Task-Agnostic Self-Supervised Learning for Physical Dynamics

Biel Castaño Segade*
Sorbonne Université, CNRS, ISIR

Joceran Gouneau*
Sorbonne Université, CNRS, ISIR

Armand Kassaï Koupaï
Sorbonne Université, CNRS, ISIR

Lise Le Boudec
Sorbonne Université, CNRS, ISIR

Patrick Gallinari
Sorbonne Université, CNRS, ISIR
Criteo AI Lab

Abstract

Physics foundation models are commonly pretrained by forecasting future states, tying representation learning to a specific downstream task. We investigate whether self-supervised learning from partial observations can instead produce task-agnostic representations that capture the hidden dynamics of physical systems. Using a shared ViT backbone for heterogeneous 2D and 3D systems, we study masked reconstruction in physical space and JEPA-style prediction in latent space. To assess what these representations encode about the underlying dynamics, we freeze the encoder and measure how well governing physical parameters can be recovered from its features. To explicitly target properties of the dynamical system that persist across an entire trajectory, we introduce *regime fingerprinting*, which exploits the fact that an early partial view and a broader view of the same trajectory are generated under the same physical regime, and trains their representations to remain identifiable across time. Combined with latent-space prediction, regime fingerprinting yields the strongest recovery of governing parameters in our experiments and compares favorably with forecasting-pretrained baselines. These results suggest that self-supervised learning can capture trajectory-level physical dynamics without using forecasting as a pretraining objective.

1 Introduction

Foundation models for physics aim to learn representations that transfer across physical systems, parameter regimes, and operating conditions. Most existing approaches use forecasting as their primary pretraining objective: given past states, predict future ones [Subramanian et al., 2023, McCabe et al., 2024, Herde et al., 2024, Hao et al., 2024]. While natural, forecasting is not necessarily aligned with the broader goal of learning reusable representations of physical dynamics. In particular, learning to predict the next state is not the same as learning a representation that makes properties of the underlying dynamical system accessible. We therefore investigate *task-agnostic* pretraining: learning representations from physical trajectories without optimizing for a particular downstream task, with forecasting considered one possible application rather than the organizing pretraining objective.

*Equal contribution

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: Representations for the Physical Sciences.

Motivated by advances in self-supervised learning in vision and video [Assran et al., 2023, Bardes et al., 2024], we explore this direction using Joint Embedding Predictive Architectures (JEPA) [Assran et al., 2023]. JEPA-style methods learn by predicting representations of unobserved parts of the data rather than reconstructing the observations themselves. Our question is whether such latent-space self-supervision can learn representations informative about the physical system generating a trajectory, without using equation names, coefficients, or other physical labels during pretraining.

To assess what these representations capture about the underlying dynamics, we probe the physical *regime*, defined here as the properties held fixed along a trajectory, including the governing equation, coefficients, forcing terms, and boundary conditions [Ohana et al., 2024]. We use governing parameters as a measurable proxy for this information. After pretraining, we freeze the encoder and train a small attention-based probe to regress these parameters. This evaluates what information is accessible in the pretrained representation, rather than what can be learned through end-to-end finetuning.

Learning across heterogeneous physical systems raises two practical questions: how to encode trajectories with different spatial dimensionalities and sets of physical fields, and which self-supervised objective best captures their dynamics. We address the first with a ViT backbone designed to handle multiple physical systems in both 2D and 3D with heterogeneous variable fields. We address the second by comparing self-supervised objectives under a shared architecture and probing protocol, including masked reconstruction in physical space and masked latent prediction. Among these, masked latent prediction provides the base on which we build.

We then exploit a structural property of physical trajectories: while the observed states evolve over time, their governing regime remains fixed. This suggests that learning representations of physical dynamics should not only capture information about individual states but also retain properties that persist across a trajectory. We therefore introduce *regime fingerprinting*, an auxiliary self-supervised objective in which a sparse early view must identify a broader view of the same trajectory among other samples [van den Oord et al., 2018]. By encouraging agreement across views generated under the same regime, this objective complements masked latent prediction with a trajectory-level learning signal. Although correspondence between views does not uniquely isolate governing parameters, it encourages the representation to retain information about the dynamics that persists over time.

We evaluate our approach on nine governing parameters spanning heterogeneous physical systems. Using frozen-encoder parameter regression, we compare masked reconstruction, masked latent prediction, and the proposed regime fingerprinting objective. We also compare with forecasting-pretrained foundation models, including MPP [McCabe et al., 2024] and PDE-Transformer [Holzschuh et al., 2025]. Our experiments yield three main findings. First, masked latent prediction learns representations informative about governing parameters without forecasting-based pretraining. Second, complementing masked latent prediction with trajectory-level regime fingerprinting improves parameter recovery on seven of nine parameters, highlighting the benefit of exploiting persistent dynamical information. Third, our approach outperforms MPP on all nine parameters and PDE-Transformer on seven, although these comparisons involve different architectures. Together, these results suggest that exploiting persistent properties of physical trajectories is a promising direction for task-agnostic representation learning.

2 Method

First we pretrain a multiphysics encoder, then freeze it and measure how well a small probe recovers a trajectory's governing parameters from its features. Three task-agnostic objectives are compared on this protocol, all sharing the backbone of section 2.1 and differing only in what they ask of the representation: masked reconstruction (MAE), masked latent prediction (JEPA), and *regime fingerprinting*.

2.1 Backbone

Our ViT encoder covers the whole mixture. It is *multi-dimensional*: 2D inputs are lifted along a third spatial axis, so one patchifier serves both 2D and 3D fields. It is *multi-resolution*, cutting fixed-shape patches from whatever grid a system provides, and *multi-channel*, patchifying each field independently so their number may vary. It also accepts *arbitrary masks*: any subset of tokens may

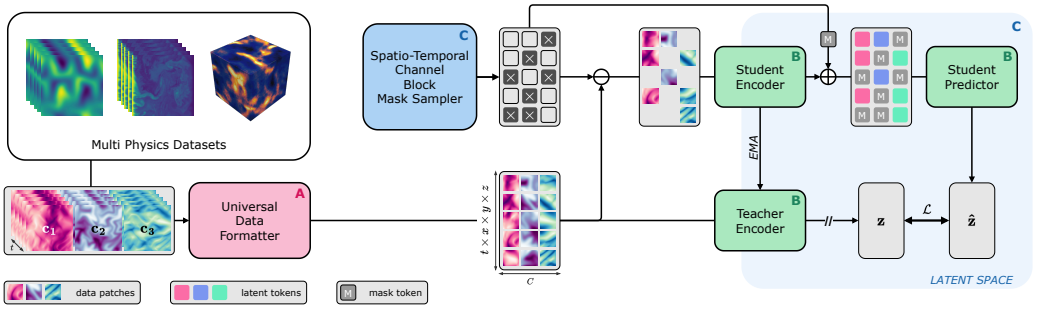

Figure 1: **Masked latent prediction on the shared backbone.** A universal data formatter (A) maps inputs with different spatial dimensions and fields to common tokens. The student (B) predicts the latents (C) an exponential-moving-average teacher produces from the unmasked input.

be hidden, not necessarily the same subset in every field, and since our objectives differ largely in what they hide, that is what lets them all run on one set of weights.

We apply axial attention [Ho et al., 2019]: one branch attends over space and time within each field, a second exchanges information across fields at each location, and a third uses learnable register tokens [Darcet et al., 2024] for global communication. The three branches run in parallel and their outputs are summed (fig. 2). The decoder uses the same axial Transformer, followed by a pointwise projection for latent targets or an unpatchifier for physical-space ones (Appendix section A).

2.2 Pretraining objectives

We investigate two self-supervised frameworks that learn from partial observations of a trajectory. *Masked autoencoding* (MAE) reconstructs missing information in physical space, whereas *Joint Embedding Predictive Architectures* (JEPA) predict missing information in latent space. For both, we explored several masking strategies and view constructions and report here the best-performing instantiation of each.

Masked reconstruction (MAE). In MAE [He et al., 2022, Zhou and Barati Farimani, 2024], the encoder receives a partially observed trajectory and the decoder reconstructs the masked patches in the original physical space. The prediction target is the physical field itself.

Masked latent prediction (JEPA). JEPA replaces physical-space reconstruction with prediction in representation space [Assran et al., 2023, Bardes et al., 2024]. We use the teacher–student formulation of Assran et al. [2023]. The student encoder E_s processes a masked view of the trajectory, while an exponential-moving-average teacher E_t processes the complete trajectory. A predictor g uses the visible student representations to predict the teacher representations at the masked positions. Let M denote the set of masked positions. The JEPA objective is

$$\mathcal{L}_\mathrm{JEPA} = \mathbb{E}_{x,M}\left[\frac{1}{|M|}\sum_{p\in M}\left\|g(E_\mathrm{s}(x_\mathrm{visible}), M)_p - \mathrm{sg}(E_\mathrm{t}(x)_p)\right\|_2^2\right],$$

where sg stops gradients through the teacher. The teacher is updated as an exponential moving average of the student, and only the student encoder is retained after pretraining.

Regime fingerprinting. Masked latent prediction scores tokens locally and does not explicitly encourage a trajectory-level representation of the physical properties. We therefore introduce *regime fingerprinting*, which exploits the fact that multiple views of the same trajectory are generated under the same physical regime ($r = (s, \mathbf{c}, \mathbf{f})$, comprising the system identity, coefficients, and forcing, independent of the initial condition). Trajectories governed by different physical regimes may appear similar at early times while diverging substantially later in their evolution; requiring an early partial view to identify the representation of the broader trajectory encourages the encoder to extract subtle, persistent regime-dependent cues.

Specifically, for each trajectory x, we draw four sparse student views and one broad teacher view. Each student view is an early temporal glimpse spanning 25% to 50% of the trajectory window while

retaining 25% of spatial positions and 50% of channels; the teacher view spans 75% to 100% of the window and retains all spatial positions and channels (section A). On each side, the trajectory representation is summarized by averaging the encoded register tokens: the student summary is passed through a two-layer MLP projection head h to produce $q = h(\bar{r}_{\mathrm{s}})$, while the teacher summary $s = \mathrm{sg}(\bar{r}_{\mathrm{t}})$ is taken directly from the EMA teacher. Both student and candidate teacher summaries are centered over the batch candidates and ℓ_2-normalized before comparison. They are aligned with an InfoNCE contrastive objective [van den Oord et al., 2018]:

$$\mathcal{L}_{\mathrm{regime}} = -\log \frac{\exp(\mathrm{sim}(q, s)/\tau)}{\sum_c \exp(\mathrm{sim}(q, s_c)/\tau)},$$

where $\mathrm{sim}(q, s) = q^\top s$ is the centered cosine similarity, $\tau = 0.1$ is a temperature parameter, and s_c ranges over the teacher summaries of all trajectories in the all-gathered batch. Crucially, each training batch is drawn from a single physical dataset, so all negative candidates s_c come from the same system with different governing parameters or initial conditions; recognizing which PDE system is being observed cannot by itself solve the matching task.

Regime fingerprinting is used only during pretraining, and the projection head h is discarded afterward. The two losses are optimized together as

$$\mathcal{L}_{\mathrm{JEPA}} + \lambda\, \mathcal{L}_{\mathrm{regime}},$$

with $\lambda = 0.03$ (section A).

2.3 Parameter Probing

We evaluate how much information about the physical regime is accessible in the learned representation by freezing the pretrained encoder and regressing the governing parameters. An attention-based pooling head [Assran et al., 2025] followed by a small MLP is trained from scratch for each dataset. All pretraining objectives use the same probe, so differences in performance reflect differences in the information made accessible by pretraining rather than downstream adaptation (section B).

3 Experiments

For pretraining we use a mixture of nine datasets from The Well [Ohana et al., 2024] (section B), spanning 2D and 3D systems with distinct governing equations, channel structures, and resolutions. To evaluate what these representations encode about the underlying dynamics, we freeze the encoder and probe each system's parameters. Because the probe has to be trained from scratch for each dataset, we limit this experiment to four representative systems spanning both 2D and 3D geometries, chosen for their diverse, well-defined continuous governing parameters across fluid, active, astrophysical, and magnetohydrodynamic regimes. Specifically, we regress nine governing parameters on these systems: Prandtl and Rayleigh for Rayleigh–Bénard, α and ζ for Active Matter, T_0, z and ρ_0 for Turbulence Gravity Cooling, and the Mach numbers M_a and M_s for MHD 64, scored by normalized RMSE (NRMSE).

3.1 Objective Study

We ran several self-supervised configurations, and report in table 1 the three that carry the comparison: MAE and JEPA, the base objectives of section 2, and JEPA with regime fingerprinting. Each base objective is reported at its best masking geometry: **channel tube** hides the same spatiotemporal patch locations in every field, forming a tube along the channel axis, and suits MAE; **multiblock** hides contiguous blocks across space and time, and suits JEPA. The fingerprinted run uses sparse early student windows against a broad teacher view (section A). All runs share the 88M-parameter encoder, pretrained for 30,000 steps and probed for 10,000. The two base objectives are close: JEPA beats MAE on 5 of 9 parameters. The trajectory-level demand shifts that tie: JEPA + regime beats them on 7 of 9. It loses to both on T_0 and ρ_0.

3.2 Comparison Against Forecasting Pretraining

We ask whether task-agnostic pretraining can expose the physical regime as effectively as models trained explicitly for forecasting. Compared with MPP [McCabe et al., 2024] and PDE-Transformer

Table 1: **Frozen-probe parameter regression** (NRMSE, lower is better; best per column within each block in bold). The blocks are separate experiments and are not comparable across the rule; section B gives the settings.

	Rayleigh-Bénard		Active Matter		Turb. Grav. Cool.			MHD 64	
	Prandtl	**Rayleigh**	α	ζ	T_0	z	ρ_0	**Ma**	**Ms**
Objective study (30,000 steps, three probe seeds)									
MAE (channel tube)	5.72E-2	1.89E-2	2.54E-2	1.37E-1	**7.75E-1**	1.49E-1	4.06E-1	3.81E-3	3.98E-2
JEPA (multiblock)	5.64E-2	1.58E-2	2.31E-2	8.31E-2	8.03E-1	1.50E-1	**3.93E-1**	5.51E-3	4.39E-2
JEPA + regime (λ=0.03)	**4.89E-2**	**1.46E-2**	**1.88E-2**	**7.79E-2**	8.77E-1	**1.47E-1**	4.16E-1	**3.34E-3**	**1.83E-2**
Against forecasting pretraining (100,000 steps, single runs)									
MPP	1.1902E-1	3.7156E-2	8.5470E-2	1.4263E-1	9.7024E-1	2.9358E-1	5.3148E-1	7.2277E-3	1.3045E-1
PDE-Transformer	**6.6177E-2**	3.1308E-2	9.4166E-2	1.6570E-1	9.5730E-1	3.2027E-1	5.3580E-1	**3.0570E-3**	1.0957E-1
Ours (JEPA + regime)	1.0356E-1	**2.5875E-2**	**7.6068E-2**	**1.0177E-1**	**8.9421E-1**	**2.3594E-1**	**4.4148E-1**	3.2672E-3	**5.2498E-2**

[Holzschuh et al., 2025], our 100,000-step encoder achieves lower parameter-regression error on all nine reported parameters for MPP and seven of nine for PDE-Transformer table 1. Despite architectural difference, we interpret these results as directional evidence that forecasting is not required to learn representations informative about the underlying dynamics.

4 Discussion and limitations

Our results support a task-agnostic approach to representation learning for physical dynamics. Self-supervised learning from partial observations produces representations that retain information about the underlying dynamical system, as revealed by frozen governing-parameter probes. To explicitly capture properties that persist across an entire trajectory, regime fingerprinting exploits correspondence between early and broader views generated under the same physical regime. Combined with latent-space prediction, it yields the strongest parameter recovery in our experiments.

Our evaluation is limited to governing-parameter recovery, which probes only one aspect of the information encoded in the representation, and the forecasting baselines use different architectures. An important next step is to determine whether these representations transfer to a broader range of physical tasks, including forecasting, adaptation, and other forms of system identification.

Acknowledgements

We acknowledge the financial support provided by DEEPNUM (ANR-21-CE23-0017-02), PHLUSIM (ANR-23-CE23-0025-02), MELISSA (ANR-24-CE23-7140-01), PEPR Sharp (ANR-23-PEIA0008, ANR, FRANCE 2030), and cluster IA PostGenAI@Paris (ANR-23-IACL-0007).

This project was provided with computer and storage resources by GENCI at IDRIS thanks to the grants AD011017018, 2025-A0191016890 on the supercomputer Jean Zay's V100/A100/H100 partitions.

We also thank the creators and maintainers of The Well benchmark for providing the diverse physics simulation datasets used in this work.

References

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15619–15629, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning, June 2025. URL `http://arxiv.org/abs/2506.09985`.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. URL `https://openreview.net/forum?id=QaCCuDfBk2`.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024. URL `https://openreview.net/forum?id=2dnO3LLiJ1`.

Zhongkai Hao, Chang Su, Songming Liu, Julius Berner, Chengyang Ying, Hang Su, Anima Anandkumar, Jian Song, and Jun Zhu. DPOT: Auto-regressive denoising operator transformer for large-scale PDE pre-training. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 17616–17635. PMLR, 2024.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022.

Maximilian Herde, Bogdan Raonić, Tobias Rohner, Roger Käppeli, Roberto Molinaro, Emmanuel de Bézenac, and Siddhartha Mishra. Poseidon: Efficient foundation models for PDEs. In *Advances in Neural Information Processing Systems*, volume 37, pages 72525–72624. Curran Associates, Inc., 2024. doi: 10.52202/079017-2311.

Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers, 2019. URL `https://arxiv.org/abs/1912.12180`.

Benjamin Holzschuh, Qiang Liu, Georg Kohl, and Nils Thuerey. PDE-transformer: Efficient and versatile transformers for physics simulations. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 23562–23602. PMLR, 2025. URL `https://proceedings.mlr.press/v267/holzschuh25a.html`.

Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, Daniel Haziza, Luca Wehrstedt, Jeremy Reizenstein, and Grigory Sizov. xformers: A modular and hackable transformer modelling library. `https://github.com/facebookresearch/xformers`, 2022.

Michael McCabe, Bruno Régaldo-Saint Blancard, Liam Parker, Ruben Ohana, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Siavash Golkar, Geraud Krawezik, Francois Lanusse, Mariel Pettee, Tiberiu Tesileanu, Kyunghyun Cho, and Shirley Ho. Multiple physics pretraining for spatiotemporal surrogate models. In *Advances in Neural Information Processing Systems*, volume 37, pages 119301–119335. Curran Associates, Inc., 2024. doi: 10.52202/079017-3791.

Michael McCabe, Payel Mukhopadhyay, Tanya Marwah, Bruno Régaldo-Saint Blancard, François Rozet, Cristiana Diaconu, Lucas Meyer, Kaze W. K. Wong, Hadi Sotoudeh, Alberto Bietti, Irina Espejo, Rio Fear, Siavash Golkar, Tom Hehir, Keiya Hirashima, Geraud Krawezik, Francois Lanusse, Rudy Morel, Ruben Ohana, Liam Parker, Mariel Pettee, Jeff Shen, Kyunghyun Cho, Miles Cranmer, and Shirley Ho. Walrus: A cross-domain foundation model for continuum dynamics. In *Proceedings of the 43rd International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2026. URL `https://arxiv.org/abs/2511.15684`.

Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina J. Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Keaton Burns, 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. The well: A large-scale collection of diverse physics simulations for machine learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 44989–45037. Curran Associates, Inc., 2024. doi: 10.52202/079017-1430. Datasets and Benchmarks Track.

M. S. Rautela, A. Most, S. Mansingh, B. C. Love, A. Scheinker, D. Oyen, N. Debardeleben, E. Lawrence, and A. Biswas. MORPH: PDE foundation models with arbitrary data modality, 2025. URL https://arxiv.org/abs/2509.21670.

Shashank Subramanian, Peter Harrington, Kurt Keutzer, Wahid Bhimji, Dmitriy Morozov, Michael Mahoney, and Amir Gholami. Towards foundation models for scientific machine learning: Characterizing scaling and transfer behavior. In *Advances in Neural Information Processing Systems*, volume 36, pages 71242–71262. Curran Associates, Inc., 2023. doi: 10.52202/075280-3119.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018. URL https://arxiv.org/abs/1807.03748.

Anthony Zhou and Amir Barati Farimani. Masked autoencoders are PDE learners. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=rZNuiFwXVs.

A Architecture details

Patchification. The repeat patchifier implements the shared-projection design in the Method. Inputs contain $T{=}10$ frames. For 3D inputs, it extracts patches of shape $(2, 16, 16, 16)$. For 2D fields, it repeats the singleton spatial axis to the patch depth before extraction. Thus a patch of spatial side k presents k^3 values to the input projection in both cases. Walrus and MORPH [McCabe et al., 2026, Rautela et al., 2025] retain the singleton third axis for 2D fields, yielding k^2 2D patches and k^3 3D patches that require dimensionality-specific projections. Patchification is applied independently to each physical channel. Each flattened patch is mapped to the embedding dimension by a two-layer GELU MLP.

Positional encoding and registers. Axial positional encodings over temporal, spatial, and channel axes are normalized, scaled by 50 with temperature 1024, summed, and added to patch embeddings. Learnable register tokens [Darcet et al., 2024] are prepended to each sequence; fingerprinting summaries are their means.

Backbone, predictor, and heads. The encoder is a pre-norm Transformer with 16 layers, width 512, and 16 attention heads. It uses RMSNorm, SwiGLU MLPs with expansion ratio 4, and no dropout, for 88.3M parameters. The masked-latent predictor inserts a learned mask token at each masked position and attends over the full grid with the same Transformer configuration. The student encoder and predictor together contain 177M trained parameters; the predictor is discarded after pretraining. The fingerprinting head is a two-layer MLP with hidden width 384 and is also discarded after pretraining. The parameter probe pools the frozen encoder's concatenated patch and register tokens with one learnable query.

Masking geometries. Four observation patterns appear in this paper, all expressed on the same encoder and weights. *Channel tube* hides the same spatiotemporal patch locations in every physical field, so a masked location is masked in all fields at once. *Multiblock* hides contiguous spatiotemporal blocks, with analogous blocks masked across fields. *Downsampled view* replaces the student input with an average-pooled and upsampled copy of the window, so the student sees the full extent at reduced resolution while the teacher sees the original field. *Sparse view* keeps a random subset of positions and of channels inside a restricted time span; this is what the fingerprinting student views use. The structured multiblock sampler removes contiguous blocks until it reaches ratio 0.85. Block scales are drawn uniformly from 0.40 ± 0.05 with aspect-ratio jitter 0.10. Masks preserve continuity across axes: nearby spatial regions remain masked over several steps with small displacements, analogous blocks are masked across channels, and spatial blocks stay contiguous within each frame. This creates coherent partial observations that require long-range and cross-channel context.

Attention over irregular token sets. Figure 2 shows one block. Block masking yields irregular visible sets: locations retain different variables, and variables retain different numbers of tokens. We flatten tokens and partition their indices into attention groups $\mathcal{C}$, one partition per axial stage. Within a stage,

$$A_{ij} = \begin{cases} 0, & \exists\, C \in \mathcal{C} \text{ such that } i, j \in C, \\ -\infty, & \text{otherwise.} \end{cases}$$

Same-variable groups implement spatio-temporal attention, and same-location groups implement channel attention. For C fields of N spatiotemporal tokens each, joint attention over all field-location tokens would cost $O((CN)^2)$, whereas the factorized paths cost $O(CN^2 + NC^2)$; with $N \gg C$, as in every system here, the first term dominates and cost is close to linear in the number of fields. A memory-efficient block-diagonal kernel handles these paths [Lefaudeux et al., 2022]; FlashAttention is used only for dense, unmasked paths because it cannot accept such masks. Fingerprinting uses the same packing, with one group per view, so tokens cannot attend across views or trajectories. This supports ragged multi-view training without padding.

Regime fingerprinting details. Four student views are drawn per trajectory. Each student view is an early temporal glimpse spanning 25% to 50% of the window, retaining 25% of spatial positions and 50% of channels; the teacher view spans 75% to 100% of the window and retains all positions and channels. Each view is summarized by the mean of its register tokens. In the loss of section 2, sim is cosine similarity at temperature 0.1, and student and candidate summaries are centered over the

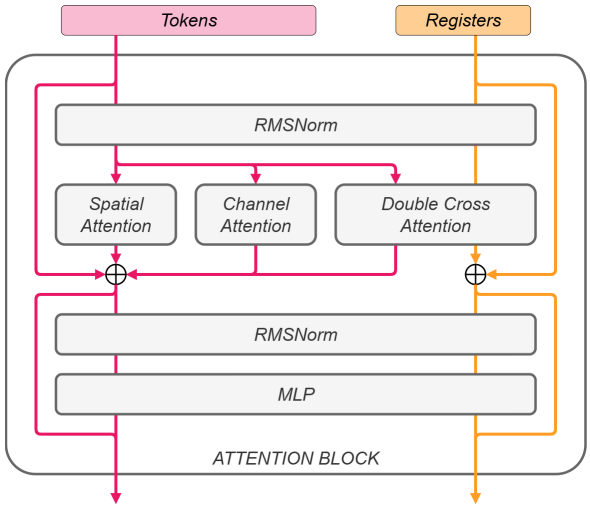

Figure 2: **One transformer block of the backbone.** Token states are updated in parallel by spatiotemporal attention within each physical field, by channel attention across fields at each location, and by a cross-attention exchange with the register tokens (drawn as global tokens), which restores the communication that factorization removes. The three paths are summed, and the same parameterization serves any number of fields.

candidates and ℓ_2 normalized before the comparison. The candidate set presented to a student view contains its own trajectory's teacher summary together with the summaries of the other trajectories in the same all-gathered batch. Because a batch is drawn from a single dataset, every candidate comes from that same system, so recognizing which system is being observed cannot by itself solve the task. The weight is swept over $\lambda \in \{0.01, 0.03, 0.1, 0.3, 1\}$ and λ=0.03 is retained. In the final 100,000-step configuration it is additionally annealed to zero between steps 70,000 and 95,000.

B Training configurations

Pretraining mixture. The nine datasets from The Well [Ohana et al., 2024] are Acoustic Scattering (Maze), Active Matter, Euler Multi-quadrants (periodic boundary conditions), Gray-Scott Reaction Diffusion, Rayleigh-Bénard, Turbulent Radiative Layer 2D, MHD 64, Supernova Explosion 64, and Turbulence Gravity Cooling. Probing uses four of them: Prandtl and Rayleigh numbers for Rayleigh-Bénard, α and ζ for Active Matter, T_0, z and ρ_0 for Turbulence Gravity Cooling, and the Mach numbers M_a and M_s for MHD 64. The low-variance subset used in section 3 is Prandtl, Rayleigh, ζ, T_0, z and ρ_0; α, M_a and M_s vary substantially across probe seeds.

Table 2 lists the two pretraining configurations: the *study configuration* underlying the objective study and the *final configuration* underlying the comparison against the external pretrained baselines. Both use AdamW with a cosine schedule, bfloat16 mixed precision, and gradient checkpointing. The baselines are pretrained for the same 100,000 steps with 32 examples per update and their native objectives and recommended optimizer settings.

Probing protocol. For each dataset, an attentive-pooling extractor (11M) [Assran et al., 2023] and a small MLP regressor are trained from scratch on frozen encoded features. Parameter targets are transformed with $\log_{10}$ and then z-scored; the provided dataset splits partition complete trajectories, so no evaluated trajectory appears in probe training. Probes are trained for 10,000 steps with 16 examples per update in the study configuration and for 20,000 steps with 16 examples per update in the final comparison. Default configurations are averaged over three probe seeds. Baseline cells are single runs, as stated per table.

Table 2: The two pretraining configurations.

Hyperparameter	Study configuration	Final configuration
Optimization steps	30,000	100,000
Global examples per update	16	32
Optimizer	AdamW	AdamW
Learning rate	2×10^{-4}	2×10^{-4}
Schedule	cosine	cosine
Warmup steps	8,000	8,000
Weight decay	10^{-4}	10^{-4}
Precision	bfloat16	bfloat16
Masking ratio	0.85	0.85
Teacher EMA decay	0.99	0.99
Fingerprinting weight λ	0.03	0.03, decayed to 0
Fingerprinting warmup steps	10,000	10,000
Fingerprinting decay window	none	70,000 to 95,000
Gradient clipping (global norm)	1.0	1.0
Learning-rate floor	10^{-7}	10^{-7}