MicroEvoEval: A Systematic Evaluation Framework for Image-Based Microstructure Evolution Prediction

Qinyi Zhang[1*], Duanyu Feng[2*], Ronghui Han[1], Yangshuai Wang[3†], Hao Wang[1†]

[1] School of Mathematics, Sichuan University, No. 24 Yihuan Road, Chengdu, China.
[2] College of Computer Science, Sichuan University, No. 24 Yihuan Road, Chengdu, China.
[3]Department of Mathematics, National University of Singapore, 10 Lower Kent Ridge Road, Singapore.
2024322010008@stu.scu.edu.cn, fengduanyuscu@stu.scu.edu.cn, hronghui@stu.scu.edu.cn, yswang@nus.edu.sg, wangh@scu.edu.cn

Abstract

Simulating microstructure evolution (MicroEvo) is vital for materials design but demands high numerical accuracy, efficiency, and physical fidelity. Although recent studies on deep learning (DL) offers a promising alternative to traditional solvers, the field lacks standardized benchmarks. Existing studies are flawed due to a lack of comparing specialized MicroEvo DL models with state-of-the-art spatio-temporal architectures, an overemphasis on numerical accuracy over physical fidelity, and a failure to analyze error propagation over time. To address these gaps, we introduce `MicroEvoEval`, the first comprehensive benchmark for image-based microstructure evolution prediction. We evaluate 14 models, encompassing both domain-specific and general-purpose architectures, across four representative MicroEvo tasks with datasets specifically structured for both short- and long-term assessment. Our multi-faceted evaluation framework goes beyond numerical accuracy and computational cost, incorporating a curated set of structure-preserving metrics to assess physical fidelity. Our extensive evaluations yield several key insights. Notably, we find that modern architectures (e.g., VMamba), not only achieve superior long-term stability and physical fidelity but also operate with an order-of-magnitude greater computational efficiency. The results highlight the necessity of holistic evaluation and identify these modern architectures as a highly promising direction for developing efficient and reliable surrogate models in data-driven materials science.

Code — https://github.com/ArcueidCroft/MircoEvoEval
Datasets — https://huggingface.co/datasets/ArcueidCroft/MicroEvoEval
Extended version — https://arxiv.org/abs/2511.08955

Introduction

Material properties are strongly governed by microstructures—the mesoscale arrangement of grains and phases (Olson 1997; Bhadeshia 2001). A key objective in materials science is to control microstructural evolution via processing techniques such as casting and annealing (Porter and Easterling 1992; Reed 2008). Accurate prediction of this evolution is critical for designing advanced materials, as it involves complex spatio-temporal dynamics highly sensitive to processing conditions (Seetharaman et al. 2021; Chen 2002; Mo et al. 2021; Jou, Leo, and Lowengrub 1997). It requires a balance of three key attributes: **numerical accuracy**, to faithfully capture the evolution of pixel-based states (Sang et al. 2023; Boettinger et al. 2000); **computational efficiency**, to enable rapid exploration of design spaces (Tandogan, Budnitzki, and Sandfeld 2025; Noguchi, Aihara, and Inoue 2024); and **physical fidelity**, to ensure that predictions remain consistent with underlying physical laws and preserve essential structural features that are often missed by conventional pixel-based metrics (Kamachali et al. 2018; Hasan et al. 2023).

Traditionally, MicroEvo has been modeled using theoretical approaches such as phase-field simulations (Chen 2002; Tourret, Liu, and LLorca 2022; Moelans, Blanpain, and Wollants 2008; Steinbach 2013), where the governing partial differential equations (PDEs) are solved using numerical methods like finite difference or spectral methods (Liu and Shen 2003; Badalassi, Ceniceros, and Banerjee 2003; Shen and Yang 2009). While accurate, these methods are often computationally expensive and may not be practical for rapid evaluations. Moreover, for complex or poorly understood materials, the PDEs corresponding to the evolution laws may be extremely challenging to formulate. These limitations have motivated the development of data-driven approaches for modeling microstructure evolution.

More recently, deep learning (DL) has opened new avenues for modeling microstructure evolution by learning complex spatio-temporal dynamics directly from image-based data. Initial efforts have shown promise by framing MicroEvo as a image-based sequence prediction task, with evaluations often focused on numerical accuracy (standard image metrics like MSE) (Lanzoni et al. 2022; Farizhandi and Mamivand 2023). These studies have frequently adapted established or specialized architectures like E3D-LSTM (Yang et al. 2021) and the recent VMamba (Jingjie et al. 2025). However, these models have often been developed in isolation, lacking a systematic comparisons. Concurrently, the broader field of general-purpose spatio-

*These authors contributed equally.
†Co-corresponding author.

temporal prediction has seen rapid advances, producing powerful models like SimVP.v2 (Tan et al. 2025) and PredFormer (Tang et al. 2025). Despite their success on other diverse tasks, their applicability to the physically-constrained domain of MicroEvo remains largely unevaluated.

Therefore, a comprehensive benchmark is currently missing to systematically evaluate and compare the diverse deep learning approaches for image-based MicroEvo prediction. This gap is multi-dimensional. Firstly, most evaluations focus on **numerical accuracy**, while overlooking **physical fidelity**, the critical aspect for understanding microstructures. Secondly, existing studies rarely offer standardized comparisons between specialized MicroEvo models and potential promising general-purpose spatio-temporal architectures. Despite these, we also find that existing evaluations typically evaluated the long-term accuracy with models trained on short-term data (Yang et al. 2021). As a result, this methodological gap obscures the process of error accumulation, making it unknown whether a model's short-term accuracy reliably translates to long-term stability.

To address these critical gaps, we present `MicroEvoEval`, the first comprehensive benchmark specifically designed for image-based microstructural evolution. Our approach is built on three key design principles. First, we define a structured suite of tasks and datasets. `MicroEvoEval` includes representative MicroEvo problems such as grain growth and spinodal decomposition, covering a range of physical complexities. The datasets, derived from high-fidelity numerical simulations (Yang et al. 2021), are further organized into separate test sets that support a joint evaluation of short-term accuracy and long-term stability. This allows for a direct analysis of how errors propagate over time in these systematized tasks. Second, we provide a standardized evaluation across diverse model types. We benchmark 5 models tailored for MicroEvo and 9 general-purpose spatio-temporal architectures, offering the first systematic comparison of their performance under consistent settings. Third, we create a holistic performance assessment. Beyond conventional metrics such as MSE and SSIM for **numerical accuracy**, our framework incorporates measures of **physical fidelity** (The Log of Error of Total Area Proportion "L-ETAP" and The Log of Error of Average Proportion of a Single Region "L-EAPSR") and **computational efficiency** (inference time), yielding a more complete and physically meaningful evaluation.

The main contributions of this work are: (1) `MicroEvoEval`**: the first standardized benchmark for image-based microstructural evolution.** It comprises representative tasks designed for short- and long-term prediction, along with a comprehensive metric suite that jointly evaluates numerical accuracy, computational efficiency, and physical fidelity. (2) **A systematic evaluation across diverse architectural paradigms.** We present the first standardized comparison between MicroEvo-specific models and state-of-the-art general-purpose spatio-temporal architectures, revealing their respective advantages and limitations. (3) **Actionable insights for future research.** We demonstrate the necessity of assessing long-term stability and physical fidelity, as short-term numerical accuracy of trained model is a poor indicator. We also show that modern architectures can give a better accuracy-efficiency trade-off, providing a potential path toward developing more reliable and practical models for materials science.

Related Work

Physics-Based Numerical Methods. Modeling microstructure evolution has traditionally relied on continuum-scale, physics-based formulations. Phase-field models, which describe spatio-temporal dynamics through PDEs for phenomena like solidification and grain growth (Chen 2002; Steinbach et al. 1996), are particularly widespread due to their flexibility. However, their high computational cost, especially with explicit time integration schemes, restricts the accessible time and length scales of simulations and hinders rapid exploration of process-parameter space (Greenwood et al. 2018). Furthermore, for complex or poorly characterized materials, deriving tractable and accurate PDEs can be a significant challenge in itself. While acceleration techniques (Guo and Xiong 2015) and alternative data-driven frameworks like Markov Random Fields (Acar and Sundararaghavan 2016) have been explored to alleviate these issues, the trade-off between fidelity and computational cost remains a primary bottleneck, motivating the search for efficient surrogate models.

Deep Learning for Microstructure Evolution. To address the computational bottleneck of numerical methods, deep learning (DL) has emerged as a promising alternative for learning MicroEvo dynamics directly from data. Many pioneering studies have adapted established deep learning architectures for this purpose. These approaches often rely on Recurrent Neural Network (RNN) variants to capture temporal dependencies. Examples include adapting classic architectures like ConvLSTM (Mao et al. 2024), ConvGRU (Lanzoni et al. 2022), and PredRNN (Farizhandi and Mamivand 2023) to forecast microstructural patterns. Other works employ a CNN-RNN structure, using CNNs as powerful feature extractors for the recurrent core. Prominent examples in this category include E3D-LSTM (Yang et al. 2021) and the more recent state-space-based model, VMamba (Jing-jie et al. 2025). Although these models have demonstrated the feasibility of data-driven MicroEvo prediction, they have been developed and evaluated on disparate tasks and mainly focus on numerical accuracy on long-term prediction, making it difficult to comprehensively compare their relative strengths and weaknesses.

General-Purpose Spatio-Temporal Prediction. Concurrently, the broader field of computer vision has produced a wealth of powerful models for general-purpose spatio-temporal prediction (i.e., video prediction). These architectures have evolved significantly, from advanced recurrent networks like PredRNN++ (Wang et al. 2018) and MAU (Chang et al. 2021) to highly efficient, non-recurrent CNN models such as SimVP (Gao et al. 2022) and its successor SimVP.v2 (Tan et al. 2025). More recently, Transformer-based architectures have become prominent for their ability to capture long-range dependencies, leading to models like PredFormer (Tang et al. 2025) and hybrids like SwinLSTM (Tang et al. 2023). The cutting edge continues

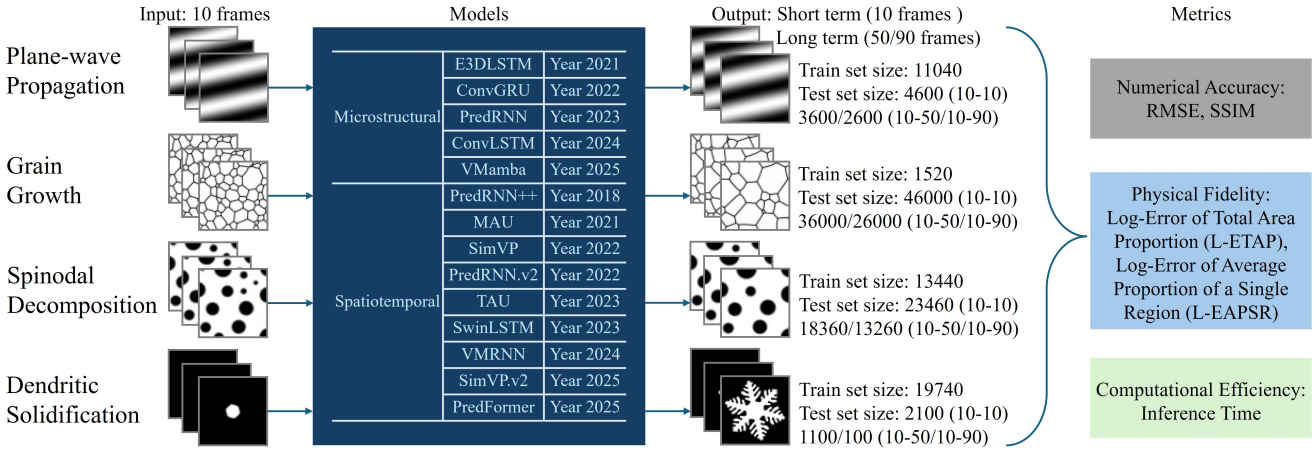

Figure 1: Schematic of the `MicroEvoEval` benchmark for microstructure evolution prediction.

to advance with refined attention mechanisms in models like TAU (Tan et al. 2023) and the integration of Mamba in VMRNN (Tang et al. 2024). Despite their proven success on different general tasks, these state-of-the-art models are not designed with inherent physical constraints. Their effectiveness in predicting physically plausible microstructure evolution, a domain with fundamentally different underlying rules from natural videos, has not been systematically investigated. Our benchmark aims to bridge this gap by evaluating both domain-specific and general-purpose models under a unified framework.

MicroEvoEval

In this section, we present the taxonomy of tasks, the dataset construction, and the metrics used for our benchmark `MicroEvoEval`.

The Taxonomy of Microstructure Evolution Tasks

At the microscale, we focus on the evolutions of microstructures governed by known PDEs, as this facilitates robust and quantitative evaluation. These governing PDEs are broadly categorized based on their physical modeling mechanisms. The first category includes equations describing periodic materials, such as the elastic wave equation for modeling wave propagation. The second involves phase-field models, including the Allen–Cahn (A–C) equation (Allen and Cahn 1979) for non-conserved order parameters (e.g., crystalline structure, ferromagnetic domains) and the Cahn–Hilliard (C–H) equation (Cahn and Hilliard 1958) for conserved quantities (e.g., composition, mass density). A third category covers multi-physics PDEs that account for coupled interactions among different physical fields. These categories cover all major PDEs governing microstructure evolution, ensuring broad physical representativeness in our study.

In this work, we select representative applications for each of these PDE categories. To ensure systematic coverage, the selected four tasks follow a physical taxonomy encompassing periodic, non-conserved, conserved, and coupled multi-physics mechanisms. Each task was designed to be non-overlapping in its governing dynamics and data domain. [1] Specifically, we consider **plane-wave propagation** for periodic structures, **grain growth** governed by the A-C equation, **spinodal decomposition** described by the C–H equation, and **dendritic solidification** as a coupled multi-physics process. It is worth noting that generating high-quality data for these PDEs is nontrivial. To ensure reproducibility and enable meaningful benchmark comparisons, we adopt the datasets and PDE formulations introduced in (Yang et al. 2021), which are widely used in the modeling of microstructural evolution (Fan et al. 2024; Jing-jie et al. 2025). We then applied specific processing to these datasets to align them with our evaluation goals in the next subsection. In the following, we first briefly describe each task and present its associated governing equation.[2]

The first task is a simple yet physically relevant toy model that captures the periodic nature of microstructures. It describes the **plane-wave propagation** of a scalar field $c(x, y, t)$, given by:

$$c(x, y, t) = \frac{1}{2}\sin(k_x x + k_y y + \omega t + \theta_0)e^{-\beta t} + \frac{1}{2}, \quad (1)$$

where $\vec{k} = (k_x, k_y)$ is the wave vector, θ_0 is a random phase, and β is a temporal decay exponent.

The second task focuses on **grain growth**, which is simulated using a multi-order-parameter phase-field model (Moelans, Blanpain, and Wollants 2008), where a set of order parameters $\{\eta_1(x), \eta_2(x), \ldots, \eta_N(x)\}$ represents N distinct grain orientations. The total free energy reads

$$F = \int \left[f(\eta_1, \eta_2, \ldots, \eta_N) + \frac{\nu}{2} \sum_{i=1}^{N} (\nabla \eta_i)^2 \right] \mathrm{d}V. \quad (2)$$

The temporal evolution of each $\eta_i(x)$ follows the time-dependent A-C equation, a representative *second-order*

[1] We adopt PDE-based datasets for their scalability and clean ground truths, as real-world microstructure evolution data remain scarce and noisy for standardized benchmarking.

[2] More details and more datasets see the extended version.

PDE:

$$\frac{\partial \eta_i}{\partial t} = -L\frac{\delta F}{\delta \eta_i}. \tag{3}$$

The third task involves **spinodal decomposition**, a spontaneous phase separation process in binary mixtures. Unlike A–C equation, this process is governed by the *fourth-order* C–H equation, which enables domain formation and coarsening without nucleation:

$$\frac{\partial c}{\partial t} = \nabla \cdot \left[Mc(1-c)\nabla \left(\frac{\partial f_{\text{chem}}}{\partial c} - \epsilon \nabla^2 c \right) \right], \tag{4}$$

where c is the molar fraction of one component in a binary alloy.

The fourth task models **dendritic solidification** (Kobayashi 1993). The system is described by a temperature field T and an order parameter φ, which distinguishes solid ($\varphi = 1$) and liquid ($\varphi = 0$) phases. The free energy functional is given by:

$$F[\varphi, T] = \int \left[\frac{1}{2}\epsilon(\theta)^2 |\nabla \varphi|^2 + f(\varphi, T) \right] \mathrm{d}r, \tag{5}$$

where anisotropy is introduced via the orientation-dependent gradient coefficient $\epsilon(\theta) = \epsilon_0 \left(1 + \delta \cos[n(\theta - \theta_0)]\right)$, with $\theta = \arctan(-\varphi_y/\varphi_x)$. The bulk free energy is modeled by a double-well potential The coupled time evolution equations are:

$$\tau \frac{\partial \varphi}{\partial t} = -\frac{\delta F}{\delta \varphi}, \tag{6a}$$

$$\frac{\partial T}{\partial t} = \nabla^2 T + K\frac{\partial \varphi}{\partial t}, \tag{6b}$$

where K is the latent heat parameter.

Tasks and Datasets

`MicroEvoEval` is built upon these four representative tasks spanning a range of physical phenomena, this section details the design of the tasks and datasets. All underlying data is derived from high-fidelity original simulations on a 256×256 grid, subsequently downsampled to a uniform 64×64 resolution from original datasets (Yang et al. 2021).

To investigate the relationship between short-term accuracy and long-term stability, we design a dual evaluation setting based on our curated datasets. The short-term setting assesses a model's ability to capture immediate dynamics (e.g., predicting the next `10` frames from `10` inputs), while the long-term setting evaluates autoregressive performance over extended horizons (up to `50` and `90` future frames).[3] This twofold structure enables a systematic analysis of error accumulation and the extent to which short-term accuracy correlates with long-term reliability. Figure 1 summarizes the benchmark tasks, datasets, and key statistics in `MicroEvoEval`.

[3]A key area of focus in this domain is assessing the performance of models when they are trained for short-term prediction but then deployed autoregressively for long-term forecasting. More implementation details see the extended version.

Plane-wave propagation. This task utilizes data from a toy model governed by Equation 1, with different $\vec{k}, \omega, \beta, \theta_0$. The original training set providing `240` sequences of `200` frames each. For our benchmark, we process their provided training split into `11040` short clips of `20` frames to form our training set for forecasting the next `10` frames given `10` input frames. We then partition their original test set to create our evaluation sets. The short-term prediction task has `4600` test cases. For the long-term prediction task predicting the next 50 frames, we make `3600` test cases, and `2600` test cases for the task predicting next 90 frames.

Grain growth. The dataset for this task is based on simulations of polycrystalline grain growth governed by the A-C equation (3). The initial polycrystalline structures were generated via Voronoi tessellation with 100 random seeds. The source data consists of `200`-frame sequences depicting grain coarsening. We adopt the designated training clips from the original study to form our training set of `1520` samples. For our evaluation sets, the short-term task involves `46000` test cases, while the long-term task involves `36000` test cases for the task predicting next 50 frames and `26000` test cases for the task predicting next 90 frames.

Spinodal decomposition. This task is based on simulations solving the Cahn-Hilliard Equation (4) with varying initial configurations sampled from $c_0 = 0.25, 0.5$, or 0.75, plus random noise $\Delta c = 0.01$ to initiate phase separation c . The original work generated `640` simulations, each `100` frames long. Following their methodology, we use staggered clips from their training simulations to create our `13440` training samples. The test simulations are then used by us to construct distinct short-term (`23460` test cases) and long-term (`18360` test cases) evaluation sets for task predicting next 50 frames while the long-term task (predicting next 90 frames) involves `13260` test cases.

Dendritic Solidification. For this task, we employ datasets from simulations of dendritic solidification based on Equation (6). The original study provides `940` training simulations, each producing `100` frames. Our training set is built from overlapping clips extracted from these simulations, resulting in `19740` samples. The provided test simulations, which notably include out-of-distribution samples to assess generalization, are partitioned by us into evaluation sets for short-term (`2100` test cases) and long-term (`1100` test cases for predicting next 50 frames and 100 test cases for predicting next 90 frames) performance assessment.

Metrics

To enable a holistic assessment, our evaluation framework integrates metrics across three dimensions: **predictive accuracy**, **physical fidelity**, and **computational efficiency**, each critical for practical applications in materials science.[4]

Numerical Accuracy. It is assessed using standard image similarity metrics that quantify pixel-wise differences between the predicted and ground truth frames. We employ two widely-used metrics used in previous studies (Yang et al. 2021; Farizhandi and Mamivand 2023): the Root Mean Squared Error (RMSE) and the Structural Similarity Index

[4]More implementation details see the extended version.

		Plane-wave propagation			Grain growth			Spinodal decomposition			Dendritic solidification		
Model domain	Model	RMSE	SSIM	L-ETAP	RMSE	SSIM	L-EASPR	RMSE	SSIM	L-ETAP	RMSE	SSIM	L-ETAP
Microstructural	E3DLSTM	0.01673	0.99315	-0.993	0.03394	0.98607	-2.488	0.00418	0.99970	-2.826	0.01275	0.99717	-2.540
	ConvGRU	0.00338	0.99944	-0.994	0.02134	0.99183	-2.586	0.00389	0.99967	-2.853	0.00267	0.99965	-2.865
	PredRNN	0.00102	0.99994	-0.995	0.02453	0.99165	-2.541	**0.00270**	0.99985	**-2.916**	0.00574	0.99911	-2.737
	ConvLSTM	0.00328	0.99965	**-0.996**	0.03082	0.98707	-2.492	0.00423	0.99969	-2.857	0.00545	0.99912	-2.726
	VMamba	0.00151	0.99991	-0.993	**0.00871**	**0.99926**	**-2.875**	0.00382	0.99983	-2.902	**0.00259**	**0.99990**	**-2.942**
Spatiotemporal	PredRNN++	0.00108	0.99993	-0.993	0.03071	0.98795	-2.545	0.00285	0.99981	-2.844	0.00662	0.99883	-2.702
	MAU	0.00867	0.99732	-0.977	0.05939	0.95131	-1.830	0.00773	0.99876	-2.494	0.00752	0.99828	-2.553
	SimVP	0.00471	0.99905	-0.993	0.02563	0.99139	-2.600	0.00492	0.99961	-2.678	0.00646	0.99887	-2.665
	PredRNN.v2	0.00157	0.99986	-0.993	0.03954	0.97898	-2.421	0.00425	0.99964	-2.844	0.00561	0.99915	-2.739
	TAU	0.00271	0.99942	-0.992	0.02127	0.99463	-2.633	0.00421	0.99973	-2.864	0.00441	0.99954	-2.741
	SwinLSTM	0.00236	0.99938	-0.995	0.03269	0.98502	-2.570	0.00354	0.99932	-2.868	0.00764	0.99877	-2.612
	VRMNN	**0.00081**	**0.99998**	-0.995	0.02230	0.99602	-2.618	0.00347	**0.99987**	-2.843	0.00668	0.99962	-2.673
	SimVP.v2	0.00578	0.99920	-0.993	0.02366	0.99339	-2.603	0.00442	0.99954	-2.755	0.00495	0.99945	-2.638
	PredFormer	0.00917	0.98603	-0.972	0.02453	0.99165	-2.696	0.01273	0.96767	-2.243	0.01261	0.95548	-2.443

Table 1: The performance of different models in short-term prediction (10-10) of `MicroEvoEval`.

Measure (SSIM). For each frame, RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N_x N_y}\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}\left(p_g(i,j) - p_p(i,j)\right)^2}, \quad (7)$$

where $p_g(i,j)$ and $p_p(i,j)$ are the pixel values of the ground truth g and prediction p, respectively. N_x, N_y are the width and height of the image in pixels. For each frame, SSIM provides a measure of perceived structural similarity and is defined as:

$$\text{SSIM} = \frac{\left(2\bar{p}_g\bar{p}_p + c_1\right)\left(2\sigma_{gp} + c_2\right)}{\left(\bar{p}_g^2 + \bar{p}_p^2 + c_1\right)\left(\sigma_g^2 + \sigma_p^2 + c_2\right)}, \quad (8)$$

where $\bar{p}_k$ and σ_k ($k = g, p$) are the average pixel value and variance, and σ_{gp} is their covariance. The constants c_1 and c_2 stabilize the division. The final metric scores are computed by averaging the frame-wise results over all predicted frames and test samples.

Physical Fidelity. Recognizing that pixel-level metrics may not adequately capture the preservation of essential physical properties, we introduce two custom metrics to assess physical fidelity. We define that $\Omega_{i,t}$ is the frame at time t for sample $i = 1, \cdots, N$.

For tasks where the overall phase fraction is a key evolving property, such as plane-wave propagation, spinodal decomposition, and dendritic solidification, we introduce the **Log-Error of Total Area Proportion (L-ETAP)**. This metric is physically significant as it evaluates a model's ability to conserve the total mass or volume of the evolving phase, a fundamental physical constraint. Specifically, we track this by computing the phase area fraction $A(\Omega_{i,t})/S$, where $A(\Omega_{i,t})$ is the total area of the evolving phase (e.g., the white pixels for dendrites) and S is the total image area. Any deviation in this fraction indicates unphysical mass gain or loss over time. L-ETAP therefore captures the degree to which a model maintains global conservation laws across a temporal sequence. It is defined as:

$$\text{L-ETAP} = \frac{1}{2}\log_{10}\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T}\left(\frac{A(\Omega^g_{i,t}) - A(\Omega^p_{i,t})}{S}\right)^2\right]. \quad (9)$$

For the grain growth task, where the kinetics of grain coarsening are critical, we introduce the **Log-Error of Average Proportion of a Single Region (L-EAPSR)**. This metric's physical meaning lies in its ability to assess whether a model accurately captures the change in average grain size by measuring the error in the number of grains over time. $C(\Omega_{i,t})$ counts the number of grains (connected regions) at each frame, the reciprocal $1/C$ approximates the average area of a single grain. L-EAPSR thus reflects how well the model captures this key dynamical feature of the underlying phase-field evolution. It is defined as:

$$\text{L-EAPSR} = \frac{1}{2}\log_{10}\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T}\left(\frac{1}{C^g_{i,t}} - \frac{1}{C^p_{i,t}}\right)^2\right]. \quad (10)$$

Let $C^g_{i,t} = C(\Omega^g_{i,t})$ and $C^p_{i,t} = C(\Omega^p_{i,t})$. Consequently, the final values for both metrics will be less than zero. A result that is more negative (more smaller), signifies that the model better maintains the physical properties.

Computational Efficiency. Finally, to evaluate the practical applicability of each model, we measure its computational efficiency. This is quantified by the average inference time required to predict a complete short-term sequence (forecasting `10` frames from `10` input frames).[5]

Experiment

Experimental Setup

To conduct this evaluation, we assess a diverse set of 14 models, comprising 5 architectures specifically tailored for MicroEvo and 9 state-of-the-art general-purpose spatio-temporal models. The domain-specific MicroEvo models include E3D-LSTM (Yang et al. 2021), ConvGRU (Lanzoni et al. 2022), PredRNN (Farizhandi and Mamivand 2023), ConvLSTM (Mao et al. 2024), and VMamba (Jing-jie et al. 2025). The general-purpose spatio-temporal architectures consist of PredRNN++ (Wang et al. 2018), MAU (Chang et al. 2021), SimVP (Gao et al. 2022), PredRNN.v2 (Wang

[5]The computational time scales linearly with the number of predicted frames, and we report this short-term inference time.

		Plane-wave propagation			Grain growth			Spinodal decomposition			Dendritic solidification		
Model domain	Model	RMSE	SSIM	L-ETAP	RMSE	SSIM	L-EASPR	RMSE	SSIM	L-ETAP	RMSE	SSIM	L-ETAP
Long term prediction (10-50)													
Microstructural	E3DLSTM	0.11997	0.78022	-0.667	0.11065	0.87022	-1.750	0.01951	0.99546	-1.809	0.05015	0.97352	-1.188
	ConvGRU	0.01845	0.98973	-0.714	0.08843	0.91765	-1.760	0.01691	0.99727	-2.139	0.02720	0.99092	-1.708
	PredRNN	0.00819	0.99742	**-0.730**	0.22172	0.46442	-0.685	0.01758	0.99655	-1.817	0.25915	0.71776	-0.032
	ConvLSTM	0.11670	0.73850	-0.622	0.10027	0.89229	-1.564	0.02520	0.99230	-1.783	0.04143	0.98008	-1.278
	VMamba	0.00817	0.99709	-0.728	**0.02986**	**0.99008**	**-2.283**	**0.01342**	**0.99848**	**-2.221**	**0.01922**	**0.99609**	**-1.983**
Spatiotemporal	PredRNN++	**0.00754**	**0.99803**	-0.704	0.09926	0.89755	-1.764	0.02294	0.99393	-1.916	0.04285	0.97866	-1.320
	MAU	0.15833	0.65653	-0.411	0.16699	0.68662	-1.093	0.05056	0.96809	-1.144	0.09578	0.92582	-0.813
	SimVP	0.02054	0.98538	-0.725	0.09435	0.89925	-1.899	0.02279	0.99565	-1.820	0.03226	0.98725	-1.777
	PredRNN.v2	0.01160	0.99545	-0.708	0.16927	0.48566	-0.487	0.02493	0.99300	-1.510	0.04015	0.98058	-1.402
	TAU	0.01396	0.99389	-0.727	0.08250	0.92942	-1.868	0.01643	0.99809	-2.213	0.02374	0.99233	-1.797
	SwinLSTM	0.02046	0.98421	-0.726	0.12624	0.81917	-1.805	0.01745	0.99635	-2.000	0.04339	0.98029	-1.513
	VRMNN	0.01113	0.99683	-0.728	0.06248	0.96615	-1.975	0.01888	0.99753	-1.911	0.03521	0.98984	-1.651
	SimVP.v2	0.02234	0.98652	-0.724	0.08931	0.91645	-1.851	0.01950	0.99679	-2.063	0.02615	0.99109	-1.762
	PredFormer	0.02892	0.95577	-0.695	0.08450	0.91212	-2.022	0.05575	0.93378	-1.332	0.05868	0.92531	-1.380
Long term prediction (10-90)													
Microstructural	E3DLSTM	0.17089	0.58738	-0.489	0.14800	0.76496	-1.450	0.03642	0.98552	-1.441	0.16485	0.84943	-0.212
	ConvGRU	0.03635	0.95884	-0.596	0.12688	0.82782	-1.425	0.02791	0.99304	-1.843	0.08216	0.95175	-1.070
	PredRNN	0.02630	0.97109	-0.598	0.31882	0.26338	0.060	0.03744	0.98465	-1.301	0.42713	0.43134	0.296
	ConvLSTM	0.16098	0.54013	-0.442	0.13749	0.79176	-1.173	0.05765	0.96745	-1.268	0.12132	0.90480	-0.489
	VMamba	**0.01652**	**0.98745**	-0.629	**0.04727**	**0.97322**	**-2.011**	**0.02138**	**0.99627**	-1.944	**0.05542**	**0.97682**	**-1.437**
Spatiotemporal	PredRNN++	0.02281	0.98039	**-0.633**	0.13494	0.80531	-1.458	0.05207	0.97385	-1.437	0.12542	0.90127	-0.485
	MAU	0.18475	0.51672	-0.340	0.20358	0.53944	-0.871	0.09445	0.92043	-0.760	0.18592	0.82421	-0.349
	SimVP	0.03558	0.95929	-0.621	0.13397	0.80088	-1.605	0.03582	0.98976	-1.556	0.07325	0.95754	-1.179
	PredRNN.v2	0.03477	0.96081	**-0.633**	0.20170	0.28880	-0.492	0.05702	0.96639	-0.888	0.11749	0.91178	-0.587
	TAU	0.02623	0.97736	-0.626	0.12098	0.84532	-1.557	0.02517	0.99590	**-1.958**	0.05931	0.96924	-1.209
	SwinLSTM	0.05263	0.91687	-0.594	0.17355	0.66420	-1.408	0.03240	0.98954	-1.620	0.09784	0.93640	-0.956
	VRMNN	0.03528	0.96877	-0.615	0.09091	0.91929	-1.697	0.03202	0.99288	-1.524	0.09203	0.94758	-1.140
	SimVP.v2	0.03878	0.95755	-0.617	0.12863	0.82407	-1.536	0.03004	0.99268	-1.811	0.06307	0.96618	-1.149
	PredFormer	0.05832	0.87576	-0.560	0.12084	0.82722	-1.741	0.09589	0.89108	-0.990	0.13028	0.85560	-0.823

Table 2: The performance of different models in long-term prediction (10-50/10-90) of `MicroEvoEval`.

et al. 2022), TAU (Tan et al. 2023), SwinLSTM (Tang et al. 2023), VMRNN (Tang et al. 2024), SimVP.v2 (Tan et al. 2025), and PredFormer (Tang et al. 2025). All models were trained for a maximum of 200, 300, 300, and 400 epochs of each task with the best performing checkpoint selected based on performance in a held-out validation set.[6]

Main Results

Our experiments provide a deep understanding of MicroEvo prediction by investigating four core questions: (1) whether a model's **short-term accuracy** is a reliable indicator of its **long-term stability**; (2) whether standard image metrics (numerical accuracy) are sufficient for evaluation compared to our curated metrics for **physical fidelity**; (3) how domain-specific MicroEvo models perform relative to general-purpose spatiotemporal architectures; and (4) which architectural classes offer the best trade-off between performance and **computational efficiency**.[7]

Short-term performance is a poor indicator of long-term stability. Comparing Table 1 and Table 2, nearly all models exhibit a significant degradation in performance during long-term autoregressive forecasting. For instance, while VMRNN achieves the best short-term performance on the Plane-wave Propagation task, its advantage diminishes significantly in long-term prediction. Another example is PredRNN, which, despite its strong short-term result on Spinodal Decomposition, suffers a catastrophic performance collapse on the long-term Grain Growth task (SSIM of 0.263), demonstrating severe error accumulation. In contrast, VMamba consistently maintains top-tier performance across both short- and long-term horizons, especially in the most challenging 90-frame predictions, highlighting its superior stability. Therefore, for MicroEvo, the results from short-term training are unreliable, the long-term evaluation is essential for identifying robust models.

Physical fidelity metrics provide indispensable insights beyond standard image metrics. This is evidenced by frequent divergences where the model that performs best for numerical accuracy such as RMSE and SSIM is not the best performer on our physical fidelity metrics. For instance, in the 90-frame long-term prediction for Spinodal Decomposition, VMamba achieves the best numerical accuracy with the lowest RMSE (0.02138) and highest SSIM (0.99627). However, the best physical fidelity, as measured by L-ETAP, is achieved by a different model, TAU (-1.958). This divergence illustrates that minimizing pixel-level error does not ensure optimal adherence to the underlying physical constraints of the system, such as phase fraction conservation. Therefore, relying solely on standard vision metrics is insufficient, and physical fidelity metrics are necessary to ensure that predictions are physically meaningful.

The architecture of model is more critical than

[6]Further implementation details see the extended version.

[7]Additional accuracy metrics, performance-cost figure, statistical variance analyses, cases and out-of-distribution evaluations see the extended version.

domain-specific and general-purpose methods. The overall top-performing model is VMamba, a Microstructural model that leverages a modern architecture (Mamba). It consistently excels, particularly in long-term predictions and complex tasks. However, another top model, especially in short-term tasks, is VMRNN, a Spatiotemporal model that also integrates a Mamba-like structure. The common thread between these top performers is their advanced architecture, which contrasts with the generally weaker long-term stability of older RNN-based models from both categories (e.g., PredRNN, ConvLSTM). This suggests the most promising path forward is combining state-of-the-art architectural designs with domain-specific considerations.

Model	Plane-wave propagation	Grain growth	Spinodal decomposition	Dendritic solidification
E3DLSTM	0.113	0.103	0.111	0.113
ConvGRU	0.044	0.195	0.039	0.044
PredRNN	0.096	0.087	0.097	0.096
ConvLSTM	0.095	0.089	0.097	0.095
VMamba	**0.021**	**0.006**	**0.007**	**0.021**
PredRNN++	0.093	0.088	0.097	0.098
MAU	0.087	0.093	0.090	0.094
SimVP	0.090	0.084	0.096	0.094
PredRNN.v2	0.088	0.088	0.095	0.097
TAU	0.086	0.082	0.090	0.097
SwinLSTM	0.097	0.098	0.105	0.106
VMRNN	0.044	0.034	0.035	0.039
SimVP.v2	0.102	0.083	0.091	0.093
PredFormer	0.085	0.083	0.086	0.089

Table 3: The time of different models in short-term prediction (10-10) of `MicroEvoEval`.

VMamba not only delivers state-of-the-art performance but also operates with an order-of-magnitude greater efficiency than most other models. Shown in Table 3, with an inference time of just 0.006s for the Grain Growth task, compared to the 0.08-0.1s range for most competitors, VMamba shows great potential of a trade-off between accuracy and speed. Its superior performance and efficiency make it a highly promising candidate for practical applications. This further highlights the transformative potential of modern architectures like Mamba.

Visual Analysis

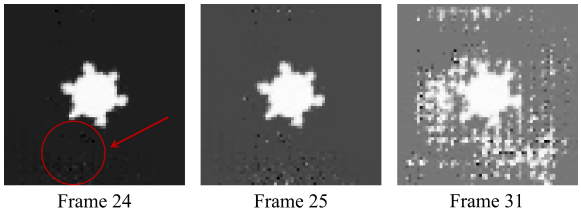

Figure 2: Case study on dendritic solidification.

In addition to quantitative metrics, we use visual analysis to explore long-term stability and to show the findings of our physical fidelity metrics.

Figure 2 shows why short-term accuracy is a poor indicator of long-term stability. Using the Dendritic Solidification task as an example, we observe that in an early prediction frame (Frame 24), a model like PredRNN introduces low-amplitude noise artifacts into the background, as highlighted by the red circle. Although this initial error may be minor, its effect becomes catastrophic during forecasting. This leads to a complete breakdown of the physical structure in Frame 31, where the prediction has diverged into a meaningless state.

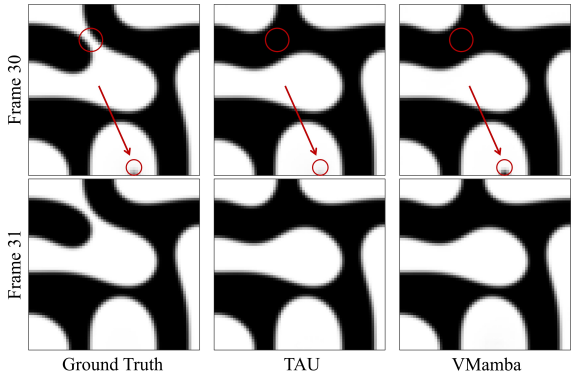

Figure 3: Case study on spinodal decomposition.

Figure 3 shows how our physical fidelity metrics can further distinguish between two strong models. Both models achieve high numerical scores, including the eventual dissolution of the smaller droplet highlighted by the red arrow. However, they differ in the speed of this process. The ground truth shows the droplet gradually shrinking from Frame 30 to Frame 31. TAU's prediction closely mimics this gradual dissolution rate. In contrast, VMamba predicts an overly accelerated dynamic where the droplet disappears more rapidly than in the ground truth.

Conclusion

In this work, we introduced `MicroEvoEval`, the first comprehensive benchmark to address critical gaps in evaluating deep learning for microstructure evolution. Our framework systematically compares domain-specific and general-purpose models using a multi-faceted evaluation of short- and long-term stability, numerical accuracy, physical fidelity and efficiency. Our results demonstrate the necessity of both long-term evaluation and physical fidelity metrics, as short-term numerical accuracy is a poor indicator of overall performance. Furthermore, we show that modern architectures, particularly state-space models, deliver superior performance and stability. These findings provide guidance for future research, pointing towards the development of more robust, physics-informed, and computationally efficient models for materials science. [8]

[8] Directions and limitations see the extended version.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (No. 11971336, 12571435).

References

Acar, P.; and Sundararaghavan, V. 2016. A Markov random field approach for modeling spatio-temporal evolution of microstructures. *Modelling and Simulation in Materials Science and Engineering*, 24(7): 075005.

Allen, S.; and Cahn, J. 1979. A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening. *Acta Metallurgica*, 27(6): 1085–1095.

Badalassi, V. E.; Ceniceros, H. D.; and Banerjee, S. 2003. Computation of multiphase systems with phase field models. *Journal of computational physics*, 190(2): 371–397.

Bhadeshia, H. 2001. The importance of microstructure in steels. *Science and Technology of Welding and Joining*, 6(2): 67–75.

Boettinger, W. J.; Coriell, S. R.; Greer, A.; Karma, A.; Kurz, W.; Rappaz, M.; and Trivedi, R. 2000. Solidification microstructures: recent developments, future directions. *Acta materialia*, 48(1): 43–70.

Cahn, J. W.; and Hilliard, J. E. 1958. Free energy of a nonuniform system. I. Interfacial free energy. *The Journal of Chemical Physics*, 28(2): 258–267.

Chang, Z.; Zhang, X.; Wang, S.; Ma, S.; Ye, Y.; Xinguang, X.; and Gao, W. 2021. Mau: A motion-aware unit for video prediction and beyond. *Advances in Neural Information Processing Systems*, 34: 26950–26962.

Chen, L.-Q. 2002. Phase-field models for microstructure evolution. *Annual Review of Materials Research*, 32: 113–140.

Fan, S.; Hitt, A. L.; Tang, M.; Sadigh, B.; and Zhou, F. 2024. Accelerate microstructure evolution simulation using graph neural networks with adaptive spatiotemporal resolution. *Machine Learning: Science and Technology*, 5(2): 025027. Open Access.

Farizhandi, A. A. K.; and Mamivand, M. 2023. Spatiotemporal prediction of microstructure evolution with predictive recurrent neural network. *Computational Materials Science*, 223: 112110.

Gao, Z.; Tan, C.; Wu, L.; and Li, S. Z. 2022. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3170–3180.

Greenwood, M.; Shampur, K.; Ofori-Opoku, N.; Pinomaa, T.; Wang, L.; Gurevich, S.; and Provatas, N. 2018. Quantitative 3D phase field modelling of solidification using next-generation adaptive mesh refinement. *Computational Materials Science*, 142: 153–171.

Guo, Z.; and Xiong, S. 2015. On solving the 3-D phase field equations by employing a parallel-adaptive mesh refinement (Para-AMR) algorithm. *Computer Physics Communications*, 190: 89–97.

Hasan, M. M.; Eger, Z. E.; Senthilnathan, A.; and Acar, P. 2023. Microstructure-sensitive material design with physics-informed neural networks. In *AIAA SCITECH 2023 Forum*, 0539.

Jing-jie, L.; Zhu, C.-s.; Tian-yu, L.; Zi-hao, G.; Shuo, L.; Hang, C.; and Jin-tao, M. 2025. Research on spatiotemporal prediction model of grain microstructure evolution based on VMamba network. *Computational Materials Science*, 252: 113793.

Jou, H.-J.; Leo, P. H.; and Lowengrub, J. S. 1997. Microstructural evolution in inhomogeneous elastic media. *Journal of Computational Physics*, 131(1): 109–148.

Kamachali, R. D.; Schwarze, C.; Lin, M.; Diehl, M.; Shanthraj, P.; Prahl, U.; Steinbach, I.; and Raabe, D. 2018. Numerical benchmark of phase-field simulations with elastic strains: precipitation in the presence of chemo-mechanical coupling. *Computational Materials Science*, 155: 541–553.

Kobayashi, R. 1993. Modeling and numerical simulations of dendritic crystal growth. *Physica D: Nonlinear Phenomena*, 63(3-4): 410–423.

Lanzoni, D.; Albani, M.; Bergamaschini, R.; and Montalenti, F. 2022. Morphological evolution via surface diffusion learned by convolutional, recurrent neural networks: Extrapolation and prediction uncertainty. *Physical Review Materials*, 6(10): 103801.

Liu, C.; and Shen, J. 2003. A phase field model for the mixture of two incompressible fluids and its approximation by a Fourier-spectral method. *Physica D: Nonlinear Phenomena*, 179(3-4): 211–228.

Mao, H.; Xie, C.; Pan, J.; Cao, Q.; Zhang, X.; Luo, Y.; Du, Y.; and Ning, H. 2024. Spatiotemporal prediction of solidified dendrites based on convolutional long-short-term neural network. *Materials Today Communications*, 41: 110634.

Mo, Y.; Tang, X.; Meng, L.; Qiao, J.; and Yao, X. 2021. Spatial–Temporal evolution of shear banding in bulk metallic glasses. *Materials Science and Engineering: A*, 800: 140286.

Moelans, S.; Blanpain, B.; and Wollants, P. 2008. An introduction to phase-field modeling of microstructure evolution. *Calphad*, 32(2): 268–294.

Noguchi, S.; Aihara, S.; and Inoue, J. 2024. Microstructure estimation by combining deep learning and phase transformation model. *ISIJ International*, 64(1): 142–153.

Olson, G. 1997. Computational design of hierarchically structured materials. *Science*, 277(5330): 1237–1242.

Porter, D.; and Easterling, K. 1992. *Phase Transformations in Metals and Alloys*. Chapman & Hall.

Reed, R. 2008. *The Superalloys: Fundamentals and Applications*. Cambridge University Press.

Sang, S.; Xu, C.; Fan, J.; Miao, D.; Side, C.; and Wang, Z. 2023. Accurate prediction of microstructure of composites using machine learning. *Advanced Theory and Simulations*, 6(2): 2200674.

Seetharaman, V.; et al. 2021. Microstructure evolution modeling in materials processing: Current status and future directions. *Acta Materialia*, 203: 116498.

Shen, J.; and Yang, X. 2009. An efficient moving mesh spectral method for the phase-field model of two-phase flows. *Journal of computational physics*, 228(8): 2978–2992.

Steinbach, I. 2013. Phase-field model for microstructure evolution at the mesoscopic scale. *Annual Review of Materials Research*, 43(1): 89–107.

Steinbach, I.; Pezzolla, F.; Nestler, B.; Seeßelberg, M.; Prieler, R.; Schmitz, G. J.; and Rezende, J. L. 1996. A phase field concept for multiphase systems. *Physica D: Nonlinear Phenomena*, 94(3): 135–147.

Tan, C.; Gao, Z.; Li, S.; and Li, S. Z. 2025. SimVPv2: Towards simple yet powerful spatiotemporal predictive learning. *IEEE Transactions on Multimedia*.

Tan, C.; Gao, Z.; Wu, L.; Xu, Y.; Xia, J.; Li, S.; and Li, S. Z. 2023. Temporal attention unit: Towards efficient spatiotemporal predictive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18770–18782.

Tandogan, I. T.; Budnitzki, M.; and Sandfeld, S. 2025. A multi-physics model for the evolution of grain microstructure. *International Journal of Plasticity*, 185: 104201.

Tang, S.; Li, C.; Zhang, P.; and Tang, R. 2023. Swinlstm: Improving spatiotemporal prediction accuracy using swin transformer and lstm. In *Proceedings of the IEEE/CVF international conference on computer vision*, 13470–13479.

Tang, Y.; Dong, P.; Tang, Z.; Chu, X.; and Liang, J. 2024. Vmrnn: Integrating vision mamba and lstm for efficient and accurate spatiotemporal forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5663–5673.

Tang, Y.; Qi, L.; Xie, F.; Li, X.; Ma, C.; and Yang, M.-H. 2025. Video Prediction Transformers without Recurrence or Convolution. arXiv:2410.04733.

Tourret, D.; Liu, H.; and LLorca, J. 2022. Phase-field modeling of microstructure evolution: Recent applications, perspectives and challenges. *Progress in Materials Science*, 123: 100810.

Wang, Y.; Gao, Z.; Long, M.; Wang, J.; and Yu, P. S. 2018. Predrnn++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In *International conference on machine learning*, 5123–5132. PMLR.

Wang, Y.; Wu, H.; Zhang, J.; Gao, Z.; Wang, J.; Yu, P. S.; and Long, M. 2022. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 2208–2225.

Yang, K.; Cao, Y.; Zhang, Y.; Fan, S.; Tang, M.; Aberg, D.; Sadigh, B.; and Zhou, F. 2021. Self-supervised learning and prediction of microstructure evolution with convolutional recurrent neural networks. *Patterns*, 2(5).