PhySiViT: A Physics Simulation Vision Transformer

Jessica Ezemba
Carnegie Mellon University
Pittsburgh, PA, USA
jezemba@andrew.cmu.edu

James Afful
Iowa State University
Ames, IA, USA
affulj@iastate.edu

Mei-Yu Wang
Pittsburgh Supercomputing Center
Pittsburgh, PA, USA
mwang7@psc.edu

Abstract

Modern scientific computing generates massive simulation data across physics domains, yet researchers lack general-purpose tools for efficient analysis. While vision transformers like CLIP and DINO have revolutionized natural image analysis, no equivalent exists for physics simulation data. This project trains a custom Vision Transformer on "The Well" dataset, a 15 TB collection of diverse physics simulations. Using only 7 million images (compared to >100 million for CLIP/DINOv2), we trained our physics foundation model in 22 hours on a single Cerebras CS-3 server. Despite reduced training scale, our model demonstrates competitive classification performance while exceeding at physics-specific tasks: temporal forecasting (R^2 = 0.33 vs. DINOvs2's 0.23) and physics clustering (silhouette score = 0.232 vs. DINOv2's 0.195). Model weights are available at https://huggingface.co/JessicaE/physics-vit-standard. This work demonstrates that efficient, domain-focused foundation models can achieve better performance in specialized scientific domains.

CCS Concepts

• **Computing methodologies** → **Machine learning**; *Image representations*; • **Mathematics of computing** → *Time series analysis*; *Computational physics.*

Keywords

vision transformers, scientific machine learning, embeddings, temporal forecasting, simulation data

ACM Reference Format:

Jessica Ezemba, James Afful, and Mei-Yu Wang. 2025. PhySiViT: A Physics Simulation Vision Transformer. In *Proceedings of The International Conference for High Performance Computing, Networking, Storage and Analysis (SC '25).* ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Problem and Proposed Solution

Modern scientific computing generates massive physics simulation data, but researchers lack general-purpose analysis tools. Existing ML methods require domain-specific solutions, limiting cross-disciplinary insights [2].

SC '25, St. Louis, MO

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

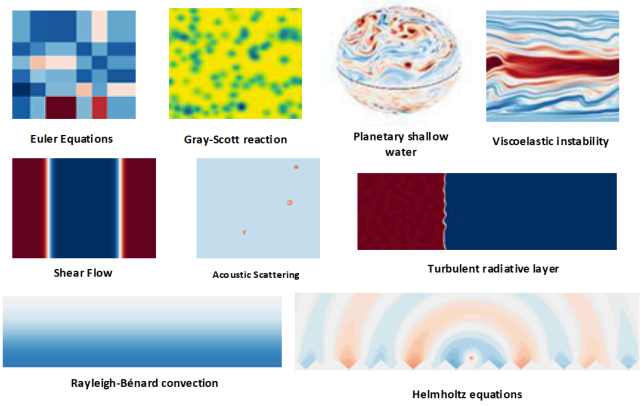

Figure 1: Simulation Classes Selected from the Well Dataset.

While vision transformers like CLIP [4] and DINOv2 [3] transformed natural image analysis, no foundation model exists for physics simulations. This project addresses these limitations by training a Vision Transformer on "The Well" dataset [2], using ≈7 million physics simulation images trained over 22 hours on a Cerebras CS-3 server. We evaluate performance against CLIP and DINOv2 across downstream tasks, testing whether domain-specific training creates effective foundation models for scientific computing.

1.1 Methods

We processed "The Well" dataset, selecting 2D spatiotemporal simulations yielding ≈7 million PNG images from 11 physics domains (Fig 1). Images underwent expand-to-square transformation with gray padding, 224×224 resizing, and trajectory-aware splitting to prevent data leakage. We also adopted physics-specific color augmentation (70% scientific colormaps, 30% grayscale) to force spatial pattern learning.

Our ViT-Huge model (1280 dimensions, 32 layers, 16 heads) trained for 78,372 steps using Adam with cosine decay. We compared against DINOv2 Giant and CLIP Large across: **Classification** (logistic regression, accuracy), **Temporal Forecasting** (ridge regression predicting t+1 from t, R^2/MSE), and **Embedding Quality** (silhouette scoring on visualizations).

2 Results

Table 1 presents the comparative performance of PhySiViT against established foundation models across three physics evaluation tasks. While DINOv2 and CLIP achieved marginally higher classification accuracy (0.99 vs. 0.98), PhySiViT outperformed both models on physics-domain tasks.

Table 1: Performance comparison on evaluation tasks.

Model	Accuracy (↑)	R^2 (↑)	MSE (↓)	Silhouette (↑)
PhySiViT	0.98	**0.33**	**0.57**	**0.23**
DINO V2 Giant	**0.99**	0.23	0.62	0.20
CLIP-ViT Large	**0.99**	0.22	0.63	0.19

Most notably, PhySiViT achieved 43% better temporal forecasting performance ($R^2 = 0.33$ vs. DINOv2's 0.23) and better physics clustering quality (silhouette score = 0.23). Qualitative analysis revealed that while DINOv2 and CLIP embeddings showed similar patterns from natural image training, PhySiViT embeddings exhibited distinct physics-informed structure that better separated physical domains as shown in Figure 2.

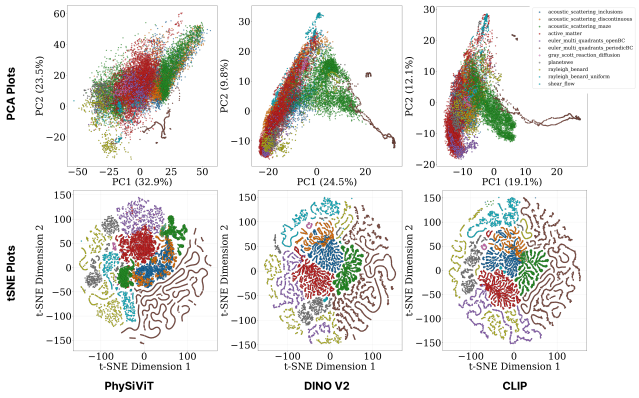

Figure 2: Embedding quality visualization comparing PhySiViT, DINOv2 Giant, and CLIP-ViT Large.

Our physics-specialized Vision Transformer was trained on 7 million images in 22.83 hours using a single Cerebras [1] CS-3 node, achieving 70× speedup over CPU and 12.5× over GPU systems. These results demonstrate that despite dramatically reduced training scale compared to general-purpose models (>100 million images), domain-specific foundation models can achieve superior performance.

3 Conclusion & Future Work

PhySiViT demonstrates that domain-specific foundation models can achieve better performance on specialized scientific tasks while maintaining competitive general classification capabilities. This work establishes the viability of efficient, domain-focused foundation models for accelerating scientific discovery.

Several promising directions emerge from this work:

- **Hyperparameter tuning**: Optimize model architecture and training parameters specifically for physics domains.
- **Improving data transformations**: Develop physics-aware augmentation strategies beyond colormap randomization.
- **Comparing with other downstream tasks**: Evaluate performance on additional scientific computing applications (PDE solving, anomaly detection).
- **Finetuning DINOv2 and CLIP**: Investigate whether physics-specific finetuning can bridge the performance gap.
- **Expanding to DINOv3**: Compare against newer foundation model architectures as they become available.

This research opens pathways for developing specialized foundation models across scientific domains, potentially transforming how machine learning supports computational science and engineering applications.

Acknowledgments

This work was made possible thanks to the ByteBoost cybertraining program which is funded by the National Science Foundation Cybertraining awards: 2320990, 2320991, and 2320992, and the Neocortex project, the ACES platform, and the Ookami cluster. The Neocortex project is supported by National Science Foundation award number 2005597. The ACES (Accelerating Computing for Emerging Sciences) platform was funded by National Science Foundation award number 2112356. The Ookami cluster is supported by National Science Foundation award number 1927880. We also thank Sylvia Howland (Cerebras) for technical support for using the Cerebras CS-3 server.

References

[1] Cerebras Systems. 2024. *Cerebras CS-3: the world's fastest and most scalable AI accelerator.* https://www.cerebras.ai/blog/cerebras-cs3 Product announcement/blog.

[2] Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina J. Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Stuart B. Dalziel, Drummond B. Fielding, Daniel Fortunato, Jared A. Goldberg, Keiya Hirashima, Yan-Fei Jiang, Rich R. Kerswell, Suryanarayana Maddu, Jonah Miller, Payel Mukhopadhyay, Stefan S. Nixon, Jeff Shen, Romain Watteaux, Bruno Régaldo-Saint Blancard, François Rozet, Liam H. Parker, Miles Cranmer, and Shirley Ho. 2024. The Well: a Large-Scale Collection of Diverse Physics Simulations for Machine Learning. *arXiv preprint arXiv:2412.00568* (2024). https://arxiv.org/abs/2412.00568 NeurIPS 2024 dataset track poster (see NeurIPS virtual site).

[3] Maxime Oquab, Timothée Darcet, Julien Mairal, and Piotr Bojanowski. 2023. DINOv2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193* (2023). https://arxiv.org/abs/2304.07193 Published in Transactions on Machine Learning Research, 2024.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139).* PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html