Earth & Environment, 7(1):103, 2026.

Yongshuo Zong, Oisin Mac Aodha, and Timothy M Hospedales. Self-supervised multimodal learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47 (7):5299–5318, 2024.

Yicheng Zou, Dongsheng Zhu, Lin Zhu, Tong Zhu, Yunhua Zhou, Peiheng Zhou, Xinyu Zhou, Dongzhan Zhou, Zhiwang Zhou, Yuhao Zhou, et al. Intern-s1-pro: Scientific multimodal foundation model at trillion scale. arXiv preprint arXiv:2603.25040, 2026.

A Detailed Theoretical Analysis

In this section, we provide detailed proofs for the theoretical results presented in Section 3.3. We use squared prediction loss throughout the analysis, which is consistent with the MSE objective used for downstream prediction.

A.1 Preliminaries

Let $Y \in \mathbb{R}$ denote the target to be predicted. Let $\mathbf{Z}_{\text{img}}$ and $\mathbf{Z}_{\text{phy}}$ denote the learned image and physical-field representations, respectively. For an arbitrary representation $\mathbf{Z}$, define its Bayes-optimal prediction risk under squared loss as

$$\mathcal{R}^*(\mathbf{Z}) = \inf_f \mathbb{E}\left[(Y - f(\mathbf{Z}))^2\right], \tag{11}$$

where the infimum is taken over all measurable predictors f.

Under squared loss, the Bayes-optimal predictor is

$$f^*(\mathbf{Z}) = \mathbb{E}[Y \mid \mathbf{Z}], \tag{12}$$

and therefore

$$\mathcal{R}^*(\mathbf{Z}) = \mathbb{E}\left[\mathrm{Var}\left(Y \mid \mathbf{Z}\right)\right]. \tag{13}$$

This identity provides the basis for Theorems 3.1 and 3.3.

A.2 Proof of Multimodal Bayes-Risk Dominance

Proof of Theorem 3.1. Consider first the image representation $\mathbf{Z}_{\text{img}}$. By the law of total variance,

$$\begin{aligned}\mathrm{Var}\left(Y \mid \mathbf{Z}_{\text{img}}\right) = \mathbb{E}\left[\mathrm{Var}\left(Y \mid \mathbf{Z}_{\text{img}}, \mathbf{Z}_{\text{phy}}\right)|\mathbf{Z}_{\text{img}}\right] \\ + \mathrm{Var}\left(\mathbb{E}\left[Y \mid \mathbf{Z}_{\text{img}}, \mathbf{Z}_{\text{phy}}\right]|\mathbf{Z}_{\text{img}}\right).\end{aligned} \tag{14}$$

The second term on the right-hand side is non-negative. Hence,

$$\mathrm{Var}\left(Y \mid \mathbf{Z}_{\text{img}}\right) \geq \mathbb{E}\left[\mathrm{Var}\left(Y \mid \mathbf{Z}_{\text{img}}, \mathbf{Z}_{\text{phy}}\right)|\mathbf{Z}_{\text{img}}\right]. \tag{15}$$

Taking expectation over $\mathbf{Z}_{\text{img}}$ gives

$$\mathbb{E}\left[\mathrm{Var}\left(Y \mid \mathbf{Z}_{\text{img}}\right)\right] \geq \mathbb{E}\left[\mathrm{Var}\left(Y \mid \mathbf{Z}_{\text{img}}, \mathbf{Z}_{\text{phy}}\right)\right]. \tag{16}$$

Using Eq. equation 13, we obtain

$$\mathcal{R}^*\left(\mathbf{Z}_{\text{img}}, \mathbf{Z}_{\text{phy}}\right) \leq \mathcal{R}^*\left(\mathbf{Z}_{\text{img}}\right). \tag{17}$$

By symmetry, conditioning first on $\mathbf{Z}_{\text{phy}}$ yields

$$\mathcal{R}^*\left(\mathbf{Z}_{\text{img}}, \mathbf{Z}_{\text{phy}}\right) \leq \mathcal{R}^*\left(\mathbf{Z}_{\text{phy}}\right). \tag{18}$$

Combining Eqs. equation 17 and equation 18 gives

$$\boxed{\mathcal{R}^*\left(\mathbf{Z}_{\text{img}}, \mathbf{Z}_{\text{phy}}\right) \leq \min\left\{\mathcal{R}^*\left(\mathbf{Z}_{\text{img}}\right), \mathcal{R}^*\left(\mathbf{Z}_{\text{phy}}\right)\right\}.} \tag{19}$$

□

A.3 Proof of Reconstruction-Risk Advantage of Physics-Constrained Representations

We provide the detailed proof of Theorem 3.2 and establish its connection to the Stage 1 training objective.

A.3.1 Physical Admissible Set

Let

$$\mathbf{X}^*_{\text{phy}} = \left\{ \mathbf{X}^{*(1)}_{\text{phy}}, \ldots, \mathbf{X}^{*(M)}_{\text{phy}} \right\} \tag{20}$$

denote the underlying physical fields. For the m-th field, let $\mathcal{F}_m$ denote its governing PDE. We define the joint physically admissible set as

$$\mathcal{S} = \left\{ \mathbf{X} : \mathcal{F}_m \left(\mathbf{X}^{(m)} \right) = 0, \quad m = 1, \ldots, M \right\}. \tag{21}$$

We assume that the governing equations are correctly specified, such that

$$\mathbf{X}^*_{\text{phy}} \in \mathcal{S}. \tag{22}$$

We further assume that $\mathcal{S}$ is nonempty, closed, and convex, so that the Euclidean projection onto $\mathcal{S}$ is uniquely defined.

Let

$$\tilde{\mathbf{X}}_{\text{phy}} = D_{\text{phy}}(\mathbf{Z}) \tag{23}$$

denote an arbitrary unconstrained reconstruction obtained from a latent representation $\mathbf{Z}$. The corresponding physics-constrained reconstruction is defined by

$$\hat{\mathbf{X}}^*_{\text{phy}} = \Pi_{\mathcal{S}} \left(\tilde{\mathbf{X}}_{\text{phy}} \right) = \arg\min_{\mathbf{U} \in \mathcal{S}} \left\| \mathbf{U} - \tilde{\mathbf{X}}_{\text{phy}} \right\|_2^2. \tag{24}$$

A.3.2 Projection-Based Reconstruction Bound

Proof of Theorem 3.2. For any nonempty closed convex set $\mathcal{S}$, the metric projection $\Pi_{\mathcal{S}}(\mathbf{x})$ satisfies

$$\langle \mathbf{x} - \Pi_{\mathcal{S}}(\mathbf{x}), \mathbf{s} - \Pi_{\mathcal{S}}(\mathbf{x}) \rangle \leq 0, \qquad \forall \mathbf{s} \in \mathcal{S}. \tag{25}$$

Taking

$$\mathbf{x} = \tilde{\mathbf{X}}_{\text{phy}}, \qquad \mathbf{s} = \mathbf{X}^*_{\text{phy}}, \tag{26}$$

and using $\mathbf{X}^*_{\text{phy}} \in \mathcal{S}$, we obtain

$$\left\langle \tilde{\mathbf{X}}_{\text{phy}} - \hat{\mathbf{X}}^*_{\text{phy}}, \mathbf{X}^*_{\text{phy}} - \hat{\mathbf{X}}^*_{\text{phy}} \right\rangle \leq 0. \tag{27}$$

Now decompose the unconstrained reconstruction error as

$$\tilde{\mathbf{X}}_{\text{phy}} - \mathbf{X}^*_{\text{phy}} = \left(\tilde{\mathbf{X}}_{\text{phy}} - \hat{\mathbf{X}}^*_{\text{phy}} \right) + \left(\hat{\mathbf{X}}^*_{\text{phy}} - \mathbf{X}^*_{\text{phy}} \right). \tag{28}$$

Taking the squared Euclidean norm gives

$$\begin{aligned} &\left\| \tilde{\mathbf{X}}_{\text{phy}} - \mathbf{X}^*_{\text{phy}} \right\|_2^2 \\ &= \left\| \tilde{\mathbf{X}}_{\text{phy}} - \hat{\mathbf{X}}^*_{\text{phy}} \right\|_2^2 + \left\| \hat{\mathbf{X}}^*_{\text{phy}} - \mathbf{X}^*_{\text{phy}} \right\|_2^2 \\ &\quad + 2 \left\langle \tilde{\mathbf{X}}_{\text{phy}} - \hat{\mathbf{X}}^*_{\text{phy}}, \hat{\mathbf{X}}^*_{\text{phy}} - \mathbf{X}^*_{\text{phy}} \right\rangle. \end{aligned} \tag{29}$$

From Eq. equation 27,

$$\left\langle \tilde{\mathbf{X}}_{\text{phy}} - \hat{\mathbf{X}}^*_{\text{phy}}, \hat{\mathbf{X}}^*_{\text{phy}} - \mathbf{X}^*_{\text{phy}} \right\rangle \geq 0. \tag{30}$$

Therefore,

$$\begin{aligned} \left\| \tilde{\mathbf{X}}_{\text{phy}} - \mathbf{X}^*_{\text{phy}} \right\|_2^2 \geq &\left\| \tilde{\mathbf{X}}_{\text{phy}} - \hat{\mathbf{X}}^*_{\text{phy}} \right\|_2^2 \\ &+ \left\| \hat{\mathbf{X}}^*_{\text{phy}} - \mathbf{X}^*_{\text{phy}} \right\|_2^2. \end{aligned} \tag{31}$$

Since the first term on the right-hand side is non-negative,

$$\boxed{\left\|\hat{\mathbf{X}}^*_{\text{phy}} - \mathbf{X}^*_{\text{phy}}\right\|_2^2 \leq \left\|\tilde{\mathbf{X}}_{\text{phy}} - \mathbf{X}^*_{\text{phy}}\right\|_2^2 .} \tag{32}$$

Let the reconstruction loss associated with a representation $\mathbf{Z}$ be

$$\mathcal{L}_{\text{rec}}(\mathbf{Z}) = \left\|D_{\text{phy}}(\mathbf{Z}) - \mathbf{X}^*_{\text{phy}}\right\|_2^2 . \tag{33}$$

If $\mathbf{Z}^*_{\text{phy}}$ denotes a representation whose decoded field equals $\hat{\mathbf{X}}^*_{\text{phy}}$, Eq. equation 32 directly implies

$$\boxed{\mathcal{L}_{\text{rec}}\left(\mathbf{Z}^*_{\text{phy}}\right) \leq \mathcal{L}_{\text{rec}}\left(\mathbf{Z}\right).} \tag{34}$$

Since Eq. equation 34 holds pointwise, taking expectation over the data distribution preserves the inequality:

$$\boxed{\mathbb{E}\left[\mathcal{L}_{\text{rec}}\left(\mathbf{Z}^*_{\text{phy}}\right)\right] \leq \mathbb{E}\left[\mathcal{L}_{\text{rec}}\left(\mathbf{Z}\right)\right].} \tag{35}$$

□

A.4 Proof of the Advantage of PDE-Structured Physical Representations

We finally compare the proposed physical-field organization with a conventional representation that treats physical measurements only as numerical values.

Let

$$\mathbf{X} = [R, u, v, T, q, \ldots] \tag{36}$$

denote the physical quantities observed at a given state.

Let

$$\mathcal{D}(\mathbf{X}) = \left[\partial_t \mathbf{X}, \nabla \mathbf{X}, \nabla^2 \mathbf{X}, \partial_{xy} \mathbf{X}, \ldots\right] \tag{37}$$

denote the collection of the PDE operators used by the proposed physical-field modality. A conventional numerical representation uses only $\mathbf{X}$, whereas the PDE-structured representation uses

$$\mathbf{X}_{\text{PDE}} = \left(\mathbf{X}, \mathcal{D}(\mathbf{X})\right). \tag{38}$$

Proof of Theorem 3.3. Using Eq. equation 13, the minimum prediction risk based only on numerical physical values is

$$\mathcal{R}^*(\mathbf{X}) = \mathbb{E}\left[\text{Var}(Y \mid \mathbf{X})\right]. \tag{39}$$

The Bayes risk using the PDE-structured physical information is

$$\mathcal{R}^*\left(\mathbf{X}, \mathcal{D}(\mathbf{X})\right) = \mathbb{E}\left[\text{Var}\left(Y \mid \mathbf{X}, \mathcal{D}(\mathbf{X})\right)\right]. \tag{40}$$

Applying the law of total variance conditional on $\mathbf{X}$ gives

$$\begin{aligned}\text{Var}(Y \mid \mathbf{X}) = \mathbb{E}\left[\text{Var}\left(Y \mid \mathbf{X}, \mathcal{D}(\mathbf{X})\right)|\mathbf{X}\right] \\ + \text{Var}\left(\mathbb{E}\left[Y \mid \mathbf{X}, \mathcal{D}(\mathbf{X})\right]|\mathbf{X}\right).\end{aligned} \tag{41}$$

Since the second term is non-negative, we have,

$$\boxed{\mathcal{R}^*\left(\mathbf{X}, \mathcal{D}(\mathbf{X})\right) \leq \mathcal{R}^*(\mathbf{X}).} \tag{42}$$

Moreover, subtracting the two risks gives

$$\begin{aligned}&\mathcal{R}^*(\mathbf{X}) - \mathcal{R}^*\left(\mathbf{X}, \mathcal{D}(\mathbf{X})\right) \\ &= \mathbb{E}\left[\text{Var}\left(\mathbb{E}\left[Y \mid \mathbf{X}, \mathcal{D}(\mathbf{X})\right]|\mathbf{X}\right)\right].\end{aligned} \tag{43}$$

Hence, the inequality is strict whenever

$$\text{Var}\left(\mathbb{E}[Y \mid \mathbf{X}, \mathcal{D}(\mathbf{X})] \mid \mathbf{X}\right) > 0 \tag{44}$$

with positive probability.

This condition precisely means that the differential structure contains predictive information that cannot be recovered from the instantaneous numerical state alone. □

B The Detailed Illustration about the Experiment Setup

B.1 Dataset Details

We conduct experiments on five real-world datasets covering different geographic locations, atmospheric conditions, and forecasting targets. These datasets provide complementary combinations of sky imagery, irradiance measurements, photovoltaic generation, and meteorological observations, allowing us to evaluate the proposed multimodal framework under diverse physical environments.

SKIPP'D. The SKIPP'D dataset (Nie et al. (2023)) was collected at Stanford University and was designed specifically for short-term solar forecasting from ground-based sky imagery. It contains three years of observations from 2017 to 2019 and provides temporally synchronized sky images and photovoltaic (PV) power generation measurements. The benchmark version includes quality-controlled sky images downsampled to 64×64 pixels and concurrent PV generation records at a one-minute interval. In addition, the dataset releases the corresponding raw data, including high-resolution 2048×2048 sky images, high-frequency sky video recorded at 20 frames per second, and PV generation histories at one-minute resolution. This combination provides both rich visual observations of cloud evolution and direct measurements of the downstream energy-generation process, making SKIPP'D particularly suitable for evaluating multimodal PV power forecasting models. In our setting, historical sky images constitute the visual modality, while the synchronized physical measurements are used to construct the physical-field representation. The forecasting target is future PV power generation.

Folsom. The Folsom dataset (Pedro et al. (2019)) is a comprehensive solar forecasting benchmark collected in California over the period 2014–2016. It provides quality-controlled irradiance observations at one-minute resolution, including global horizontal irradiance (GHI), direct normal irradiance (DNI), and diffuse horizontal irradiance (DHI). The dataset additionally contains daytime ground-based sky images sampled at one-minute intervals, local meteorological observations, GOES-15 satellite information, and Numerical Weather Prediction (NWP) forecasts from the North American Mesoscale model. All temporal records are provided in UTC, while irradiance and meteorological quantities are stored in SI units. Compared with SKIPP'D, which directly provides PV generation as the prediction target, Folsom primarily supports solar irradiance forecasting and therefore offers a complementary evaluation setting in which the target is an atmospheric radiative quantity rather than electrical power output. Its synchronized imagery and heterogeneous meteorological observations make it particularly useful for evaluating whether physical-field representations improve forecasting beyond purely image-based models.

NREL Sky Imagery. The NREL Sky Imagery dataset (Hammond & Korgel (2026)) is collected at the Solar Radiation Research Laboratory (SRRL) of the National Renewable Energy Laboratory in Golden, Colorado, USA. The dataset uses an EKO ASI-16 all-sky imager operating at its native one-minute acquisition frequency and provides high-resolution 1920×1920 JPEG sky images. Data collection began in October 2025 and the dataset is continuously updated, providing substantially denser temporal sampling than the standard ten-minute SRRL public image gallery. Each daytime image is temporally associated with the nearest one-minute observation from the SRRL Baseline Measurement System, with a matching tolerance of approximately five minutes.

The accompanying physical measurements provide a particularly rich set of atmospheric variables. They include GHI, DNI, DHI, secondary irradiance measurements, ambient air temperature, relative humidity, average and peak wind speed, wind direction, atmospheric pressure, precipitation, solar zenith angle, solar azimuth angle, and a derived clear-sky index. These variables jointly describe radiative, thermodynamic, and dynamical aspects of the local atmosphere and therefore provide a natural basis for constructing the physical-field modality. The dataset is intended primarily for intra-hour solar irradiance forecasting, cloud-motion analysis, and cross-site transfer learning. In our experiments, historical all-sky images provide the visual stream, while synchronized irradiance and meteorological quantities constitute the physical observations used by the physical-field encoder.

MeteoNet. The MeteoNet dataset (Larvor et al. (2020)) is an open meteorological dataset released by Météo-France. It covers the period from 2016 to 2018 over two 550 km × 550 km regions in northwestern and southeastern France. The dataset integrates multiple heterogeneous data sources, including ground-station observations, precipitation radar maps, satellite imagery, numerical weather prediction outputs, and static land–sea and terrain masks. Ground observations are sampled every six minutes, radar products every five minutes at a spatial resolution of 0.01°, and satellite cloud-type maps every 15 minutes at a resolution of 0.03°.

The physical measurements include temperature, dew-point temperature, humidity, atmospheric pressure, precipitation, wind speed, and wind direction. MeteoNet further provides two- and three-dimensional forecasts from the AROME and ARPEGE models, containing near-surface and upper-atmospheric variables such as temperature, humidity, horizontal wind components, vertical velocity, pressure, and geopotential. These temporally aligned observations, remote-sensing products, and forecast fields make MeteoNet suitable for multimodal weather forecasting, precipitation nowcasting, and learning representations that jointly capture radiative, thermodynamic, moisture, and dynamical atmospheric processes.

Boreas. The Boreas dataset (Burnett et al. (2023)) is a multi-season autonomous-driving dataset collected along repeated routes near Toronto, Canada. It contains 44 driving sequences covering more than 350 km between November 2020 and November 2021, encompassing substantial seasonal and weather variations, including sunshine, rain, accumulated snow, and active snowfall. Its sensing platform combines a 128-beam Velodyne Alpha-Prime lidar, a 2448 × 2048 FLIR Blackfly S monocular camera, a 360° Navtech scanning radar, and an Applanix POS LV GNSS–INS system. The lidar and camera operate at 10 Hz, while the radar operates at 4 Hz.

The dataset provides synchronized sensor streams, intrinsic and extrinsic calibration parameters, and centimetre-level post-processed ground-truth poses. Repeated observations of the same environment under different seasons and weather conditions enable systematic evaluation of multimodal representation learning and sensor robustness under distribution shifts. Boreas primarily supports all-weather odometry, long-term metric localization, sensor fusion, and 3D object detection. Its complementary camera, lidar, radar, and navigation measurements are particularly suitable for studying how different sensing modalities respond to adverse environmental conditions.

B.2 Baseline Details

LASSO. LASSO (Ranstam & Cook (2018)) addresses overfitting and feature redundancy in regression by augmenting the least-squares objective with an ℓ_1 penalty. It directly maps the input variables to the target while shrinking weakly informative coefficients toward zero, thereby performing regression and implicit feature selection simultaneously. We include LASSO as a classical structured-data baseline because it provides an interpretable linear reference for assessing whether the gains of more complex models arise from nonlinear temporal and multimodal representation learning.

Random Forest. Random Forest (RF) (Pavlov (2000)) models nonlinear relationships through an ensemble of decision trees. Each tree is trained on a bootstrapped subset of the training data, considers a randomly sampled subset of features at each split, and produces an independent prediction; the final regression output is obtained by averaging the predictions of all trees. RF is selected as a classical structured-data baseline because it captures nonlinear feature interactions without learned temporal representations, providing a strong non-neural reference for evaluating the benefits of sequence modeling and multimodal fusion.

TabM. TabM (Gorishniy et al. (2025)) addresses supervised learning on tabular data through a parameter-efficient ensemble of MLPs. It uses BatchEnsemble-style parameter sharing to generate multiple predictions for each sample within a single model and aggregates these predictions to improve optimization, robustness, and generalization. We select TabM as a representative neural baseline for structured inputs because it offers strong regression performance without explicitly modeling temporal order, enabling us to distinguish the contribution of temporal and multimodal inductive biases from that of a powerful numerical feature learner.

LSTM. Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber (1997)) was introduced to mitigate the vanishing-gradient problem that prevents conventional recurrent networks from learning long-range dependencies. At each time step, its input, forget, and output gates regulate information written to, retained in, and read from a persistent memory cell; the resulting hidden representation is subsequently mapped to the regression target. As regression tasks are inherently sequential, we include LSTM as the canonical recurrent baseline for evaluating whether more recent architectures provide advantages beyond conventional gated temporal modeling.

TSMixer. TSMixer (Chen et al. (2023b)) is an all-MLP architecture designed for multivariate time-series forecasting. Its stacked mixer blocks alternately apply MLPs along the temporal dimension to capture dependencies across time steps and along the feature dimension to model interactions among variables, after which a projection head produces the forecasts. We include TSMixer as a lightweight temporal baseline because it directly targets regression forecasting while avoiding recurrence and self-attention, allowing us to evaluate whether simple temporal and cross-variable mixing is sufficient for the considered tasks.

xLSTM. xLSTM (Beck et al. (2024)) modernizes the original LSTM through two complementary memory mechanisms: sLSTM introduces exponential gating and enhanced memory mixing, whereas mLSTM replaces scalar memory with a matrix-valued memory and covariance-style updates. These blocks are stacked to encode the input sequence before a task-specific head generates the prediction. We select xLSTM as a recent recurrent baseline because it tests whether improved long-range memory and scalable sequence modeling can better capture the temporal dynamics underlying regression-based prediction.

BiMamba. Sky-BiMamba (Zhang et al. (2025)) addresses ultra-short-term PV power forecasting by jointly modeling historical sky images and PV measurements. It first extracts cloud appearance and motion cues from sky-image sequences using frame differencing and lightweight visual encoders, then employs bidirectional Mamba modules and temporally aware cross-modal attention to integrate visual features with PV sequences; a dynamic gating mechanism further reduces the influence of low-quality images before prediction. We include it as a recent domain-specific multimodal baseline because its task and input modalities closely match our setting, making it a particularly direct comparison for evaluating PhyMo.

AstroCLIP. AstroCLIP (Parker et al. (2024)) addresses the joint representation of galaxy images and optical spectra, which provide complementary observations of the same physical system. It first pretrains transformer-based image and spectrum encoders using modality-specific self-supervised objectives and subsequently aligns paired representations in a shared latent space through contrastive learning. AstroCLIP is selected as a recent scientific multimodal baseline because its separate encoding followed by cross-modal alignment closely matches the general paradigm considered in our work, while providing a reference for comparison with PhyMo's explicitly physics-aware representation.

Maven. Maven (Zhang et al. (2024)) addresses the substantial imbalance between abundant photometric observations and scarce but information-rich spectra in supernova science. It encodes photometric and spectroscopic observations separately, contrastively aligns their representations using approximately 0.5 million simulated supernovae, and then adapts the pretrained model to real observations for downstream classification and redshift regression. We include Maven as a recent multimodal scientific model because it demonstrates how heterogeneous observations can be unified through contrastive pretraining and transferred to regression tasks, providing a relevant comparison with our multimodal alignment strategy.

AION-1. AION-1 (Parker et al. (2026)) addresses the fragmentation of astronomical learning systems across heterogeneous instruments, surveys, and data types. It first transforms images, spectra, and scalar measurements using modality-specific tokenizers and then applies transformer-based masked modeling to their combined token sequences, producing a unified encoder transferable to property estimation, retrieval, segmentation, and spectral reconstruction. We select AION-1 as a recent omnimodal scientific foundation model because it represents a general approach to integrating heterogeneous scientific observations and therefore provides a strong comparison for assessing the effectiveness of PhyMo's dedicated physical-field organization and learning objective.

B.3 Metrics

We evaluate model performance using root mean squared error (RMSE), coefficient of determination (R^2), and mean absolute error (MAE). Let $\{y_i\}_{i=1}^N$ denote the ground-truth targets, $\{\hat{y}_i\}_{i=1}^N$ denote the corresponding predictions, and

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{45}$$

denote the mean of the ground-truth targets.

Root Mean Squared Error (RMSE). RMSE measures the square root of the average squared prediction error:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}. \tag{46}$$

Because prediction errors are squared before aggregation, RMSE places relatively larger penalties on samples with large deviations. It is therefore particularly sensitive to large forecasting errors. A lower RMSE indicates better predictive accuracy, with an ideal value of zero.

Mean Absolute Error (MAE). MAE computes the average absolute difference between predictions and ground-truth values:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|. \tag{47}$$

In contrast to RMSE, MAE penalizes prediction errors linearly and is therefore less sensitive to a small number of large deviations. It provides a direct measure of the typical absolute forecasting error in the same unit as the prediction target. Lower MAE values indicate better performance.

Coefficient of Determination (R^2). The coefficient of determination evaluates how much of the variation in the target variable is explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}. \tag{48}$$

The numerator corresponds to the residual sum of squares, while the denominator represents the total variation of the ground-truth targets around their mean. An R^2 value of 1 indicates perfect predictions, whereas $R^2 = 0$ indicates performance equivalent to predicting the target mean for every sample. Negative R^2 values are also possible and indicate that the model performs worse than this mean-value baseline. Unlike RMSE and MAE, larger R^2 values correspond to better predictive performance.

Complementarity of the Metrics. The three metrics characterize prediction quality from complementary perspectives. RMSE emphasizes large deviations, MAE reflects the average magnitude of forecasting errors with greater robustness to extreme errors, and R^2 measures the explanatory capability of the model relative to the intrinsic variability of the target. We therefore report all three metrics to provide a more comprehensive evaluation of downstream forecasting performance.

C Limitations

Despite its effectiveness, PhyMo has several limitations that motivate future work. First, owing to space constraints, our downstream evaluations primarily focus on regression-based forecasting; we will extend PhyMo to broader tasks, including classification, reconstruction, and physical-state estimation, to further assess its applicability and generalizability in real-world physical systems. Second, although our experiments cover diverse datasets and physical environments, the available physical variables and their spatiotemporal resolutions remain dataset-dependent. We will therefore evaluate PhyMo under incomplete, noisy, and

asynchronously sampled observations and extend it to additional physical domains. Finally, the three-stage training procedure and PDE residual computation introduce additional computational overhead compared with conventional data-driven models. Future work will improve efficiency through lightweight physical-field encoders, shared-stage optimization, and more efficient PDE-operator computation.

AI Use Statement

In this work, we used generative AI tool to assist with translation, implement methods with codes and provide feedback on experiments. We have not used generative AI tools for generate synthetic data sets, help develop conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose hypotheses, clean the dataset, support qualitative and thematic data analysis and interpret results. We have reviewed all AI-assisted work. For example, LLM-generated code was verified and tested for correctness by 2 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.