Physics-Aware AI at Scale: Neural Compression and Vision Transformers for Simulation Data

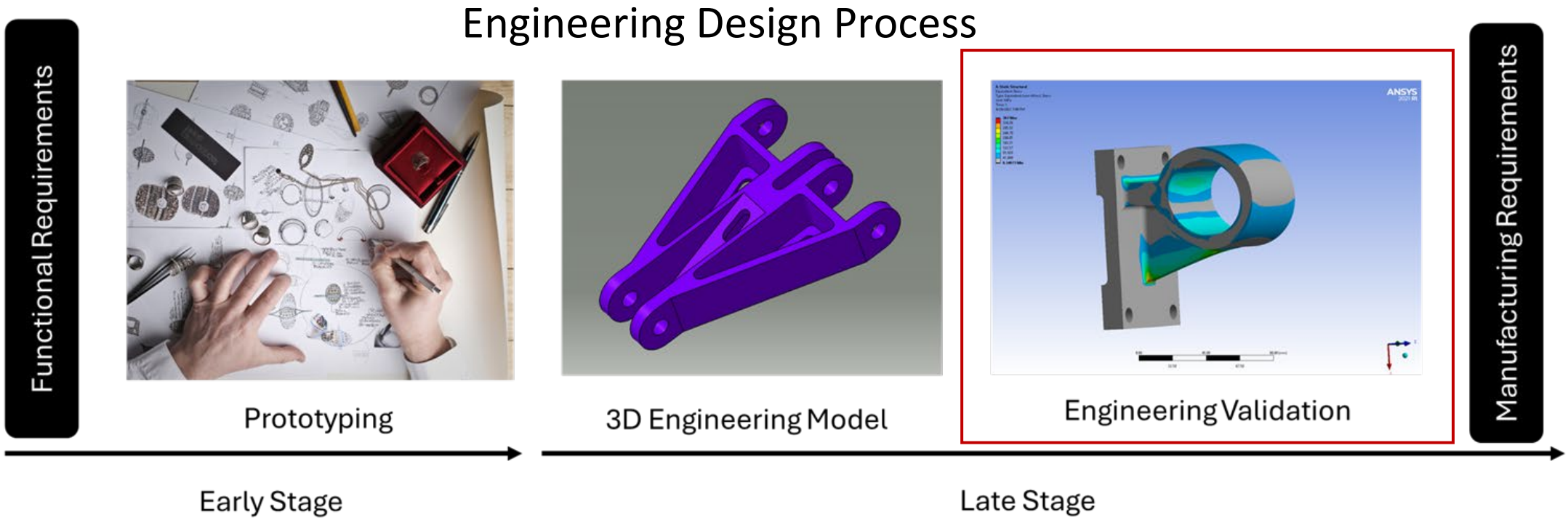
Jessica Ezemba

jezemba@andrew.cmu.edu

Jessicaezemba.com

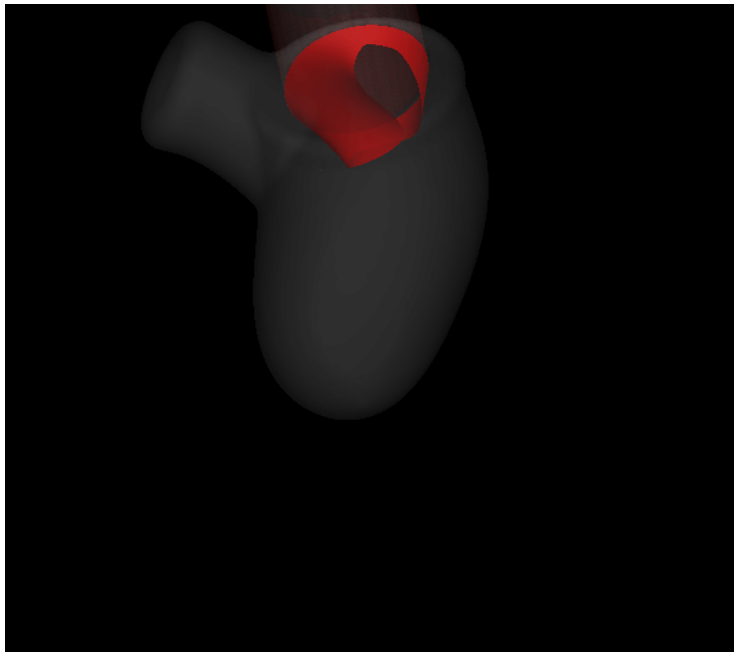
Carnegie Mellon University

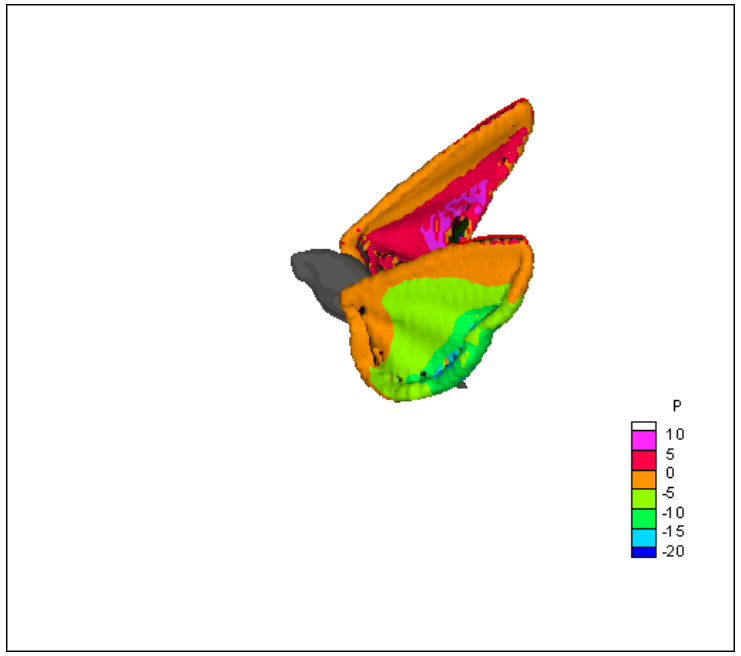
Simulation is a key tool used by engineering designers to verify initial functional requirements

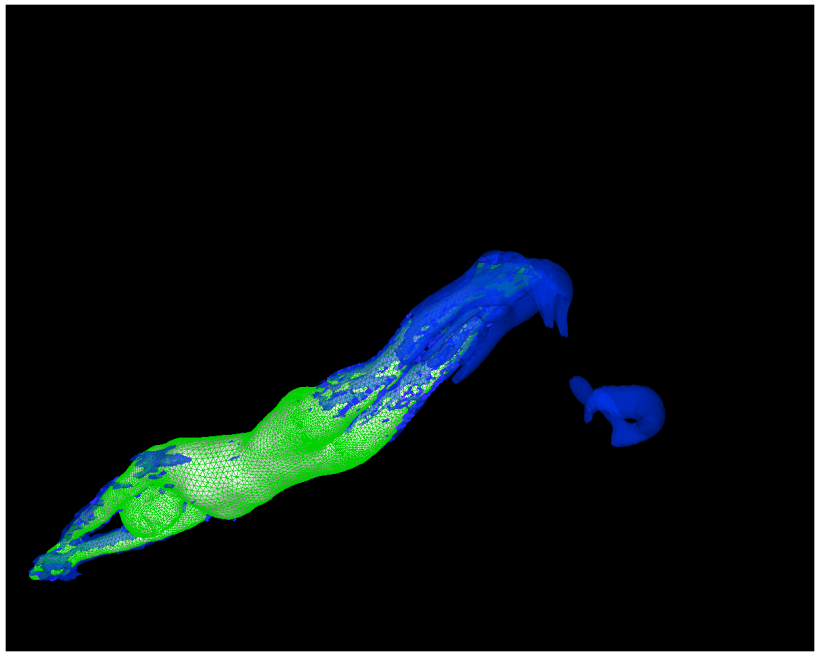


Scientific ML is a new and upcoming field that is transforming how we simulate and understand complex systems.

Physics-based simulations can exceed petabytes of data and have varied representation outputs

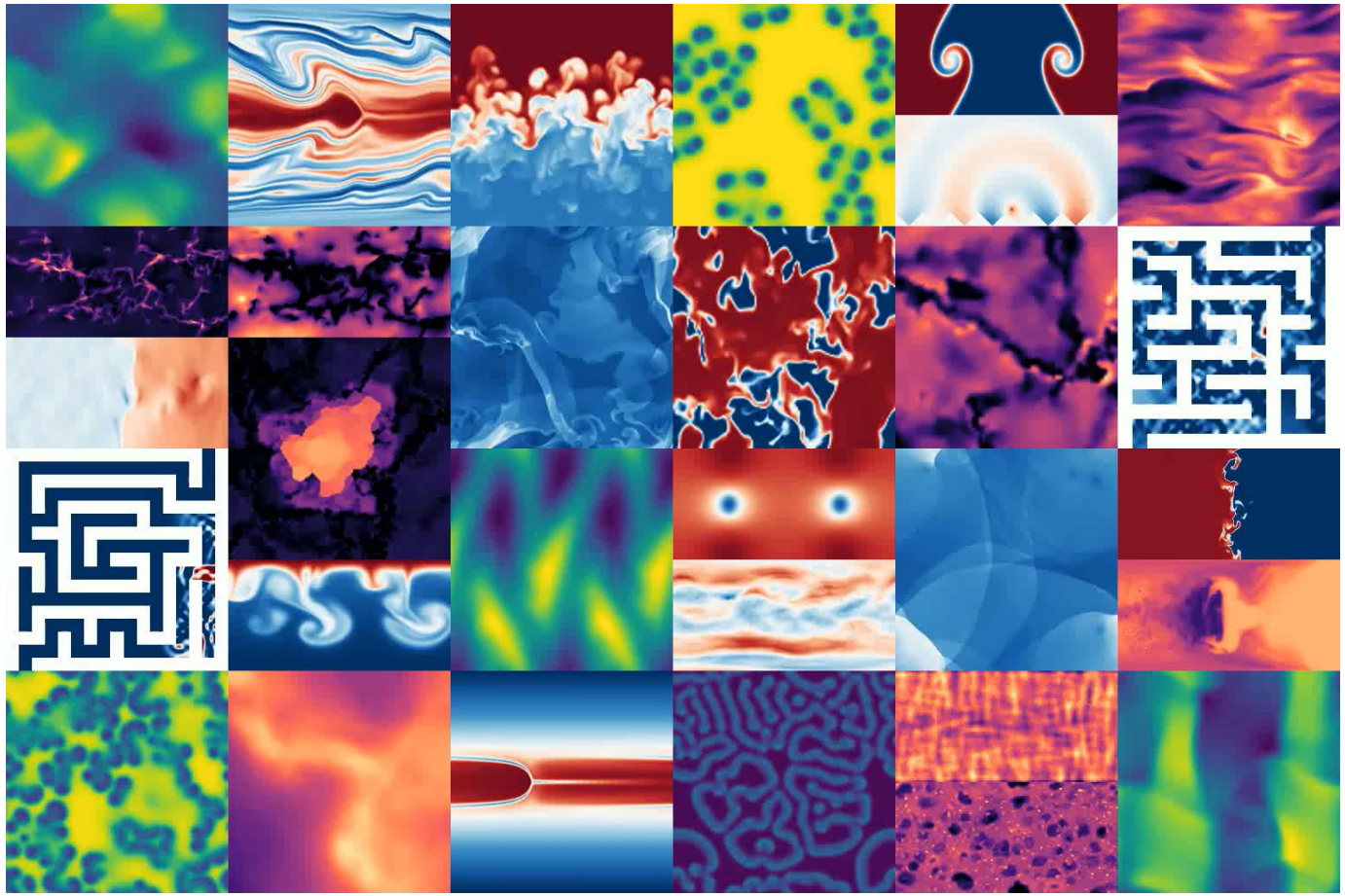






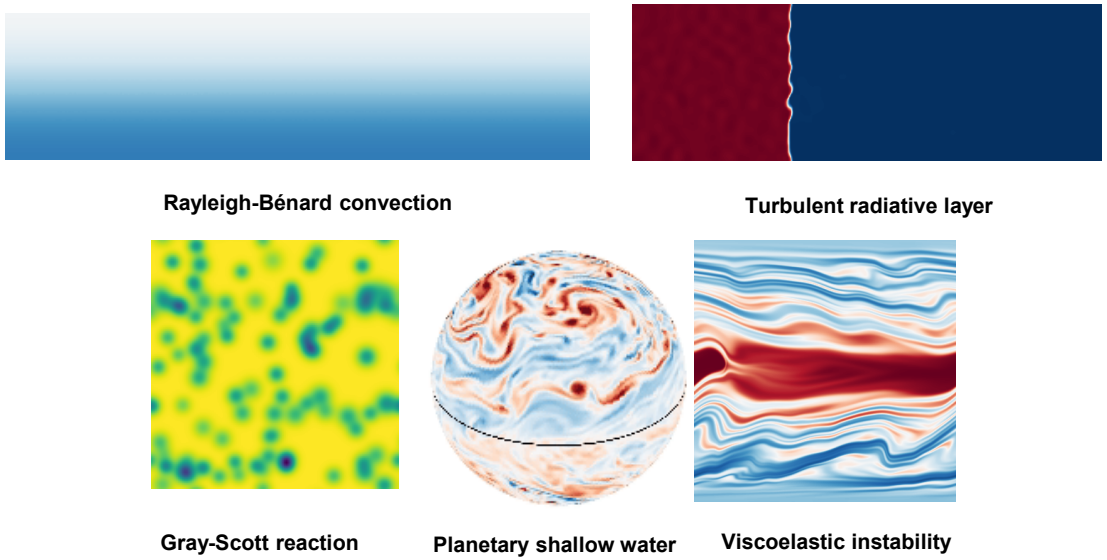
Research from Rajat Mittal Group at John Hopkins

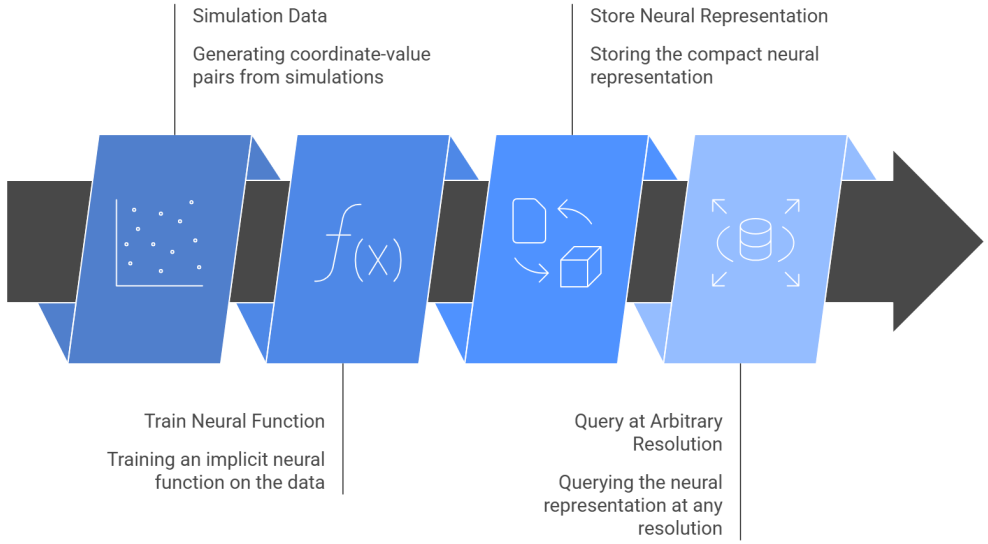
More Accessible Benchmarks have been Released but Still Exceed Terabytes of Data



Ohana, R., McCabe, M., Meyer, L., Morel, R., Agocs, F., Beneitez, M., ... & Ho, S. (2024). The well: a large-scale collection of diverse physics simulations for machine learning. *Advances in Neural Information Processing Systems*, *37*, 44989-45037.

Two paradigms have emerged to make this data tractable: learning over compact visual representations and compressing raw simulation data with neural networks.





Learning Over Reduced Representations

Implicit Neural Compression

[1] Ezemba, J., Afful, J., & Wang, M. Y. (2025). PhySiViT: A Physics Simulation Vision Transformer. Super Computing Conference 2025 (SC25)
[2]] Ezemba, J., Afful, J., & Wang, M. Y. (2025). Semantic-aware Implicit Neural Compression for Physics Simulations. Platform for Advanced Scientific Computing (PASC) Conference 2026 (Accepted)

The idea of learning over reduced representation has been used in many contexts

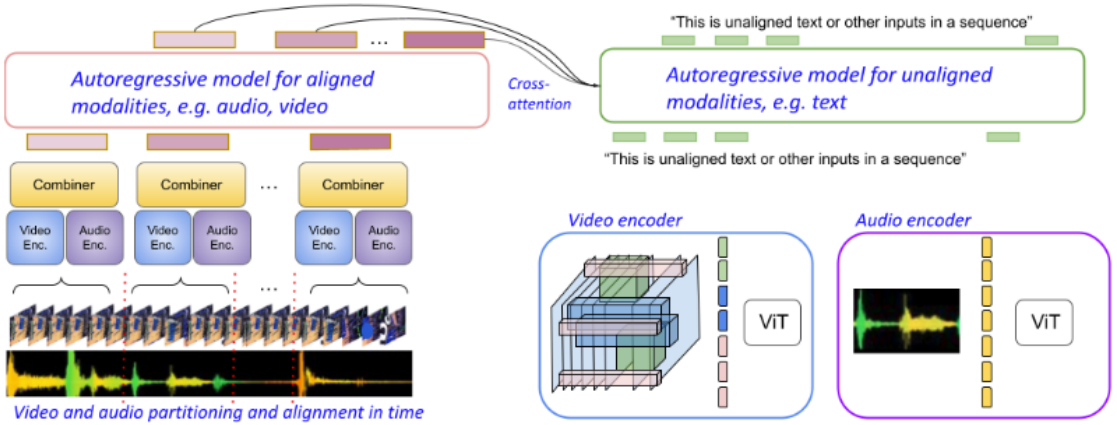
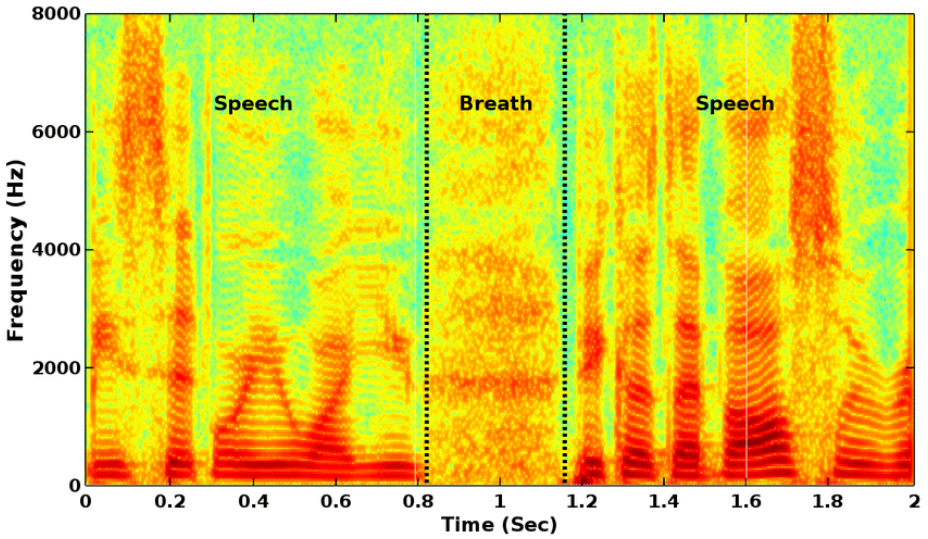


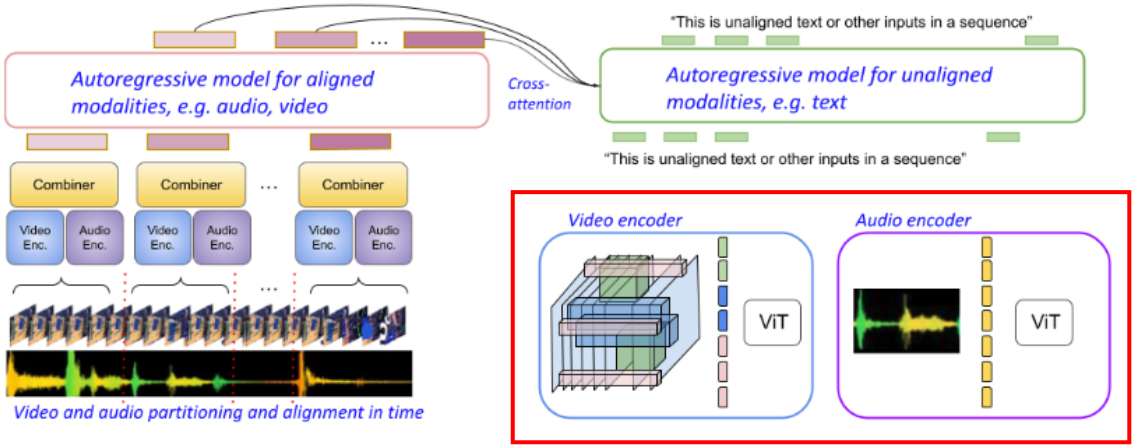
Figure 2. The Mirasol3B model architecture consists of an autoregressive model for the time-aligned modalities, such as audio and video, which are partitioned in chunks (left) and an autoregressive model for the unaligned context modalities, which are still sequential, e.g., text (right). This allows adequate computational capacity to the video/audio time-synchronized inputs, including processing them in time autoregressively, before fusing with the autoregressive decoder for unaligned text (right). Joint feature learning is conducted by the Combiner, balancing the need for compact representations and allowing sufficiently informative features to be processed in time.



Spectrogram

Piergiovanni, A. J., Noble, I., Kim, D., Ryoo, M. S., Gomes, V., & Angelova, A. (2024). Mirasol3b: A multimodal autoregressive model for time-aligned and contextual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 26804-26814).

Previous Research has Accelerated Multimodal Understanding with Audio using Image Embedding



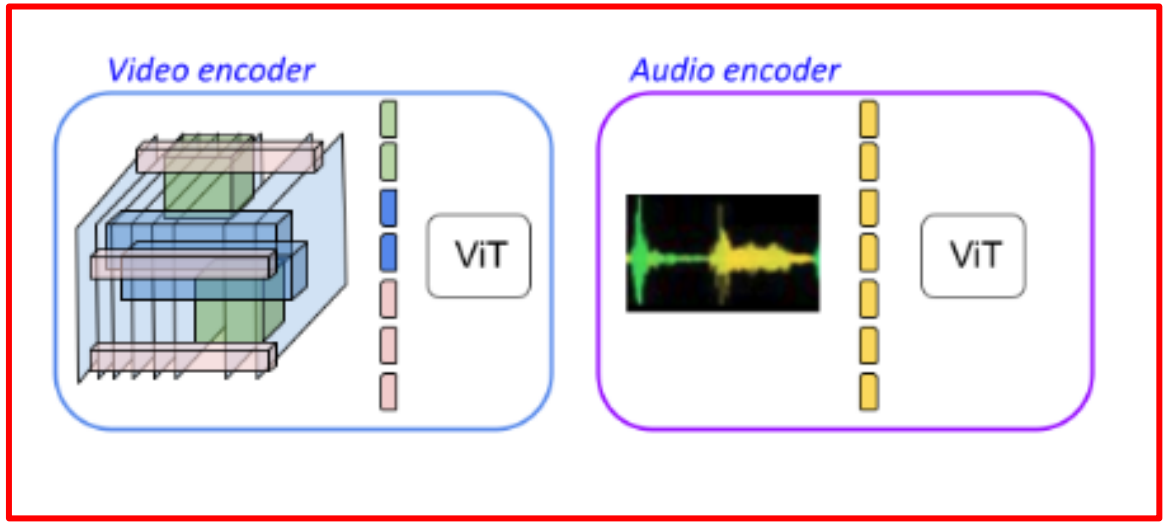


Figure 2. The Mirasol3B model architecture consists of an autoregressive model for the time-aligned modalities, such as audio and video, which are partitioned in chunks (left) and an autoregressive model for the unaligned context modalities, which are still sequential, e.g., text (right). This allows adequate computational capacity to the video/audio time-synchronized inputs, including processing them in time autoregressively, before fusing with the autoregressive decoder for unaligned text (right). Joint feature learning is conducted by the Combiner, balancing the need for compact representations and allowing sufficiently informative features to be processed in time.

Piergiovanni, A. J., Noble, I., Kim, D., Ryoo, M. S., Gomes, V., & Angelova, A. (2024). Mirasol3b: A multimodal autoregressive model for time-aligned and contextual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 26804-26814).

No existing model can learn cross-domain, physics-aware representations from raw simulation data

✓ a photo of **guacamole**, a type of food.

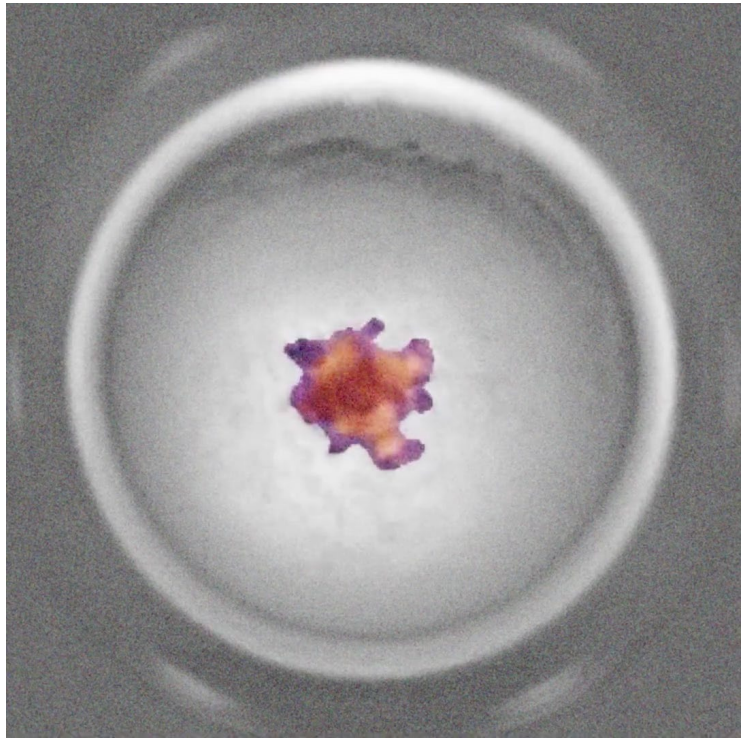
✕ a photo of **ceviche**, a type of food.

✕ a photo of **edamame**, a type of food.

✕ a photo of **tuna tartare**, a type of food.

✕ a photo of **hummus**, a type of food.

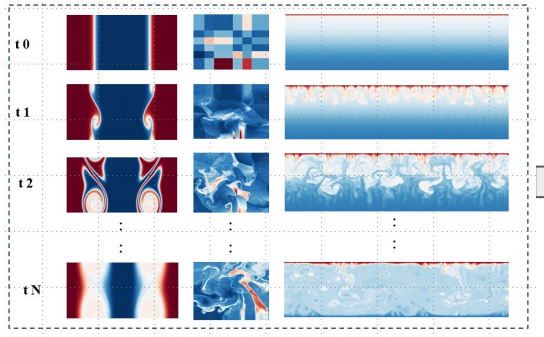
CLIP



DINO V3

BYTEBOOST
CYBERTRAINING

Design and train a physics-aware Vision Transformer (ViT) model using self-supervised learning on "The Well" simulation dataset.



Raw Simulation Trajectories

Preprocessing:

- expand_to_square padding
- physics-aware colormaps (viridis, plasma)
- temporal split by trajectory

12 Layers

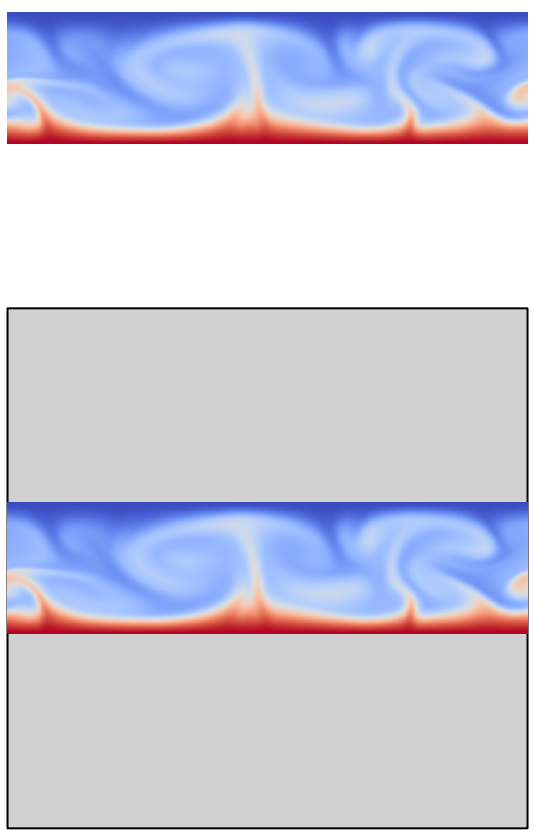
768 dim

12 heads

Evaluation:

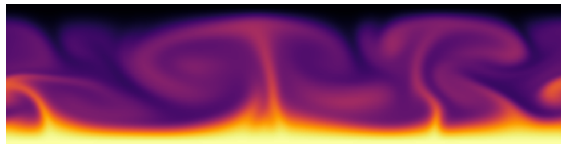
- Domain Classification
- Temporal Forecasting
- Embedding Visualization

Traditional computer vision augments (random crops, rotations) can break underlying physics meaning

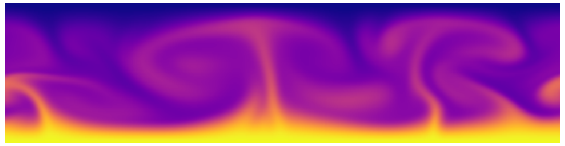


Cropping

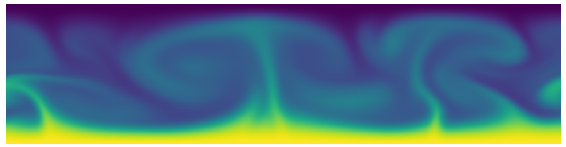
Black and White



Inferno



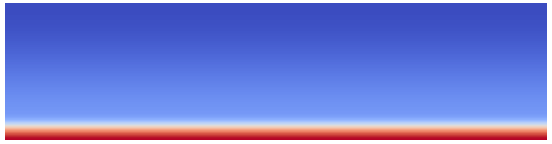
Plasma



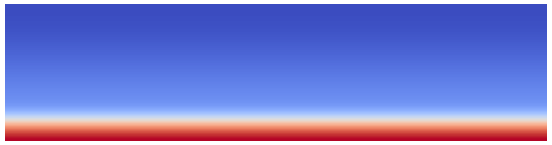
Viridis

Color Augmentation

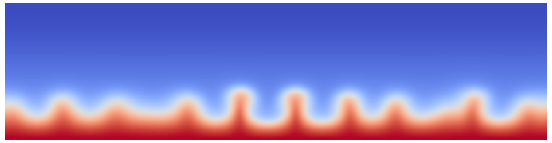
t = 2



t = 10

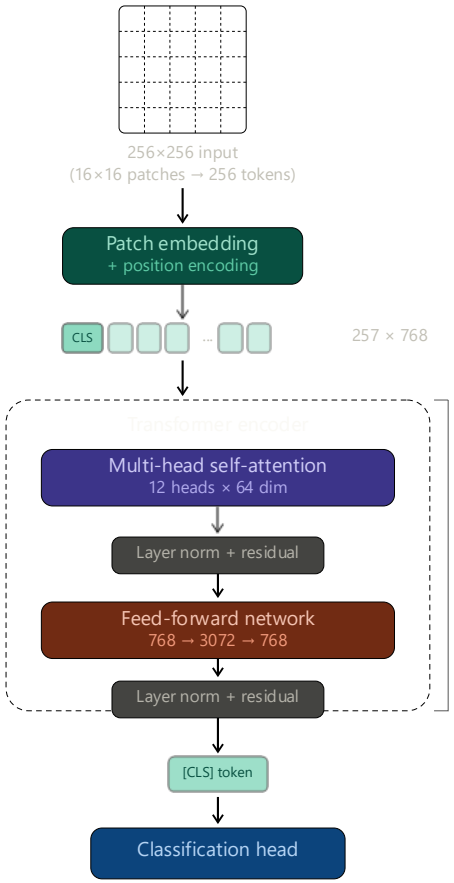


t = 50

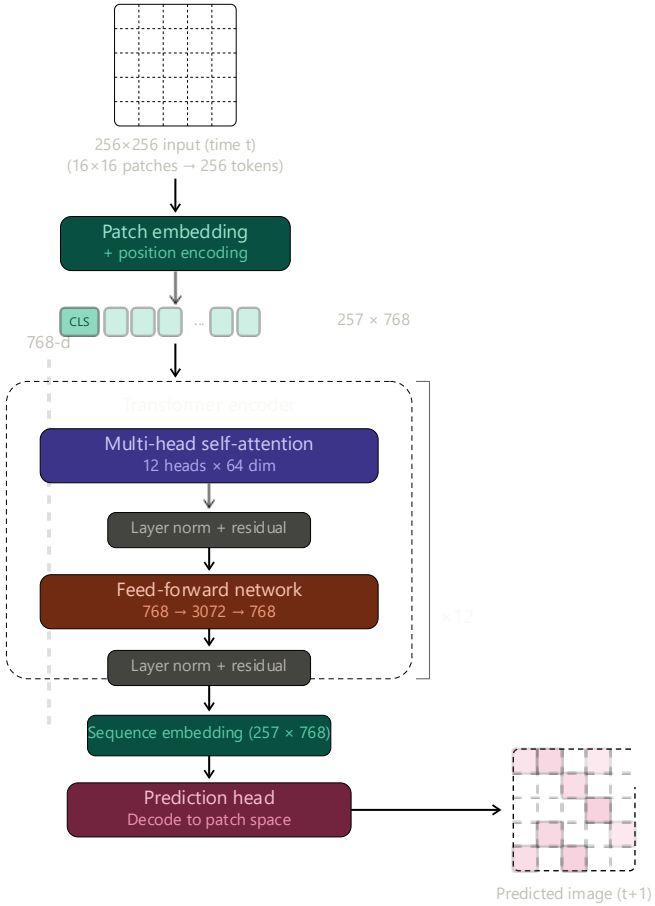


Random Augmentation

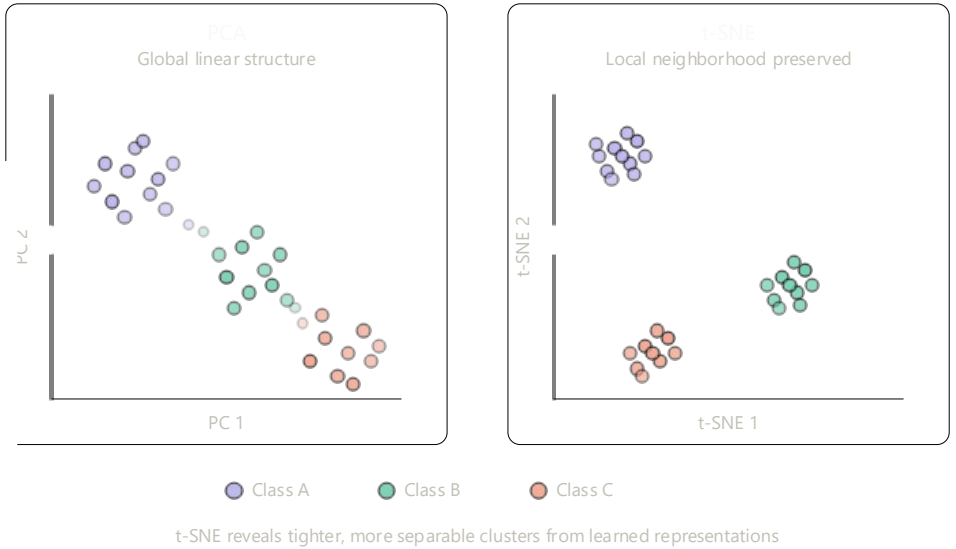
We evaluated the vision transformer on two downstream tasks and visualized the learned embedding space



Classification of Simulation



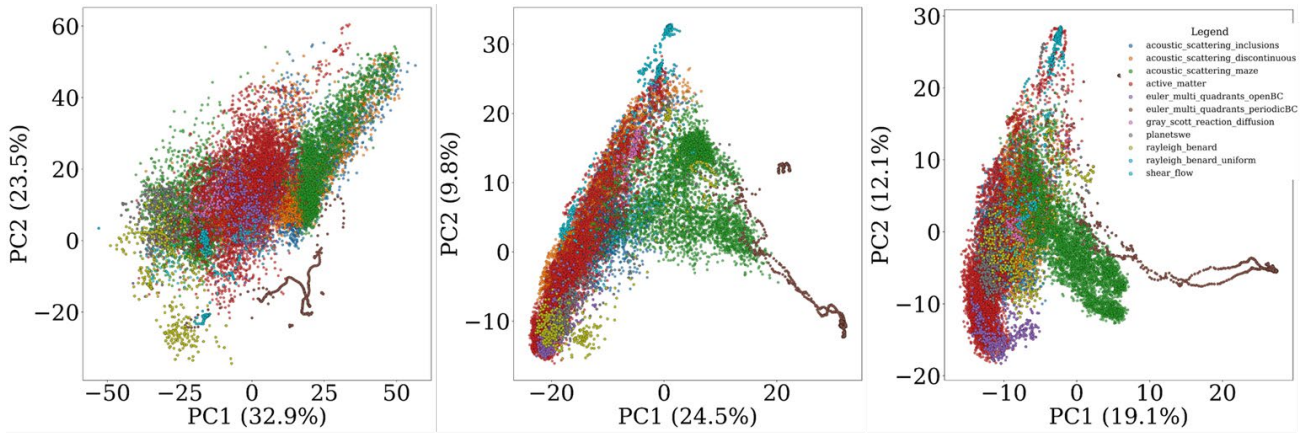
Next Time Step Prediction

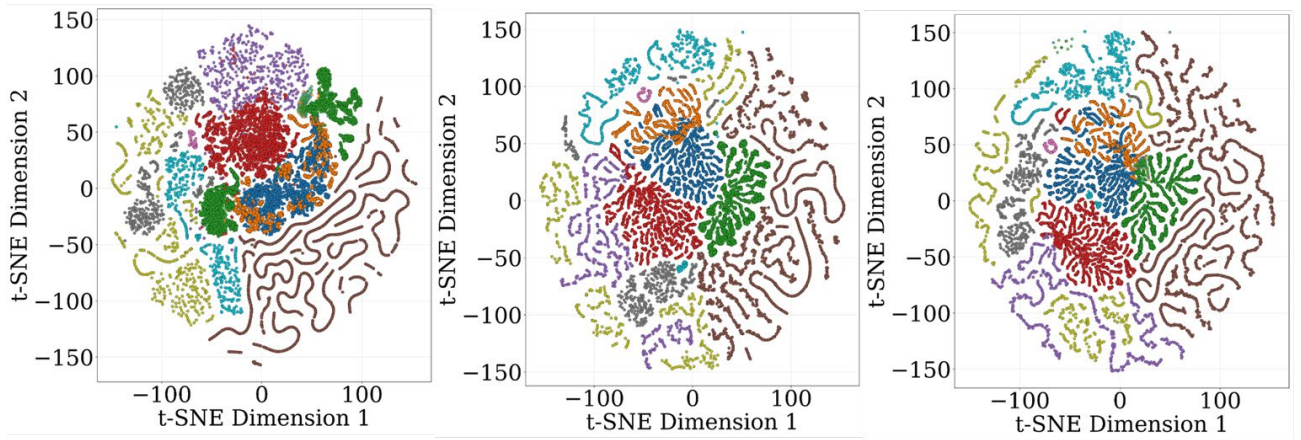


Embedding Space Visualization

BYTEBOOST
CYBERTRAINING

Physics ViT achieved competitive classification performance, demonstrating the power of targeted, physics-aware foundation models over general-purpose alternatives





PhySiViT | DINO V2 | CLIP

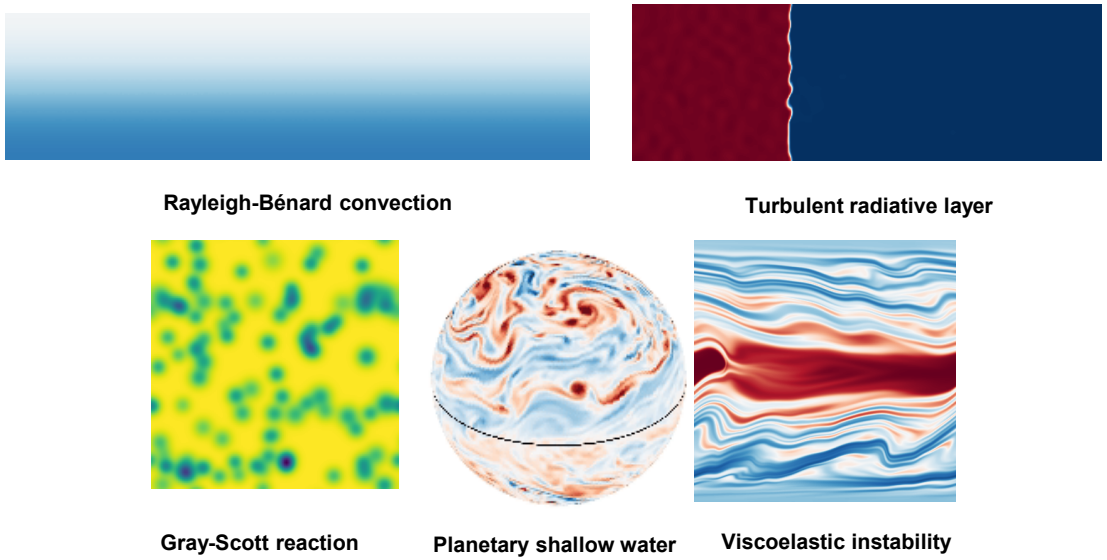
Model	Accuracy (↑)	R^2 (↑)	MSE (↓)	Silhouette (↑)
PhySiViT	0.98	**0.33**	**0.57**	**0.23**
DINOv2 Giant	**0.99**	0.23	0.62	0.20
CLIP-ViT Large	0.99	0.22	0.63	0.19

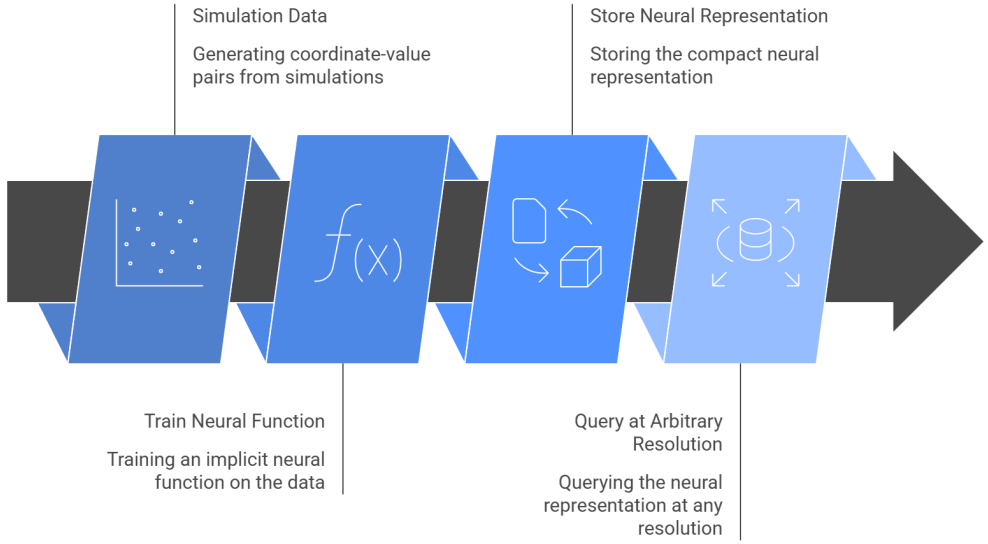
Training Speed Comparison (Single Device with 70 k Images)

Compute Device	Total Time (hours)
AMD EPYC 7702P (CPU)	14.03
NVIDIA H100 80GB (GPU)	2.50
Cerebras CS-3	**0.20**

Carnegie Mellon University

Two paradigms have emerged to make this data tractable: learning over compact visual representations and compressing raw simulation data with neural networks.





Learning Over Reduced Representations

Implicit Neural Compression

[1] Ezemba, J., Afful, J., & Wang, M. Y. (2025). PhySiViT: A Physics Simulation Vision Transformer. Super Computing Conference 2025 (SC25)
[2]] Ezemba, J., Afful, J., & Wang, M. Y. (2025). Semantic-aware Implicit Neural Compression for Physics Simulations. Platform for Advanced Scientific Computing (PASC) Conference 2026 (Accepted)

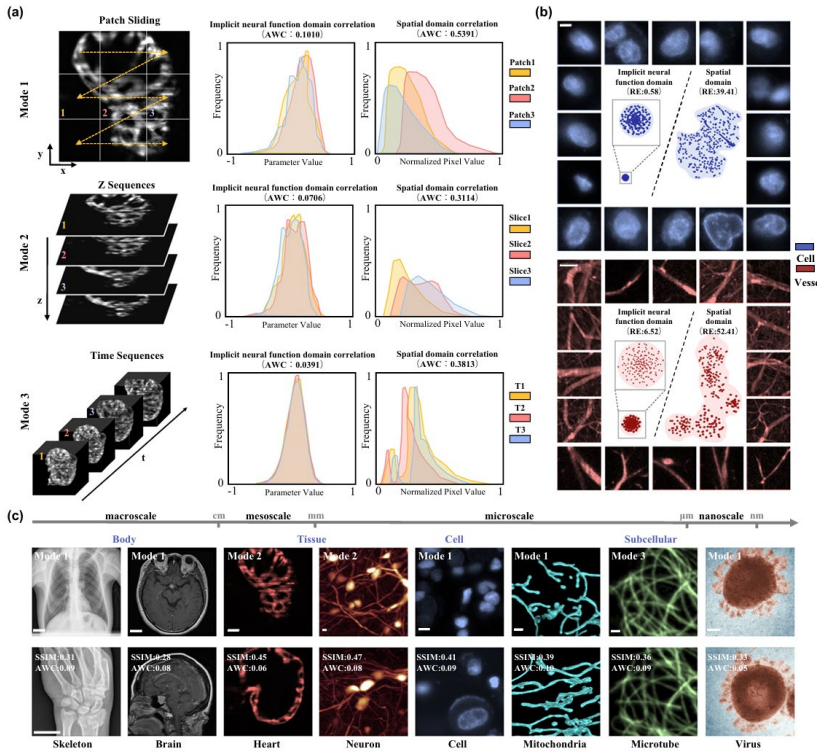
Lossy compression techniques are not new and have been used for many applications

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

Where:

- $H(X)$ = entropy (in bits per symbol)
- $p(x_i)$ = probability of symbol x_i occurring
- n = number of unique symbols

Traditional Compression

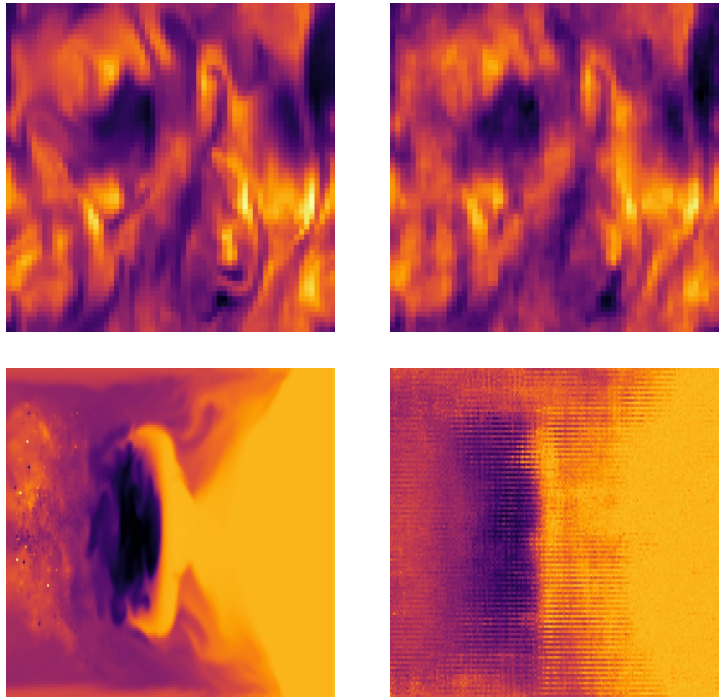


Semantic Neural Compression

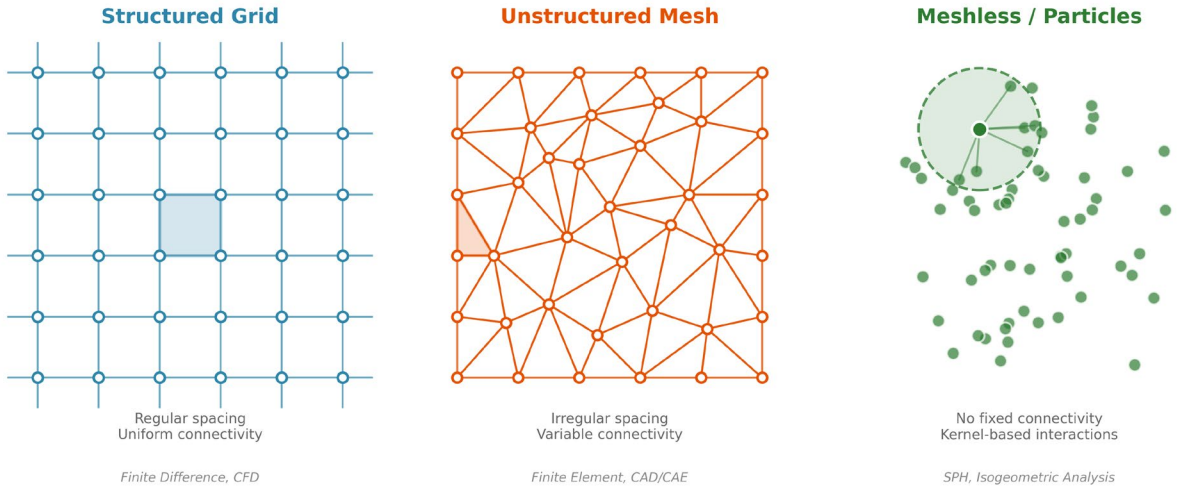
Di, S., et al (2025). A survey on error-bounded lossy compression for scientific datasets. *ACM computing surveys*

Ma, Y., et al (2024). Semantic redundancy-aware implicit neural compression for multidimensional biomedical image data. *Communications Biology*

Learn continuous implicit functions f(x,y,z,t) enabling ML workflows and automated discovery



Lossy compression
destroying physics



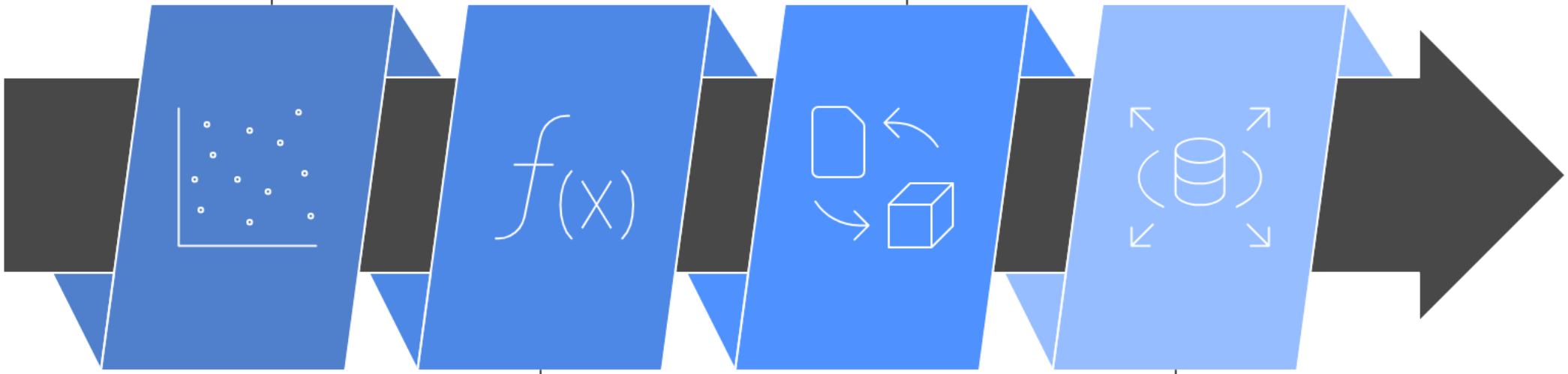
Mesh heterogeneity
(structured/unstructured/meshless)

Simulation Data

Generating coordinate-value pairs from simulations

Store Neural Representation

Storing the compact neural representation



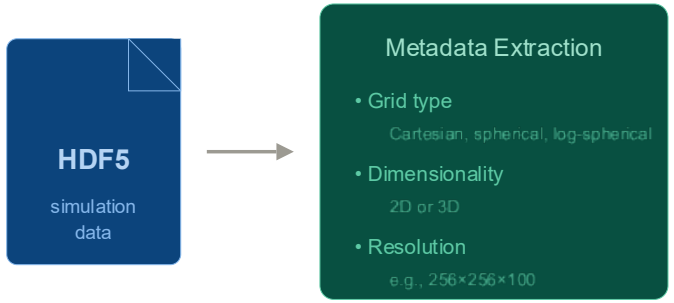
Train Neural Function

Training an implicit neural function on the data

Query at Arbitrary Resolution

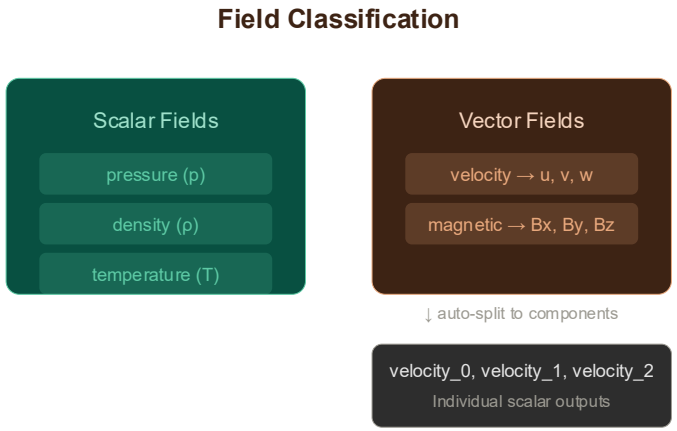
Querying the neural representation at any resolution

Different simulations have different setups

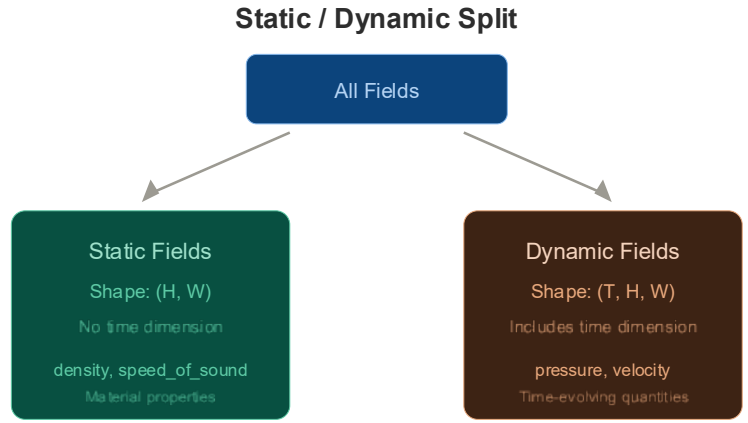


Auto-detect simulation setup from file structure

Metadata Extraction



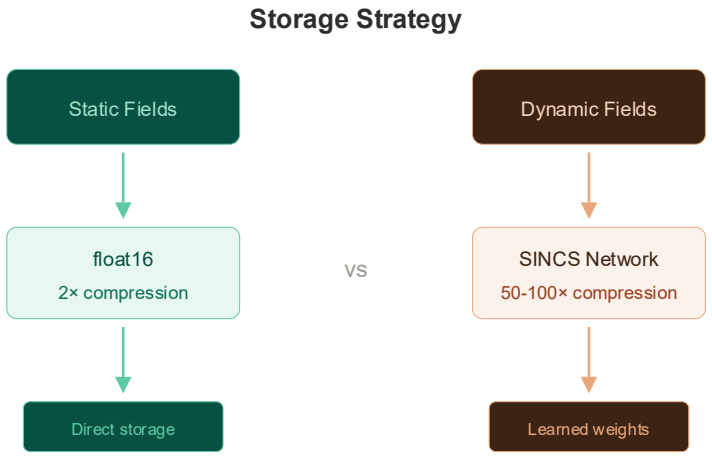
Field Classification

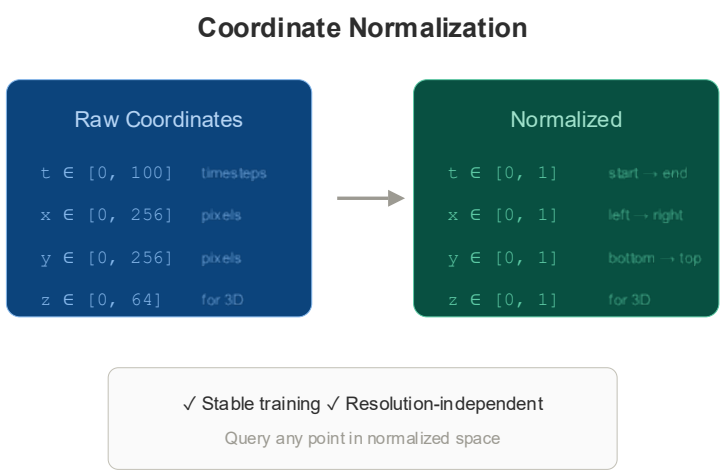


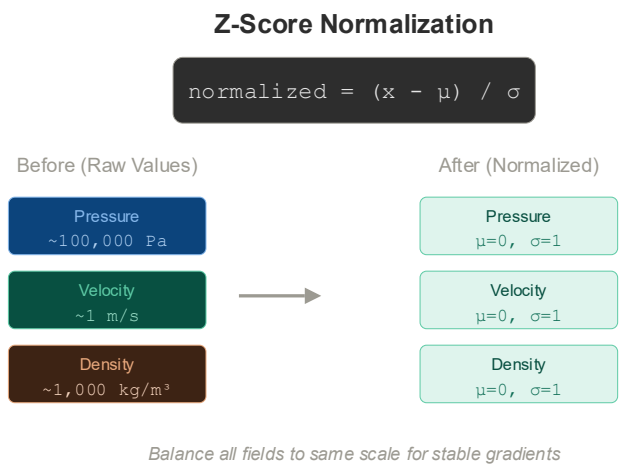
Separate what changes from what stays constant

Static/Dynamic Fields

To train an implicit neural network data has to be batched and normalized



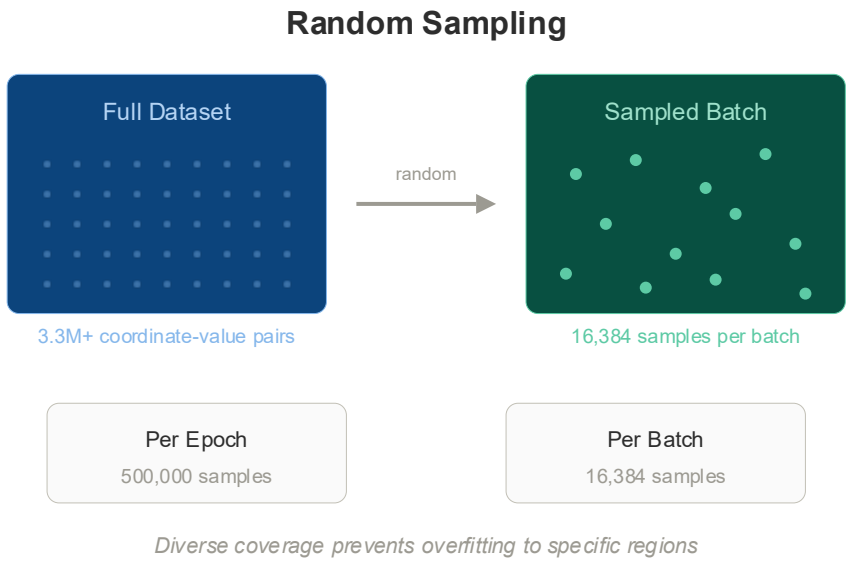


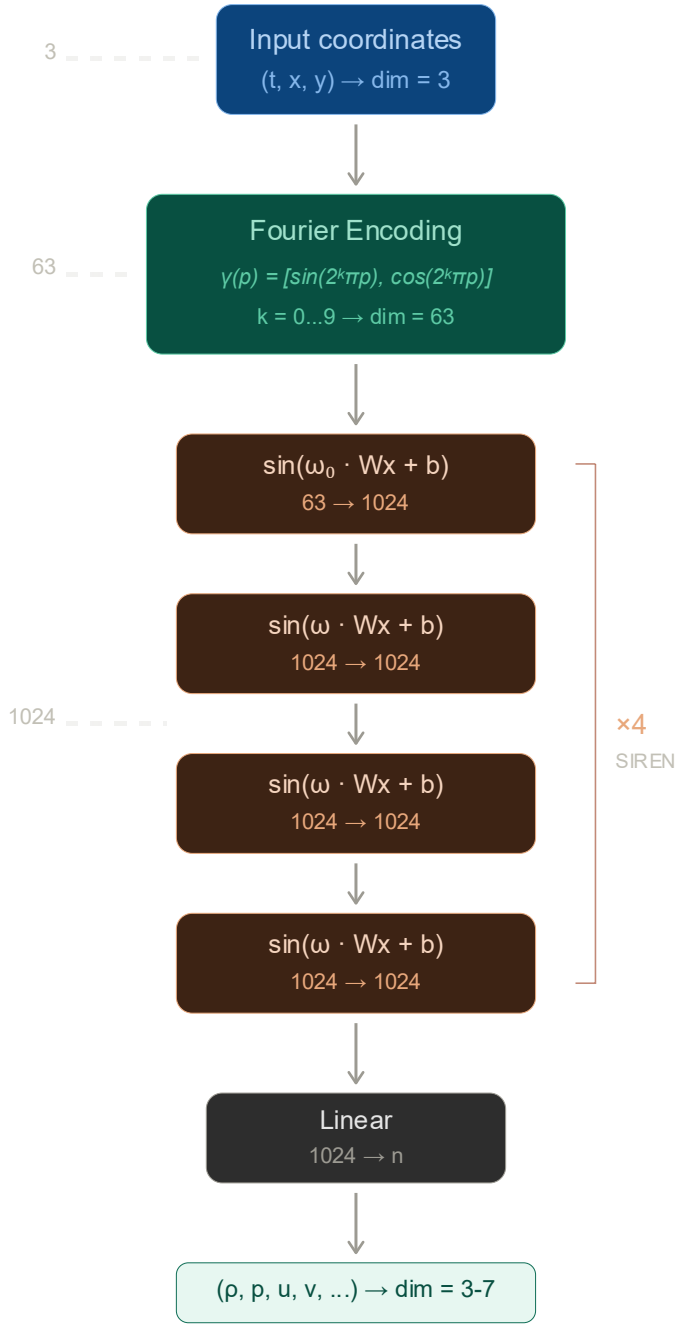


Static/Dynamic Data Store

Varied Time Scale and Metadata

Points are sampled and used to train a SINCS Network



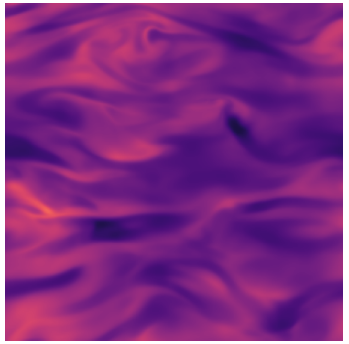


Total points in a 2D simulation: 50 timesteps × 256 × 256 = 3.3 million coordinate-value pairs

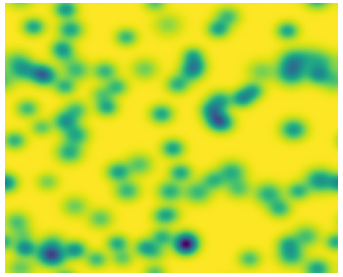
We trained 22 separate implicit neural representations, one per dataset, spanning 8 physics domains.

Domain	Class
Fluid Dynamics	Rayleigh Benard (3), Shear Flow, Active Matter, Viscoelastic Instability
Wave Phenomena	Acoustic Scattering (4)
Compressible Flows	Euler (2)
Magnetohydrodynamics	Magnetohydrodynamics
Turbulence	Turbulence Gravity, Turbulent Radiative (2)
Astrophysics	Supernova Explosion (2), Hydrodynamic Radiation, Plasma Physics
Geophysical	Shallow Water
Reaction- Diffusion	Gray Scott

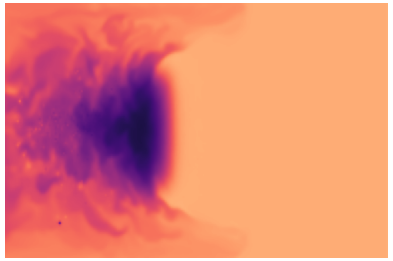
Fluid Dynamics



Magnetohydrodynamics



Reaction-Diffusion

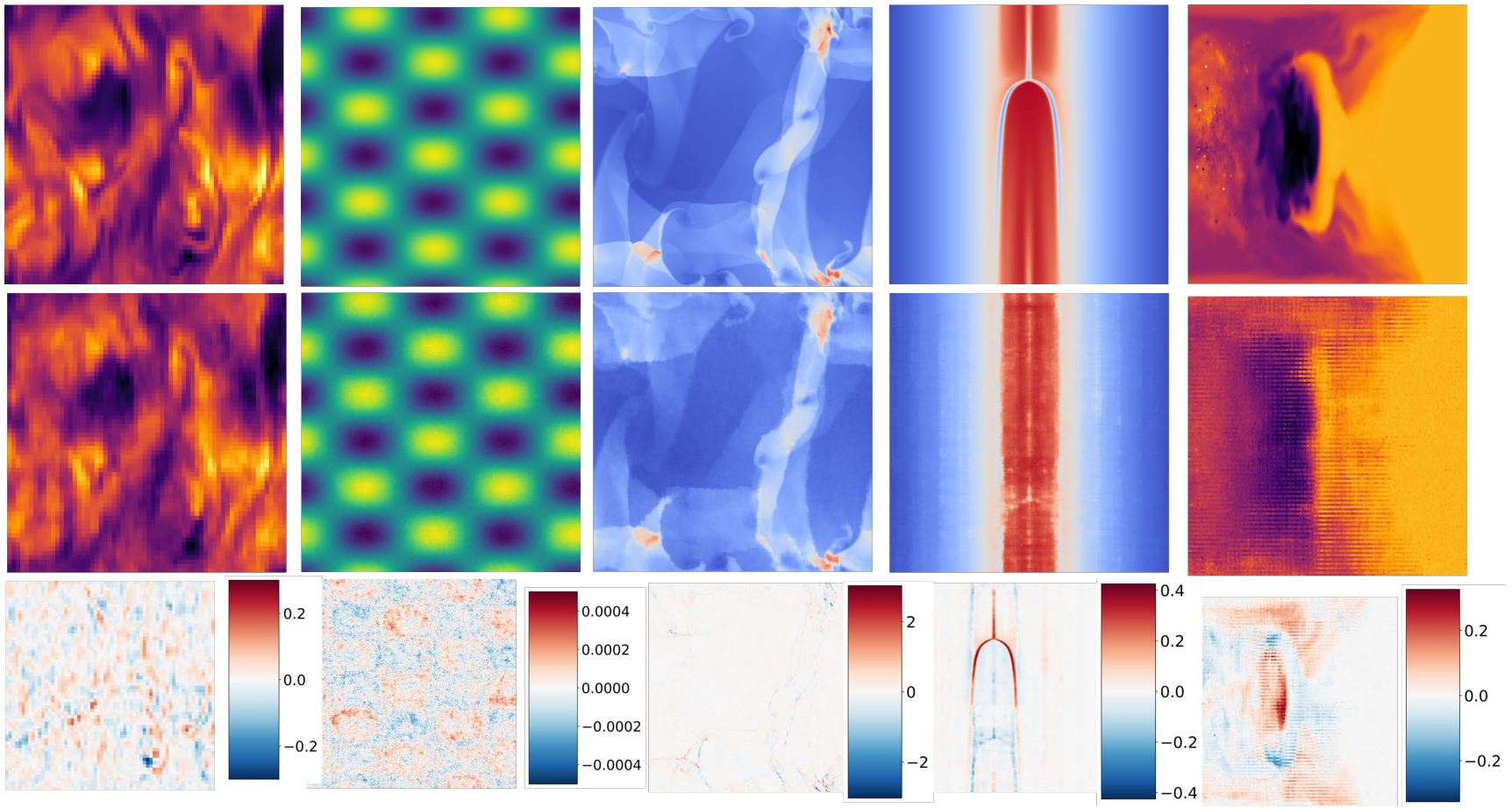


Astrophysics

Compression ratios range from 152× to 25,348× with all 22 datasets reduced to 37.6 MB each from initial sizes of 5.4 GB to 1.9 TB

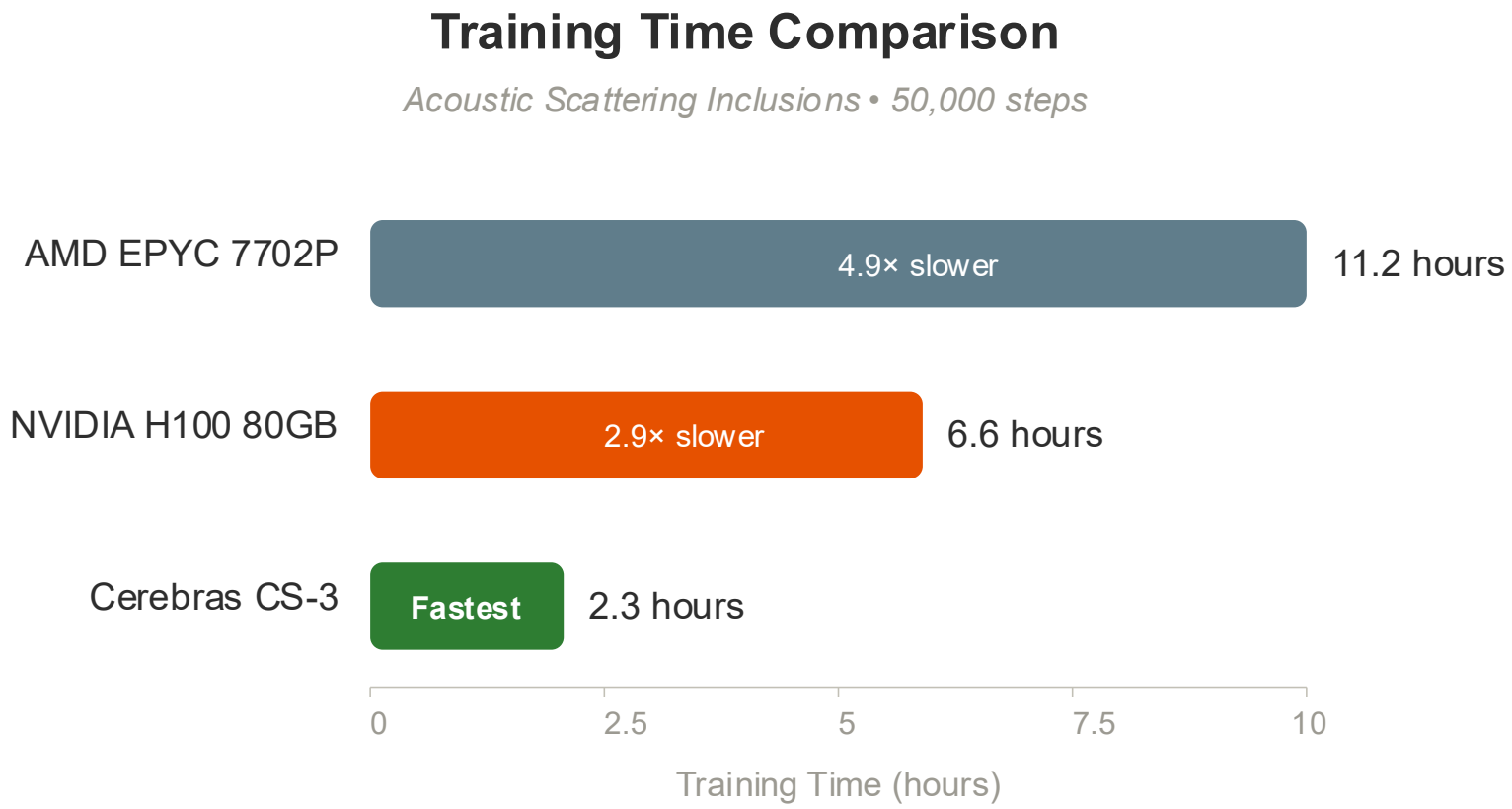
Domain	Compression Ratio	PSNR	Spectral Error
Fluid Dynamics	1483 ± 780	21.41 ± 4.30	0.49 ± 0.40
Wave Phenomena	1750 ± 679	21.49 ± 8.74	1.32 ± 0.24
Compressible Flows	25348 ± 8	14.71 ± 5.97	1.21 ± 0.03
Magnetohydrodynamics	220	14.90	0.73
Turbulence	2394 ± 1988	13.69 ± 1.75	1.09 ± 1.17
Astrophysics	1767 ± 1338	32.25 ± 9.27	0.65 ± 0.31
Geophysical	182	14.08	0.62
Reaction- Diffusion	152	23.86	0.29

Reconstruction quality varies by domain complexity

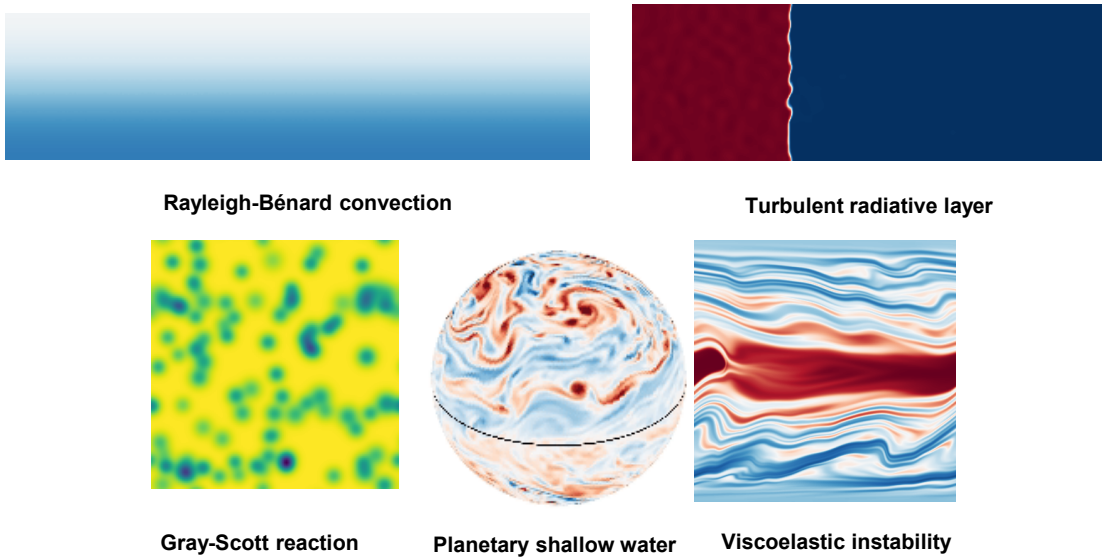


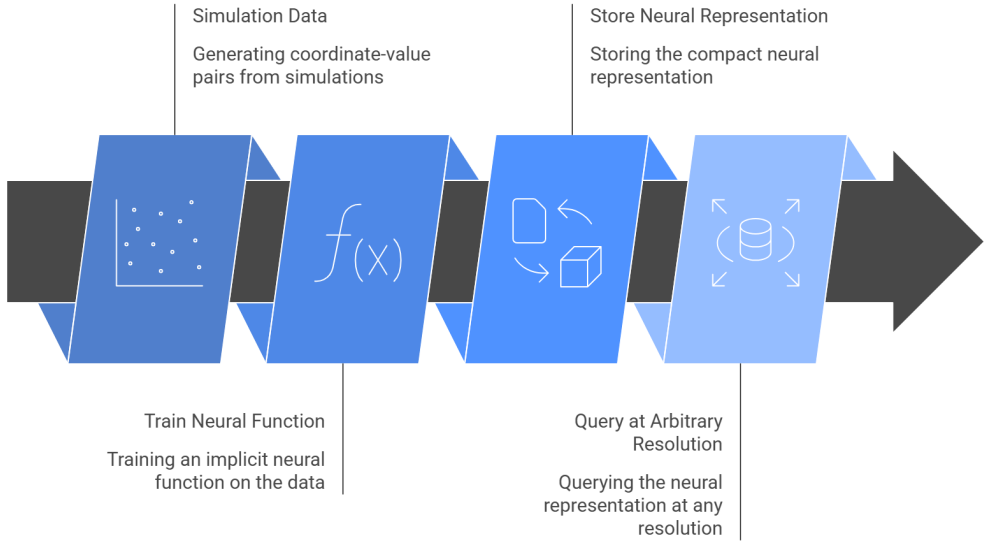
Magnetohydrodynamics Reaction-Diffusion Compressible Flows Fluid Dynamics Astrophysics

Training runs for 50,000 steps with batch size 16,384, completing in 2 to 3 hours per dataset compared to 6 to 11 hours on GPU and CPU respectively



Two paradigms have emerged to make this data tractable: learning over compact visual representations and compressing raw simulation data with neural networks.





Learning Over Reduced Representations

Implicit Neural Compression

[1] Ezemba, J., Afful, J., & Wang, M. Y. (2025). PhySiViT: A Physics Simulation Vision Transformer. Super Computing Conference 2025 (SC25)
[2]] Ezemba, J., Afful, J., & Wang, M. Y. (2025). Semantic-aware Implicit Neural Compression for Physics Simulations. Platform for Advanced Scientific Computing (PASC) Conference 2026 (Accepted)

Acknowledgements

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.

Physics-Aware AI at Scale: Neural Compression and Vision Transformers for Simulation Data

Jessica Ezemba

jezemba@andrew.cmu.edu

Jessicaezemba.com

Carnegie Mellon University