Physics-Aware Representation Learning for Physical Systems

Charles Cheng Ji [1] **Zhanhe Shi** [1] **Richard Wang** [1] **Romina Yalovetzky** [1]

Abstract

Learning representations aligned with physical dynamics remains a central challenge in scientific machine learning. We study this problem within the JEPA framework on the active matter dataset from the Well benchmark, and identify preservation of temporal structure during encoding as the key factor driving downstream physics-prediction quality.

Building on this insight, we propose two complementary approaches. H-JEPA decomposes spatial and temporal modeling through its hierarchical structure, reducing total test MSE on the regressed parameters (α, ζ) by up to 70% relative to the baseline. FA-JEPA introduces field-aware attention together with a physics-prior attention mechanism that injects domain knowledge directly into the attention process, lowering total test MSE by roughly 20% over the prior-free variant on both linear and kNN probes.

1. Introduction

Many problems in scientific domains remain fundamentally challenging due to the complexity of the underlying physical systems and the high-dimensional nature of observational data. A common approach in this setting is to learn frame-by-frame, pixel-level emulators of numerical simulations using autoregressive surrogate modeling (McCabe et al., 2023; Herde et al., 2024; McCabe et al., 2025). However, these full-field prediction models are often not only computationally expensive to train, but they are also not necessarily well aligned with downstream scientific objectives. In contrast to generative video tasks, where the goal is to produce visually accurate predictions, scientific applications typically require the estimation of underlying system parameters or physically meaningful quantities governing the dynamics.

To tackle this problem, while supervised learning provides a direct way to learn meaningful representations, it requires large amounts of labeled data. Reinforcement learning, though effective, often demands an impractically large number of trials. Self-supervised learning mitigates these challenges; however, a gap remains between physical numerical simulations and modern self-supervised methods in capturing the essential physical structure of the system.

We study this in the context of active matter, a system with complex, multiscale, and highly nonlinear dynamics driven by the collective behavior of self-propelled particles (Maddu et al., 2024). The simulations are governed by two parameters, α and ζ, and produce observable fields—concentration, velocity, and nematic orientation—that are strongly coupled and evolve through interactions not directly visible at the pixel level. As a result, even self-supervised methods struggle to capture the underlying physics, motivating architectures that incorporate domain-specific structure.

For this, we introduce two approaches that build on Joint-Embedding Predictive Architectures (JEPA) (LeCun et al., 2022). Our main contributions are two-fold:

(1) We show that capturing physical evolution requires modeling system dynamics directly, rather than framing the task as next-frame prediction. Building on (Qu et al., 2026), we introduce physics-compatible preprocessing, a ViT encoder, and a expressivity constrained 3D-CNN predictor, but still observe limitations in capturing the underlying physics.

(2) We develop novel techniques that aim at incorporating physics of the system into the architecture: a hierarchical approach, dubbed H-JEPA, that divides the task into leaning the spatial representation and the time evolution independently and Field-Aware JEPA, dubbed FA-JEPA, to augment the model architecture with a physics-aware attention mechanism that aims at capturing the physical relations between the observables.

2. Related work

Joint-Embedding Predictive Architectures (JEPA) (LeCun et al., 2022) have emerged as a promising framework for learning latent world models through predictive representation learning. A growing body of work has explored JEPA-style approaches across vision, video, and dynamical systems (Assran et al., 2025; Zhou et al., 2024; Goswami

[1]New York University. Correspondence to: All authors <{cj2220, zs3413, rw3931, ry2665}@nyu.edu>.

Preprint. May 3, 2026.

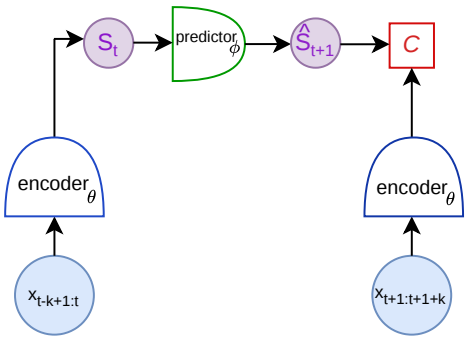


Figure 1. JEPA architecture for latent dynamical modeling. The encoder maps past observations $x_{t-k+1:t}$ to a representation S_t, which is transformed by a predictor into a future embedding $\hat{S}_{t+1}$. A target encoder processes future frames $x_{t+1:t+1+k}$ to produce a target representation C, and learning is driven by aligning predicted and target embeddings.

et al., 2025; Nam et al., 2026; Maes et al., 2026). Recent work has therefore explored simpler and more stable training objectives for JEPA-based models, including losses that encourage well-structured latent distributions, such as Gaussian-aligned embeddings (Maes et al., 2026). These developments highlight the trade-off between representation stability, expressivity, and scalability in predictive latent modeling. Recent work has further incorporated physical priors into these frameworks, including integrating physics into encoders (Bardhan et al., 2025), introducing physics-inspired loss functions (Yee & Koh, 2026). Despite these advances, systems such as active matter (Maddu et al., 2024), characterized by multiscale, nonlinear, and strongly coupled dynamics, remain challenging, as their underlying physical mechanisms are not readily captured by generic representation learning methods, highlighting the need for architectures with domain-specific inductive biases.

3. JEPA to capture dynamics

Our goal is to design and train a representation learning model to capture the temporal evolution of physical systems with self-supervised learning. As done in the recent work of (Qu et al., 2026), we leverage the JEPA framework, whose general architecture is shown in Fig. 1. Our goal is to propose novel techniques to leverage JEPA to capture the underlying physical dynamics.

Input. At each time step t, it is considered a context window of k pixel frames, $x_{t-k+1:t} = (x_{t-k+1}, \ldots, x_t)$, in the observation space. We consider different channels that can, in principle, represent different attributes of the problem as physical observables (e.g., velocity in a direction). This context is encoded onto the latent representation space to produce the state S_t that is used to predict a target state $\hat{S}_{t+1}$ that encodes the pixel frames $x_{t+1:t+1+k}$. We consider k consecutive time frames both in the context and the target.

The architecture consists in both an encoder and predictor. The encoder is used to encoder the context and the target states onto the latent representation space, while the predictor evolves in time in that space.

$$\begin{aligned} \text{Encoder:} \quad & s_t = E_\theta(x_{t-k+1:t}), \\ & s_{t+1} = E_\theta(x_{t+1:t+1+k}), \\ \text{Predictor:} \quad & \hat{s}_{t+1} = P_\phi(s_t). \end{aligned} \tag{1}$$

In this work we explore different variants of this architecture. Our goal is to train a *robust encoder* that is capable of producing a representation of the input in the latent representation that is evolved in time by the predictor to produce a state $\hat{S}_{t+1}$ that is close to the representation corresponding to the target k frames. Given a training dataset $\mathcal{T}$ consisting of context and target of k consecutive pixel frames, we measure the closeness with the square L2 distance (MSE) as

$$\mathcal{L}_{\text{pred}} = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \| P_\phi(E_\theta(x_{t-k+1:t}))_i - E_\theta(x_{t+1:t+1+k})_i \|_2^2 \tag{2}$$

For the training loss, we consider adding a regularization term as $\mathcal{L}_{\text{pred}} + \lambda \mathcal{L}_{\text{reg}}$.

In the next sections we elaborate on the different variants of this architecture that we propose and benchmark. We consider different encoding and prediction mechanisms, different regularizations, specifically we consider both VICReg (Bardes et al., 2021) and SIGReg (Balestriero & LeCun, 2025).

4. Baseline JEPA

We build on the JEPA baseline of Qu et al. (2026) and study three design choices: preprocessing, regularization, and—most importantly—the encoder. The baseline uses a 3D ConvNeXt encoder over k=16 frames and a CNN predictor, trained with VICReg (Bardes et al., 2021). Temporal information is fused early via strided convolutions, reducing T=16 $\rightarrow$ 4, so much of the network operates on already pooled features.

FFT preprocessing. We replace spatial cropping with a band-limited 2D FFT resize, which preserves periodic boundary conditions and phase relationships across channels (see Appendix Fig. 9). The operation is exact under periodic boundary conditions and preserves the relative phase across the 11 channels—relevant because the orientation and strain-rate tensors are spatial gradients of the same underlying particle field.

Encoder family. We compare three encoders: (i) 3D ConvNeXt (Liu et al., 2022) (baseline), (ii) Conv+Attn, which adds transformer blocks on top of convolutional features, and (iii) ViT3D (Assran et al., 2025), which directly tokenizes spatiotemporal patches. This isolates whether performance depends on attention depth or on the tokenization of temporal structure (see Appendix B for details).

Regularizer and training. We compare VICReg against SIGReg (Balestriero & LeCun, 2025), which has been reported to scale better at large batch sizes

5. Approaches for physics-aware JEPA

To address the limitations identified in the baseline, we propose two complementary approaches that explicitly incorporate the structure of the underlying physical dynamics.

5.1. Field-Aware JEPA

5.1.1. Model architecture

The encoder and predictor. We leverage the concept of Vision Transformers (ViT) (Assran et al., 2025; Mur-Labadia et al., 2026) as both the encoder and the predictor. In the encoder, we adopt a ViViT-style architecture (Arnab et al., 2021) and augment it with an additional attention mechanism, which we term *field-aware attention*.

Both models use 32-frame samples split into 16-frame context and 16-frame target clips, tokenize the 11 physical channels with field-specific 3D tubelets of size $2 \times 16 \times 16$, and 8 block transformer with factorized attention and a 6-block ViT predictor. For more details on how the training was performed, refer to Appendix B.5.1.

Field-Aware attention mechanism. While ViViT factorizes attention across spatial and temporal dimensions to efficiently model video data, we generalize this idea to physical systems by introducing attention across an additional dimension corresponding to the physical fields. This allows the model to capture not only the evolutions across space and time, but also the interactions between different physical quantities. The proposed method is particularly suitable for complex domains in which there are multiple interconnections between the physical attributes of the problem, as is the case with active matter systems, which we tackle in this paper. Within a transformer block in the encoder, we use three attention mechanisms: (i) *field attention*, which mixes information across physical fields (velocity, orientation, strain, concentration) at a fixed space-time location; (ii) *spatial attention*, which operates across spatial patches within a field at a fixed time; and (iii) *temporal attention*, which operates across time at a fixed spatial location within a field.

Physics-prior attention. In addition to field-level attention, we introduce a *physics-prior attention*. Our motivation is to improve the estimates of the alignment parameter ζ. We note that ζ controls the tendency of rod-like particles to align, so the encoder should be encouraged to attend more strongly to pairs of spatiotemporal patches with similar nematic orientation structure. We introduce a physics-informed attention weight that encourages more attention to the most similarly aligned patches and reduces it for the most misaligned ones. This encourages the model to preferentially exchange information between patches that are both spatially related and similarly aligned, which is expected to improve sensitivity to the alignment parameter ζ.

Inspired by prior work on attention with inductive biases (Lin et al., 2022), we consider a final attention as a convex combination of the learned attention (through data) and a physics-informed attention map derived from the nematic order tensor. We extract the nematic order tensor components $\mathbf{q} = (Q_{xx}, Q_{xy}, Q_{yx}, Q_{yy})$ from the input channels, pools them to match the patch resolution, and normalizes each resulting 4-dimensional vector $\mathbf{q}_i$ to unit length $\tilde{\mathbf{q}}_i = \frac{\mathbf{q}_i}{\mathbf{q}_i}$. Pairwise similarities are then computed as the dot product $S_{ij} = \mathbf{q}_i \cdot \mathbf{q}_j \in [-1, 1]$ between all spatial positions i, j within each time step. We define a physics score $\tau S_{ij} \in [-\tau, \tau]$, such that the scale is controlled by τ, which is a bounded learnable parameter, preventing the prior attention from becoming arbitrarily sharp.

With this, the final attention is computed as a convex combination

$$A = (1 - \lambda)A^{\text{data}} + \lambda A^{\text{phys}}, \tag{3}$$

where the physics attention weight is $A^{\text{phys}} = \text{softargmax}(\tau S)$, with τ ensuring that the final attention remains a valid probability distribution and that the physics-prior attention logits are bounded, and A^{data} is the standard (data) attention. $\lambda \in [0, 1]$ is a learnable mixing coefficient initialized near zero. This guarantees that the final attention map remains a valid probability distribution while allowing the model to adaptively incorporate the physics prior. To ensure a well-defined and stable parametrization, we introduce unconstrained scalar parameters η and ρ, and define

$$\lambda = \sigma(\eta), \qquad \tau = \tau_{\max}\, \sigma(\rho), \tag{4}$$

where $\sigma(\cdot)$ denotes the sigmoid function. This enforces $\lambda \in [0, 1]$ and $\tau \in [0, \tau_{\max}]$. Both η and ρ are learned jointly with the model parameters via gradient-based optimization. We initialize λ to a small value, such that the model begins close to standard attention, and allows it to adaptively incorporate the physics prior during training.

Regularizer. We employ SIGReg (Balestriero & LeCun, 2025) with 17 integration knots and projected over 1024 different unit vectors, following what was done in LeWorldModel (Maes et al., 2026; Balestriero & LeCun, 2025). In

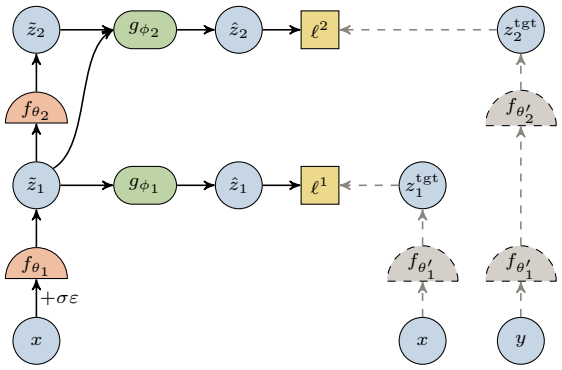


Figure 2. Hierarchical JEPA computational graph. **Left**: the online encoder (orange) processes the corrupted context $\tilde{x} = x + \sigma\varepsilon$ bottom-up into shallow / deep latents $\tilde{z}_1, \tilde{z}_2$. **Right**: EMA target encoders (gray dashed; shared parameters, stop-gradient) encode the clean context x and the paired next-window y into targets $z_1^{\text{tgt}}, z_2^{\text{tgt}}$. Predictors g_{ϕ_1}, g_{ϕ_2} (pills) produce $\hat{z}_1, \hat{z}_2$, paired against the targets through losses ℓ^1 (latent denoising) and ℓ^2 (next-window prediction); $\tilde{z}_1$ additionally skips to g_{ϕ_2} via a strided 4×4 projection. Solid / dashed arrows mark gradient / stop-gradient. For the outer-denoising ablation (S3) the right-branch input y is replaced by x.

order to use SIGReg as a collapse prevention regularizer, which relies on the model having a large enough batch size during training, we aim to match the batch size as closely to LeWorldModel as possible. In LeWorldModel, the batch size is 128. However, due to hardware constraints, we set the batch size to 16, which is a weaker setup for SIGReg. We mitigate this issue through activation checkpointing and gradient accumulation.

Summary. In contrast to other methods that impose stop-gradients across the encoder over the target (Bardes et al., 2024) and apply and MAE approach, our proposal follows the idea presented in LeWorldModel (Maes et al., 2026) of seeking more simplified and explainable architectures. The model has $\sim 36.5M$ parameters. For the details of the training refer to Appendix B.5.1.

5.2. Hierarchical JEPA

5.2.1. Model architecture

Inspired by LeCun et al. (2022), we propose Hierarchical JEPA (H-JEPA), a structural decomposition of the JEPA encoder of Section 3 into two pathways supervised by complementary self-supervised pretext tasks. The *inner pathway* (shallow encoder f_{θ_1} + inner predictor g_{ϕ_1}) retains instantaneous field structure through a latent denoising task; the *outer pathway* (deep encoder f_{θ_2} + outer predictor g_{ϕ_2}) captures temporal dynamics through paired next-window prediction. The two pretext tasks together pressure the deep latent to carry both the local field state and the trajectory-level dynamics relevant to (α, ζ).

Inputs. The context window $x = x_{t-k+1:t}$ and paired target window $y = x_{t+1:t+1+k}$ are the $k = 16$ contiguous frames of Section 3 at sliding stride 1, sharing the same trajectory. At the model boundary we apply Gaussian corruption $\tilde{x} = x + \sigma\varepsilon$ with $\sigma = 1$ and $\varepsilon \sim \mathcal{N}(0, I)$ to the context only, so the EMA target encoders always receive clean inputs. Frames are bilinearly resized to 224×224 with no augmentation.

Encoders. We split the JEPA encoder E_θ into a shallow encoder f_{θ_1} and a deep encoder f_{θ_2}, producing a shallow latent $z_1 = f_{\theta_1}(x_{t-k+1:t}) \in \mathbb{R}^{d_1 \times T_1 \times H_1 \times W_1}$ and a deep latent $z_2 = f_{\theta_2}(z_1) \in \mathbb{R}^{d_2 \times H_2 \times W_2}$; The deep encoder collapses time, so z_2 is a per-window spatial feature map. The EMA target encoders $f_{\theta'_1}, f_{\theta'_2}$ are EMA copies of the online ones with stop-gradient on their outputs (Grill et al., 2020; Chen & He, 2020), a self-distillation construction also used by V-JEPA (Bardes et al., 2024). Concretely, $f_{\theta'_1}, f_{\theta'_2}$ run on clean inputs with no gradient and are updated as $\theta'_k \leftarrow \tau\,\theta'_k + (1-\tau)\,\theta_k$ with $\tau = 0.996$.

Predictors and tasks. Each pathway is trained with its own pretext task, mediated by a dedicated predictor g_{ϕ_l}. The *inner predictor* g_{ϕ_1} recovers the EMA target's encoding of the clean context from a Gaussian-corrupted view of the same window, a latent-space denoising objective. Concretely, g_{ϕ_1} is a 3D residual ConvNet on the (T_1, H_1, W_1) grid: $\hat{z}_1 = g_{\phi_1}(\tilde{z}_1)$, $z_1^{\text{tgt}} = f_{\theta'_1}(x)$, where $\tilde{z}_1 = f_{\theta_1}(\tilde{x})$ is the encoded corrupted context $\tilde{x} = x + \sigma\varepsilon$. The *outer predictor* g_{ϕ_2} predicts the deep latent of the paired (next-window) target from the corrupted current-window pair — the standard JEPA next-step latent-prediction objective. Concretely, g_{ϕ_2} first reduces $\tilde{z}_1$ to the deep grid (temporal mean followed by a strided 4×4 Conv2d), concatenates the result with $\tilde{z}_2$, and runs a 2D residual stack: $\hat{z}_2 = g_{\phi_2}(\tilde{z}_1, \tilde{z}_2)$, $z_2^{\text{tgt}} = f_{\theta'_2}\big(f_{\theta'_1}(y)\big)$.

Learning objective and regularizer. The default H-JEPA aligns each predictor output with its EMA target through a VICReg pair loss ℓ_{V} (Bardes et al., 2021), with coefficients $(\text{sim}, \text{std}, \text{cov}) = (2, 40, 2)$ matching the baseline of (Qu et al., 2026):

$$\mathcal{L} = w_1\,\ell_{\text{V}}(\hat{z}_1, z_1^{\text{tgt}}) + w_2\,\ell_{\text{V}}(\hat{z}_2, z_2^{\text{tgt}}). \tag{5}$$

We additionally study an MSE + SIGReg variant of this objective— replacing ℓ_{V} with an MSE pair loss and adding an encoder-side SIGReg term (Balestriero & LeCun, 2025; Maes et al., 2026)—whose full formulation is given in Appendix C.3.

Summary. H-JEPA has ~ 4.0 M trainable parameters (enc_1 0.27 M, enc_2 2.20 M, inner 0.66 M, outer 0.84 M); the EMA target encoders add 2.47 M non-trainable parameters.

6. Benchmarks

Dataset. We use the active matter dataset from the Polymathic AI Well benchmark from the HuggingFace repository[1], which models systems of self-driven agents that convert chemical energy into mechanical work, giving rise to emergent collective dynamics (Maddu et al., 2024). The dataset consists of simulations of N rod-like particles immersed in a Stokes fluid and provides spatiotemporal fields such as concentration, velocity, and nematic orientation. Each trajectory is parameterized by underlying physical quantities, including the active dipole strength α and the alignment interaction strength ζ.

Our goal is to leverage the self-supervised JEPA framework to learn an encoder that produces latent representations capturing the physical dynamics of the system. Training is fully unsupervised, with α and ζ predicted only via post-hoc probes on the frozen encoder. The model learns representations whose temporal evolution—modeled by a predictor—matches future states, decoupling representation learning from downstream evaluation.

Figures of merit. We report z-score-normalized MSE for (α, ζ), both per parameter and averaged, which we refer as total. The encoder is used as a frozen feature extractor. We evaluate representations with two probes: a linear probe solved via closed-form least squares using `torch.linalg.lstsq`, and a kNN probe with inverse-distance weighting, sweeping k and distance metrics $\in \{\text{euclidean}, \text{cosine}\}$. Using the train/validation/test split, we select the configuration (checkpoint, k, metric) with lowest validation MSE and report the corresponding test performance. The same checkpoint is used for both α and ζ within each probe. See Appendix (Fig. 5, Fig. 6) for details.

Baseline and its extensions. Our main baseline methods is (Qu et al., 2026) and we developed the extensions mentioned in Section 4. With a fixed setup (refer to Appendix B), we evaluate architectural choices: SIGReg fails to converge, while Conv+Attn variants provide only marginal gains, likely due to the loss of temporal structure in convolutional tokenization. In contrast, a ViT3D encoder operating on spatiotemporal patches significantly improves performance (linear MSE 0.107, k-NN 0.120), highlighting the importance of the tokenizer in capturing system dynamics. Additional implementation details, ablations, and training curves are provided in Appendix B.

The empirical story collapses to one sentence: every change that helped acted on the tokenizer or its inputs. FFT preserves the periodic phase information the crop discarded; ViT3D injects temporal context *into* each token rather than after the strided convs have removed it. Adding attention depth without changing the tokens—the Conv+Attn family—did not move the baseline. We refer to this ViT3D-based variant as *Temporal* (denoted as *ViT3D-d6, FFT* in Appendix B).

[1] https://polymathic-ai.org/the_well/datasets/active_matter/

6.1. Results summary

We summarize the results and highlight H-JEPA, which achieves the best performance and outperforms prior JEPA-based approaches on this task. While JEPA provides a strong and flexible framework for self-supervised representation learning, H-JEPA improves upon it by better aligning the learned representations with the underlying physical dynamics. It reduces total test MSE on (α, ζ) by up to 70% relative to the baseline, significantly narrowing the gap to supervised methods without requiring large labeled datasets.

	Linear MSE ↓			KNN MSE ↓			
Method	α	ζ	Total	α	ζ	Total	k
Baseline	0.059	0.460	0.260	0.081	0.570	0.325	5
Temporal	0.016	0.197	0.107	**0.009**	0.231	0.120	20
H-JEPA	**0.014**	**0.173**	**0.093**	**0.009**	**0.189**	**0.099**	20
FA-JEPA	0.033	0.291	0.162	0.033	0.550	0.292	4
Supervised	**0.027**	**0.077**	**0.052**	–	–	–	–

Table 1. z-score MSE on the test dataset per configuration and probe. Bold marks the best results obtained. The baseline corresponds to executing (Qu et al., 2026). H-JEPA and FA-JEPA correspond to the best variant of the these methods (discussed in Sec.6.2 and 6.3). While the top of the table reports self-supervised models, we also include results from a supervised model for comparison (see Appendix B.7), which did not use any probing as it directly predicts $(\alpha, \zeta) \in \mathbb{R}^2$.

6.2. Hierarchical JEPA

Ablation design. We sweep two orthogonal axes against a single fixed training schedule (Appendix C.4; 10 epochs AdamW, batch 12, bf16-mixed on one A100 40 GB). The *structural axis* toggles components of the hierarchy: **S1** is the full H-JEPA at $w_1 = 0.3$; **S2** ablates the inner pathway ($w_1 = 0$); **S3** additionally redirects the outer target to $y \equiv x$, collapsing both levels into denoising. The performance gap between S2 and S3 reflects the benefit of paired (next-window) supervision over same-window denoising. The *inner-weight axis* sweeps $w_1 \in \{0.3, 0.6, 1.0\}$ at $w_2 = 1$ under the default VICReg pair loss (Eq. 5). An MSE + SIGReg variant of the loss is examined separately in Appendix C.3.

The probing is performed on the deep-encoder-EMA features $\phi(x) = \text{mean}_{H_2, W_2}(\tilde{z}_2) \in \mathbb{R}^{128}$ (spatial mean pool of $\tilde{z}_2$ over (H_2, W_2)). We report z-score-normalized test MSE across (α, ζ); raw-unit numbers and the per-row sweep are

in Appendix C.

ID	w_1	Linear MSE ↓			KNN MSE ↓		
		α	ζ	Total	α	ζ	Total
S1	0.3	0.021	0.201	0.111	0.011	0.223	0.117
S2	–	0.025	0.244	0.135	0.013	0.281	0.147
S3	–	0.102	0.387	0.244	0.062	0.337	0.199
best	1.0	**0.014**	**0.173**	**0.093**	**0.009**	**0.189**	**0.099**

Table 2. H-JEPA frozen-encoder probes on the held-out test split, z-score MSE for (α, ζ) at the final (10th) epoch under the default VICReg objective; lower is better, best in bold. **S1**: full H-JEPA. **S2**: only-outer ablation ($w_1 = 0$). **S3**: outer-denoising ablation ($y \equiv x$). **best**: best inner weight $w_1^* \in \{0.3, 0.6, 1.0\}$ selected independently for the linear and kNN probes by their respective validation MSE (both happen to agree on $w_1^* = 1.0$). The full per-row sweep is in Appendix C.

What the table shows. Table 2 shows: (i) The inner pathway is useful: removing it (S1 → S2) inflates kNN total MSE from 0.117 to 0.147 (+26%). (ii) Paired (next-window) supervision substantially beats multi-level same-window denoising: redirecting the outer target to $y \equiv x$ (S2 → S3) almost doubles linear MSE (0.135 → 0.244), confirming that the temporal coupling carried by the paired target is the dominant learning signal. (iii) The inner-weight sweep is monotone in w_1—more shallow supervision helps—with the best-validated configuration at $w_1^* = 1.0$.

6.3. Capturing physics with FA-JEPA

We analyze the proposed architecture FA-JEPA with and without the physics-prior attention on the orientation field. For the latter, we use a learnable blend initialized at $\lambda = 0.1$ and $\tau = 1.0$, while retaining 0.1 dropout in the encoder and predictor; its best checkpoint achieved validation loss 0.6643. The model without this physics-prior attention serves as a regularized non-physics ablation with the same FA-JEPA backbone and 0.1 dropout, but with the orientation physics bias disabled; its best checkpoint achieved validation loss 1.4986. Both runs use Fourier Transform resizing to 224×224 and remain well below the 100M parameter limit. For more details refer to Appendix B.5.1.

The results are shown in Table 3. The FA-JEPA variant with standard (data) attention improves upon the baseline results (Table 1). Once the physics-prior attention was incorporated, we saw further improvements that improve the baseline in both linear and kNN probing. Interestingly, in linear probing, while there is a decay in performance in the estimation of α, the improvement over ζ is significant enough that the total MSE is improved. For the kNN, the MSE for both α and ζ improved. We compare the results on training and validation sets in Fig.12 in the Appendix.

Config	Linear MSE ↓			KNN MSE ↓			k
	α	ζ	Total	α	ζ	Total	
Data attn only	**0.017**	0.385	0.201	0.060	0.652	0.356	10
Physics attn	0.033	**0.291**	**0.162**	**0.033**	**0.550**	**0.292**	4

Table 3. z-score MSE on the test set for FA-JEPA with data attention only (first row) and with also the physics-prior attention (second row). Each row corresponds to the encoder checkpoint selected based on validation performance for the corresponding probe. For k-NN, the reported results also use the best k. The distance metric used is the l_2 cosine.

7. Discussion

7.1. Baseline

Two ablations are needed to firm up the claim that tokenization, not depth, drives the ViT3D win. First, Conv+Attn ×6 was cut at epoch 7 by HPC quota with the loss still descending, so its row mixes architecture and undertraining; matching ViT3D's ∼30-epoch budget would tell whether the deeper attention head plateaus or overtakes the FFT-only baseline. Second, ViT3D itself uses six blocks—the same depth as Conv+Attn ×6—so the tokenizer's contribution should be isolated by running ViT3D with a single block, and with no token mixing at all (patch embed + linear). If both still beat every conv variant, the win is unambiguously in the spatiotemporal patching; if the shallow variant collapses, depth and tokenisation are co-dependent and the current takeaway weakens.

7.2. FA-JEPA

FA-JEPA introduces a field-aware attention mechanism that models interactions across physical fields, together with a physics-prior attention module tailored to capture spatial structure in the orientation field.

Field-aware attention. We analyze the 4×4 field-attention matrix averaged over validation samples, encoder blocks, and attention heads. Each row corresponds to the queries and columns to the key/value physics fields. Fig. 3 shows attention relative to a uniform baseline.

Velocity, orientation, and strain-rate predominantly attend to other fields rather than themselves, indicating that the model captures cross-field coupling rather than treating channels independently. Moreover, the attention patterns in Fig. 3 suggest that FA-JEPA emphasizes the same coupled variables that appear in the kinetic-theory formulation of active suspensions (Maddu et al., 2024).

The strongest directed enrichment is from orientation to velocity. When updating orientation tokens, the model attends to velocity tokens above the uniform baseline by approximately 0.18. This corresponds to orientation assigning roughly 43% of its field-attention mass to velocity. This is

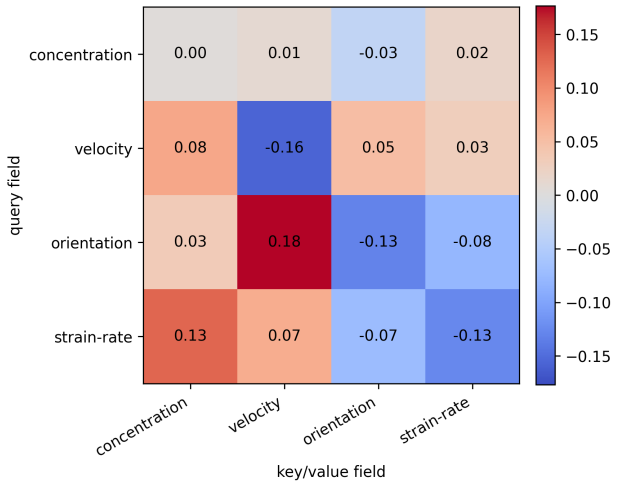


Figure 3. Field-attention enrichment matrix for FA-JEPA, computed as mean attention minus the uniform baseline $1/4$. Positive values indicate above-uniform attention, highlighting cross-field coupling.

consistent with the active-suspension equations, in which orientational dynamics are coupled to the local flow-gradient tensor $\nabla u + 2\zeta D$. The reverse direction is weaker; this asymmetry is physically plausible due to the directional nature of field attention.

Physics-prior attention. Incorporating the physics prior improves performance and induces structured spatial attention. To analyze this, we loaded the trained FA-JEPA corresponding to the one used to report the results in Table 3 and passed validation samples through the frozen encoder. For each encoder block, we extracted the spatial attention corresponding to the orientation field. We then recomputed three attention matrices: the data attention, the physics-prior attention with the learned $\tau \sim 1.001$ (Eq.4), and the effective blended attention actually used by the model $\lambda \sim 0.099$ (Eq.3). These matrices were averaged globally to obtain a compact summary of the dominant orientation-field spatial attention structure. Fig. 4 shows the learned data attention (left), the physics-prior attention (middle), and the difference between the effective blended attention and the learned data attention (right) across 196 spatial patches.

Compared to the learned data attention, which exhibits broad, weakly structured patterns, the physics-prior attention introduces more localized and organized interactions. For a given query patch, attention is no longer distributed uniformly across the same dominant key locations; instead, certain spatial patches are attended to more than others, revealing structured relations between orientation patterns across space. The effective blended attention, which is the actual attention used by the encoder after weighting the prior with the learned mixing coefficient, remains close to the learned data attention but shows a visible shift toward this structured pattern. This suggests that the physics-prior term aids spatial attention by introducing orientation-aware structure without completely overriding the learned attention mechanism.

Limitations. One major limitation, which we did not fully explore, is controlling the power of the predictor. In training, we achieved low prediction loss. Due to the highly nonlinear nature of the predictor, however, the encoder is not forced to produce representations in the representation space that can be extracted by linear and kNN probes as effective as other methods discussed in this work. Another limitation is the size of the representation, which exceeds 2M dimensions. To better align with the task of linear and kNN probing, global pooling may be ineffective in comparison to directly limiting the size of the representation.

7.3. H-JEPA

Why H-JEPA is effective. Auxiliary self-supervised pretext tasks are widely used to enrich learned representations (Zhou et al., 2022; Balestriero et al., 2023). In our case we attribute H-JEPA's gain over the JEPA baseline (Table 1) specifically to weight sharing of the shallow encoder f_{θ_1}: gradient reaches θ_1 along three paths—the inner predictor g_{ϕ_1}, the outer predictor g_{ϕ_2} via the $\tilde{z}_1$ skip, and g_{ϕ_2} via f_{θ_2}—so the inner denoising loss and the outer next-window loss both shape the same parameters. The inner pretext pressures θ_1 to retain frame-local instantaneous structure; the outer pretext pressures the same parameters to retain content predictable across paired next-window targets. Features satisfying both lie in a tighter intersection than features satisfying either alone, which is why probing on $\tilde{z}_2$ benefits even though no probe ever reads $\tilde{z}_1$: θ_1 lies in $\tilde{z}_2$'s forward path, and any reshaping induced by the inner loss propagates upward. The ablations corroborate this. Removing the inner pretext (S1 $\rightarrow$ S2) inflates kNN total MSE by 26%, and the inner-weight sweep is monotone in w_1 throughout $[0.3,\ 1.0]$, with more shallow supervision continuing to help rather than competing with the outer loss—as expected if the two pretexts shape a shared trunk along complementary axes.

Limitations. We identify three caveats specific to H-JEPA. First, the inner pretext is fixed to Gaussian latent denoising; alternative same-window pretexts would reshape what "instantaneous structure" the trunk retains, and we have not measured this sensitivity. Second, the inner loss and the shallow-to-outer skip are coupled in our ablation grid (S2 disables the inner loss but keeps the skip), so we measure their joint effect rather than the marginal contribution of each. Third, ζ's test MSE remains an order of magnitude above α's across all rows: the pretraining loss is uniform across latent components, while probing mean-pools $\tilde{z}_2$ over (H_2, W_2) and plausibly washes out the local struc-

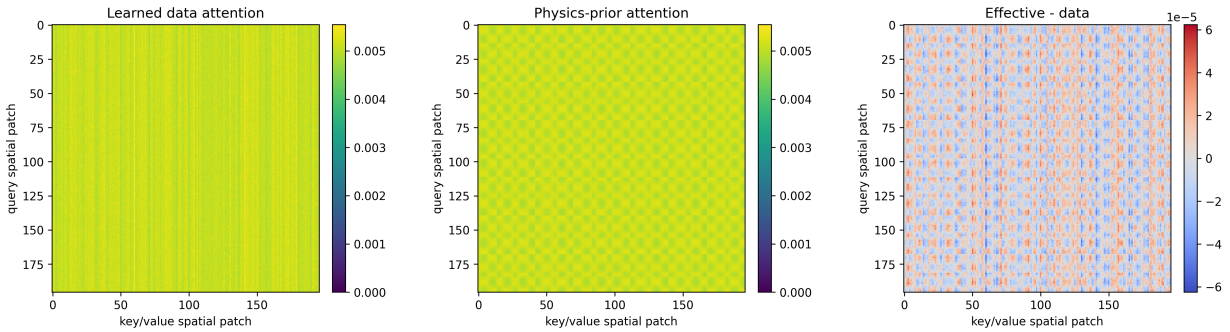


Figure 4. All-block average of orientation-field spatial attention. The learned data attention shows broad vertical bands, indicating key spatial patches are attended by many query patches. The physics-prior attention introduces a more structured pattern based on orientation similarity. The difference between the effective blended attention and the learned data attention shows how the learned physics-prior mixing slightly redirects attention toward this structured pattern.

tures encoding ζ—a training-loss/probing-task mismatch shared with other JEPA-style methods that our architectural changes alone cannot lift.

8. Conclusions

We presented physics-aware extensions of the JEPA framework for learning representations of complex dynamical systems from high-dimensional observations. Our results show that though standard JEPA baselines are good general-use models, introducing physics-compatible preprocessing and spatiotemporal tokenization allows for significant improvements in representation quality and downstream parameter inference.

Building on this foundation, we proposed two complementary approaches. H-JEPA improves representation learning by decomposing spatial and temporal modeling through hierarchical supervision. Overall, H-JEPA achieves the strongest performance, reducing total test MSE relative to the baseline by approximately 64% for the linear probe and 70% for the kNN probe, while maintaining a small parameter count ($\sim$4M). FA-JEPA introduces field-aware attention and a physics-prior attention mechanism, enabling the model to capture interactions between physical observables and incorporate domain knowledge directly into the attention process. Together, these results highlight that aligning representation learning with the structure of the underlying physics is key to modeling complex systems.

Our findings suggest that future progress in scientific representation learning will depend less on scaling generic architectures and more on designing inductive biases that reflect the governing dynamics. We hope this work contributes to bridging data-driven methods and physical modeling, and provides a foundation for more effective learning in scientific domains.

Accessibility

Code and checkpoints are publicly available. The implementation uses standard frameworks and runs on common GPUs. Documentation and evaluation details ensure reproducibility; further details are in the appendix.

Software and Data

The model checkpoints are available on Hugging Face at DL-Project-Active-Matter repository. The codebase is available at GitHub repository.

Statement of Contributions

C. Ji implemented the baseline method and developed and benchmarked its extensions. Z. Shi developed, implemented, and evaluated H-JEPA and its variants. R. Wang developed and implemented FA-JEPA. R. Yalovetzky developed the physics-prior attention mechanism and performed post-processing and analysis of the FA-JEPA results and its variants. All authors contributed to the conceptualization of the work, interpretation of results, and writing of the manuscript.

Impact Statement

This work develops physics-aware representation learning methods for modeling complex dynamical systems. By improving the alignment between learned representations and underlying physical processes, the approach may support advances in scientific discovery and simulation-based modeling. Potential applications include materials science, fluid dynamics, and biological systems. While the methods are intended for research use, their application in broader domains highlights the importance of reliability and interpretability in scientific machine learning.

References

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., and Schmid, C. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6836–6846, 2021.

Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Balestriero, R. and LeCun, Y. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian, Y., Schwarzschild, A., Wilson, A. G., Geiping, J., Garrido, Q., Fernandez, P., Bar, A., Pirsiavash, H., LeCun, Y., and Goldblum, M. A cookbook of self-supervised learning, 2023. URL https://arxiv.org/abs/2304.12210.

Bardes, A., Ponce, J., and LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., and Ballas, N. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Bardhan, J., Agrawal, R., Tilak, A., Neeraj, C., and Mitra, S. Hep-jepa: A foundation model for collider physics. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*, 2025.

Chen, X. and He, K. Exploring simple siamese representation learning, 2020. URL https://arxiv.org/abs/2011.10566.

Goswami, R. G., Krishnamurthy, P., LeCun, Y., and Khorrami, F. Osvi-wm: One-shot visual imitation for unseen tasks using world-model-guided trajectory generation. *arXiv preprint arXiv:2505.20425*, 2025.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. Bootstrap your own latent: A new approach to self-supervised learning, 2020. URL https://arxiv.org/abs/2006.07733.

Herde, M., Raonić, B., Rohner, T., Käppeli, R., Molinaro, R., De Bezenac, E., and Mishra, S. Poseidon: Efficient foundation models for pdes. *Advances in Neural Information Processing Systems*, 37:72525–72624, 2024.

LeCun, Y. et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1): 1–62, 2022.

Lin, T., Wang, Y., Liu, X., and Qiu, X. A survey of transformers. *AI open*, 3:111–132, 2022.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Maddu, S., Weady, S., and Shelley, M. J. Learning fast, accurate, and stable closures of a kinetic theory of an active fluid. *Journal of Computational Physics*, 504: 112869, 2024.

Maes, L., Lidec, Q. L., Scieur, D., LeCun, Y., and Balestriero, R. Leworldmodel: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

McCabe, M., Blancard, B. R.-S., Parker, L. H., Ohana, R., Cranmer, M., Bietti, A., Eickenberg, M., Golkar, S., Krawezik, G., Lanusse, F., et al. Multiple physics pretraining for physical surrogate models. *arXiv preprint arXiv:2310.02994*, 2023.

McCabe, M., Mukhopadhyay, P., Marwah, T., Blancard, B. R.-S., Rozet, F., Diaconu, C., Meyer, L., Wong, K. W., Sotoudeh, H., Bietti, A., et al. Walrus: A cross-domain foundation model for continuum dynamics. *arXiv preprint arXiv:2511.15684*, 2025.

Mur-Labadia, L., Muckley, M., Bar, A., Assran, M., Sinha, K., Rabbat, M., LeCun, Y., Ballas, N., and Bardes, A. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

Nam, H., Lidec, Q. L., Maes, L., LeCun, Y., and Balestriero, R. Causal-jepa: Learning world models through object-level latent interventions. *arXiv preprint arXiv:2602.11389*, 2026.

Qu, H., Morel, R., McCabe, M., Bietti, A., Lanusse, F., Ho, S., and LeCun, Y. Representation learning for spatiotemporal physical systems. *arXiv preprint arXiv:2603.13227*, 2026.

Yee, B. and Koh, P. Pi-jepa: Label-free surrogate pretraining for coupled multiphysics simulation via operator-split latent prediction. *arXiv preprint arXiv:2604.01349*, 2026.

Zhou, G., Pan, H., LeCun, Y., and Pinto, L. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., and Kong, T. ibot: Image bert pre-training with online tokenizer, 2022. URL https://arxiv.org/abs/2111.07832.

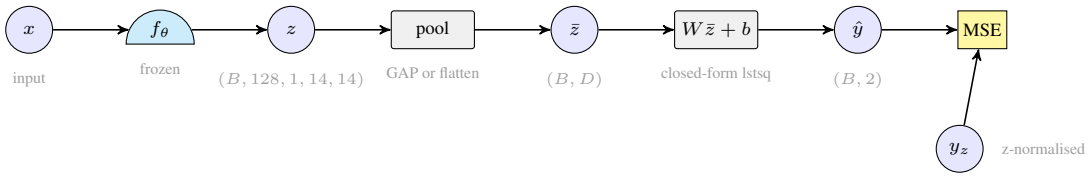


Figure 5. Linear probe. The encoder is frozen (cyan); features z are spatially pooled into $\bar{z} \in \mathbb{R}^D$ and a closed-form least-squares fit predicts the z-normalised parameter vector $\hat{y} = W\bar{z} + b$. No SGD; the regulariser is the implicit minimum-norm solution.

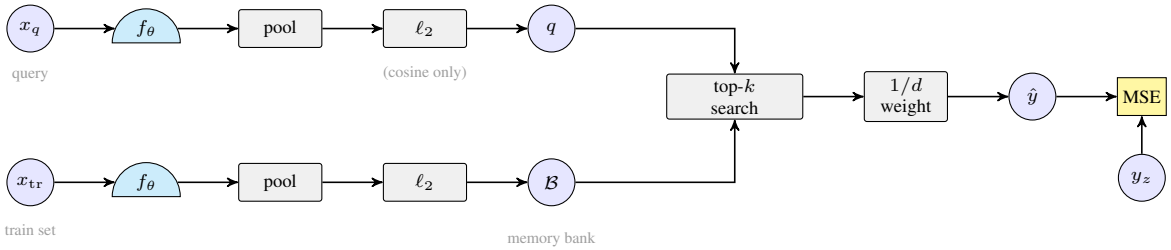


Figure 6. k-NN probe. The query x_q and every training sample x_{tr} pass through the same frozen encoder and pool, producing the query feature q and a memory bank $\mathcal{B}$. For the cosine metric both are ℓ_2-normalised; for Euclidean they are not. The k nearest training points contribute to $\hat{y}$ with inverse-distance weights.

A. Probing methods

We consider linear and kNN probing, whose diagrams are shown in Fig. 5 and 6.

B. Baseline and its extensions

As discussed in the main text, we developed and benchmarked multiple extensions over the baseline, whose results are shown in 4.

B.1. Precision and batch size

A single A100 with 40 GB of VRAM cannot fit fp32 training at bs=8. We confirm empirically that VICReg collapses at bs=2 (Appendix Figure 7, grey): the linear probe MSE decreases for the first 11 epochs and then climbs as the projected features lose variance. Switching to bfloat16 mixed precision restores bs=8 within the same memory budget; this configuration (blue in Appendix Figure 7) converges to linear MSE 0.243 and serves as the reference point for the subsequent ablations.

B.2. FFT preprocessing

Replacing the spatial crop with the band-limited FFT resize (Section 3) drops linear MSE $0.243 \to 0.212$ and k-NN MSE $0.217 \to 0.174$ (orange in Appendix Figures 7–8). The gain is modest on `active_matter`—whose native 256×256 grid needs no crop in the first place—but transfers across every encoder variant we tried, so we fix FFT preprocessing for the rest of this section.

Encoder family. We test three encoders that span an axis of *when* temporal context enters the representation: (i) the **3D ConvNeXt** baseline, fusing time inside the conv stack; (ii) **Conv+Attn**, appending $L \in \{1, 6\}$ pre-norm transformer blocks at the top of the conv stack—the attention here mixes only the post-conv map, after T has collapsed to 4 and the spatial grid to 4×4 (Appendix Figure 10); (iii) **ViT3D**, replacing the conv stack entirely with a single $4\times16\times16$ patch embedding and six transformer blocks operating on the resulting spatiotemporal tokens (Appendix Figure 11). Both Conv+Attn$\times6$ and ViT3D place six transformer blocks in the encoder, so the comparison isolates one question: whether the deciding factor is the depth of attention or what the tokens encode.

Config	Linear MSE ↓			k-NN MSE ↓				
	mean	α	ζ	mean	α	ζ	k	d
VICReg baseline	0.2594	0.0591	0.4596	0.3251	0.0812	0.5691	5	euc
VICReg + FFT (ep 19)	0.2201	0.0367	0.4034	0.3044	0.0369	0.5718	10	euc
SIGReg, FFT (did not conv.)	0.6306	0.1905	1.0707	0.5496	0.2026	0.8965	50	cos
Conv+Attn, FFT (ep 17)	0.2512	0.0492	0.4532	0.2589	0.0528	0.4650	5	euc
Conv+Attn $\times 6^{\star}$ (ep 7)	0.2667	0.0350	0.4983	0.2569	0.0350	0.4788	10	euc
ViT3D-d6, FFT (ep 29)	**0.1065**	**0.0161**	**0.1970**	**0.1199**	**0.0088**	**0.2310**	20	cos

Table 4. Test MSE ↓ at the best-validation epoch. Bold marks the best per column. *Conv+Attn $\times 6$ was cut at epoch 7 by HPC quota; its row is the best of four checkpoints, not a converged run. Per-parameter columns confirm ζ is the harder target by across every configuration.

B.3. Encoder ablations

With bfloat16, bs=8, and FFT fixed, we vary only the encoder. Each row of Table 4 isolates one architectural choice from Section 3.

SIGReg. Under the same encoder, SIGReg did not converge: the linear probe MSE *rises* throughout pretraining (red dotted in Appendix Figures 7–8). We revert to VICReg for all subsequent runs.

Conv+Attn. A single transformer block on top of the conv stack edges linear MSE to 0.251 and k-NN to 0.259 by epoch 17, barely improving on the FFT-only baseline. Stacking six pre-norm blocks at the same position yields a steeper trajectory—linear MSE descends $0.753 \to 0.546 \to 0.370 \to 0.267$ across epochs $1, 3, 5, 7$—but HPC quota cut training at epoch 7, before the curve could overtake the FFT-only baseline. Both variants exhibit the defect predicted in Section 5: the attention operates only on the post-conv spatial map, so depth alone cannot recover the temporal axis the strided convolutions have already collapsed.

ViT3D. Replacing the conv stack with a $4\times16\times16$ patch embedding followed by six transformer blocks drops linear MSE to $\mathbf{0.107}$ and k-NN to $\mathbf{0.120}$—roughly $3\times$ better than every CNN variant we tried. The *same* six transformer blocks were inert on top of a conv stack and dominant on spatiotemporal patches; the difference is the tokenizer.

B.4. Compute accounting

The largest model, *Temporal* ViT3D, has a total parameter count of 15.37 M for encoder, 7.12 M for decoder, a total of 22.5 M. A single ablation row trains in roughly 20 GPU-h on one A100 40 GB, using $\sim$ 36 GB of the VRAM; the 6 training rows (all trained from scratch) plus the 8 final-probe runs ($\sim$ 2 GPU-h each) total $\leq$ 270 GPU-h, within the project's per-student quota.

B.5. FA-JEPA

B.5.1. Training

Data preprocessing. We set up $k = 16$ time frames and perform spatial resizing such that the dimensions are 224×224 pixels per frame. The number of channels is 11. Bilinear interpolation is used in the baseline model (Qu et al., 2026) setup. In order to preserve the periodic boundary conditions, we opt for the Fast Fourier Transform to preprocess the data.

Grouping of channels. We group the channels according to their respective physical fields: concentration as a scaler, velocity as a 2D vector, and orientation and strain rate as 2×2 tensors.

Patching. To prepare the data as embeddings to input inside the encoder transformer, we employ a 3D convolution for each of the four field stems. The tubelet size is $2 \times 16 \times 16$, compressing 2 time frames and 16×16 pixel regions in both vertical and horizontal spatial directions. Thus, the path embeddings become $(B, T, H, W, F, D) = (16, 8, 14, 14, 4, 384)$ and we flatten the intermediate dimensions to finally obtain the input to the encoder to be $(16, 6272, 384)$. Note that these dimensions are the same for the four fields such that there is one token per patch and field combination.

Encoder. It takes the output of the patching and outputs $(16, 6272, 384)$. This means that we have 6272 tokens and the

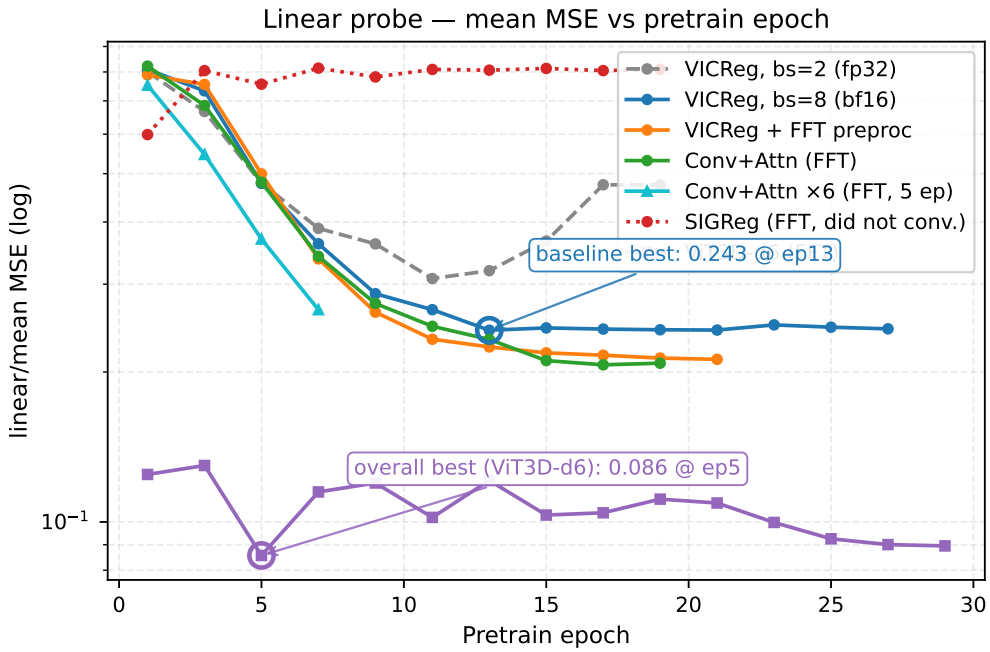


Figure 7. linear mean MSE ↓ vs. pretrain epoch. Batch size = 2 (grey, dashed) collapses past epoch 11; batch size = 8 with bfloat16 (blue) is the baseline. SIGReg (red, dotted) diverges. ViT3D (purple) is approximately 3× better than every CNN variant.

representation space has a dimension of 384. When performing the probing, we implemented global average pooling that takes 6272 tokens. Thus, each probing sample has 384 dimensions.

Attention mechanisms. The order of these attention mechanisms has an impact on the encoder output. We propose to order these factorized blocks as field, spatial, then temporal attention so each local token first integrates coupled physical variables, then propagates this field-aware state across space and finally across time for prediction. We incorporate a factorized attention for two reasons: (1) to allow for self-attention over fields, space, and time, while (2) reducing computational complexity. The encoder has six heads per attention layer.

Compute accounting. FA-JEPA was trained on a single NVIDIA GB10 with a Grace-Blackwell GPU with 128GB of unified memory. All models used bf16 mixed precision. Across different dropout, batch sizes, sampling stride settings, the amount of total memory required ranged from 70GB (no dropout) to 115GB (with dropout).

B.6. Benchmark

Physics-prior attention impact on α and ζ estimation. We observe in Fig.12 that the physics-prior attention improve the performance of the total MSE. Interestingly, in linear probing, while there is a decay in performance in the estimation of α, the improvement over ζ is significant enough that the total MSE is improved. For the kNN, the MSE for both α and ζ improved.

Impact of dropout. We study the impact of leveraging the dropout technique in the model. We compare the performance using dropout (0.1) both in the encoder and predictor and with no dropout at all. In particular for the kNN probing we plot the total MSE (in validation) as a function of the k in Fig.13 for a few different epochs. In this case we consider stride 16, while for the final results shown in the table we did probing with stride 1 for those models that showed best results.

Dropout slows down specialization (panel (a)), making the early representations (first epochs), which are more general, perform better in predicting the (α, ζ) and as the epochs go by, the training overfits the JEPA objective and the prediction performance degrades. In contrast, without dropout, at later epochs, the MSE reduces. The model learns sharper structure over time. We have observed that the best kNN probe is at the last epoch we trained (number 12). This indicates a tradeoff between global linear accessibility and local metric fidelity: dropout regularizes the representation space in a way that benefits linear readout, but appears to reduce the fine-grained neighborhood structure required by kNN. In terms of the best

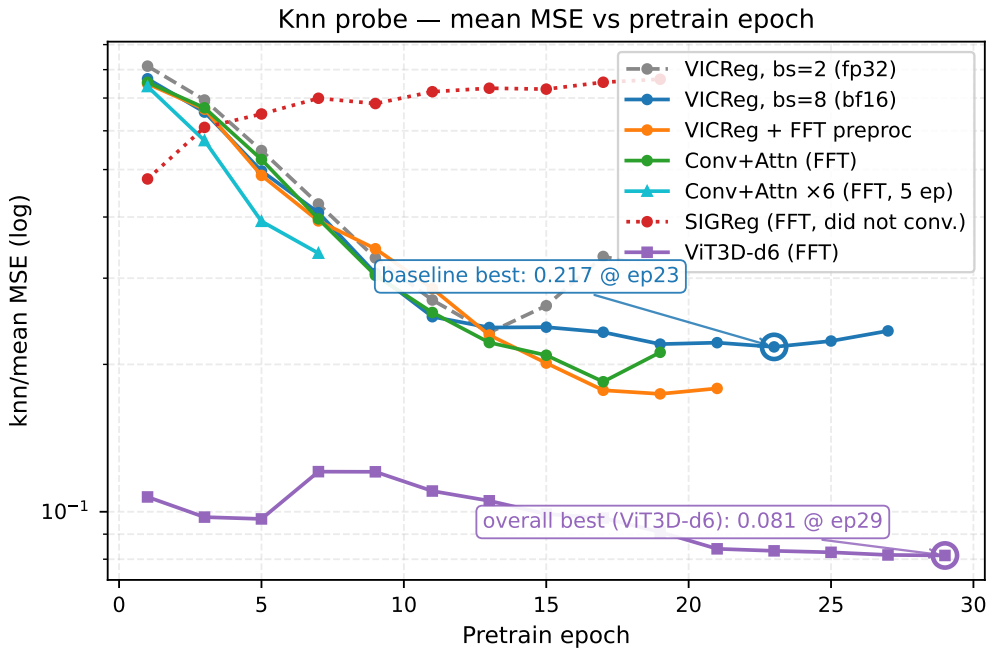


Figure 8. knn mean MSE ↓ vs. pretrain epoch (best k and metric per epoch). FFT (orange) opens a gap on the bilinear baseline after epoch 13; ViT3D continues to improve past epoch 25.

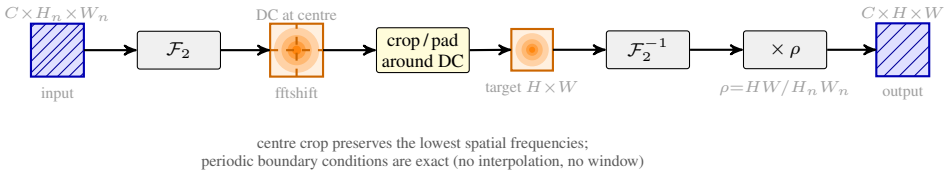


Figure 9. FFT-based resize. $\mathcal{F}_2$ is the 2-D Fourier transform; the centre crop or zero-pad happens around the DC bin so periodic boundary conditions are preserved exactly.

k, there does not seem to be a correlation with the epoch.

We study the impact of dropout on the performance of the linear probing in Fig. 14. With dropout, linear probing exhibits stable and consistent convergence across epochs, with closely aligned training, validation, and test performance, indicating improved generalization of the learned representations. Without dropout, linear probing exhibits unstable convergence with large fluctuations across epochs, including a pronounced spike in validation error approaching the baseline (MSE ≈ 1), and a significant train–validation gap, indicating overfitting and poor generalization.

B.7. Supervised Baseline

We use the existing encoder and predictor in FA-JEPA to train a supervised model as the baseline. Instead of JEPA, we simply keep one encoder and change the predictor's output dimension to 2, which corresponds to α and ζ predictions. With this baseline, we achieved an MSE of 0.04585 on the validation set and 0.05226 on the test set.

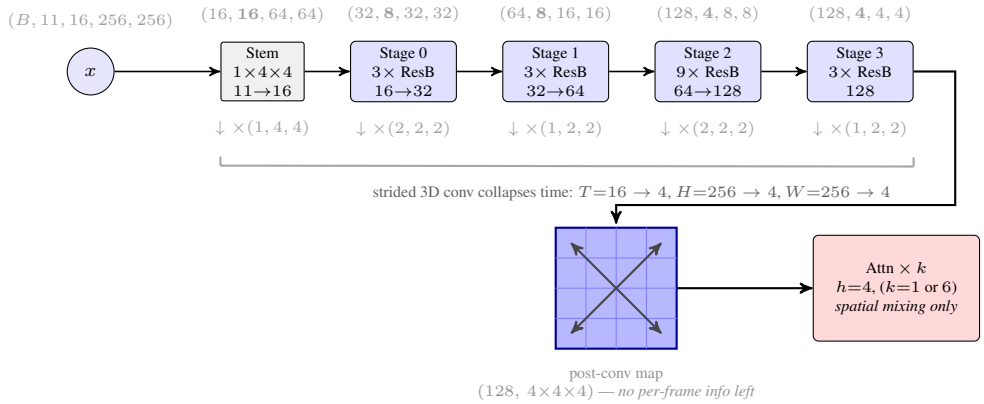


Figure 10. Conv+Attn loses temporal information. **Top:** the strided 3D convolution collapses $T = 16 \to 4$ and $H, W = 256 \to 4$ before the attention runs. **Bottom:** the attention block sees a flat spatial map, so only spatial mixing is left.

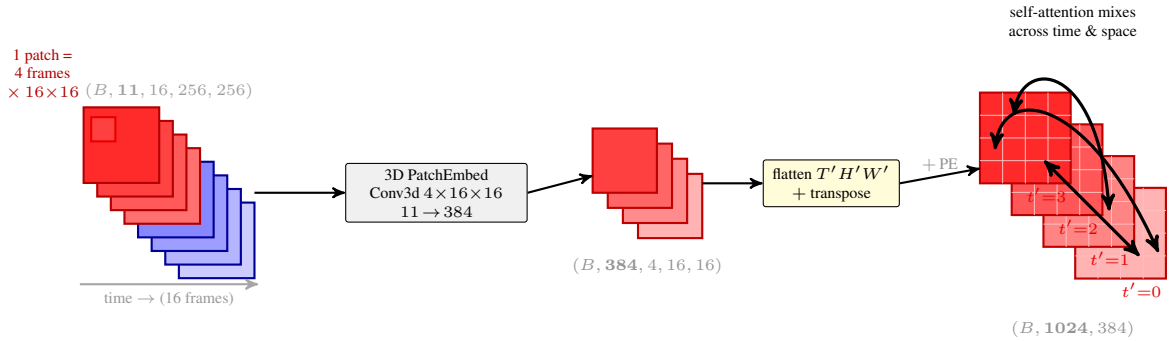


Figure 11. ViT3D extracts temporal information at the tokenizer. Each red tube spans 4 frames × 16 × 16 pixels and becomes one 384-dimensional token, so every token already carries temporal context *before* attention runs. The resulting $4 \times 16 \times 16 = 1024$-token cube is then mixed across both axes by six transformer blocks (omitted here).

C. Hierarchical JEPA — implementation details

C.1. Forward pass

The complete two-level forward pass of HJEPA is

$$\tilde{z}_1 = f_{\theta_1}(\tilde{x}), \qquad \tilde{z}_2 = f_{\theta_2}(\tilde{z}_1), \tag{6}$$

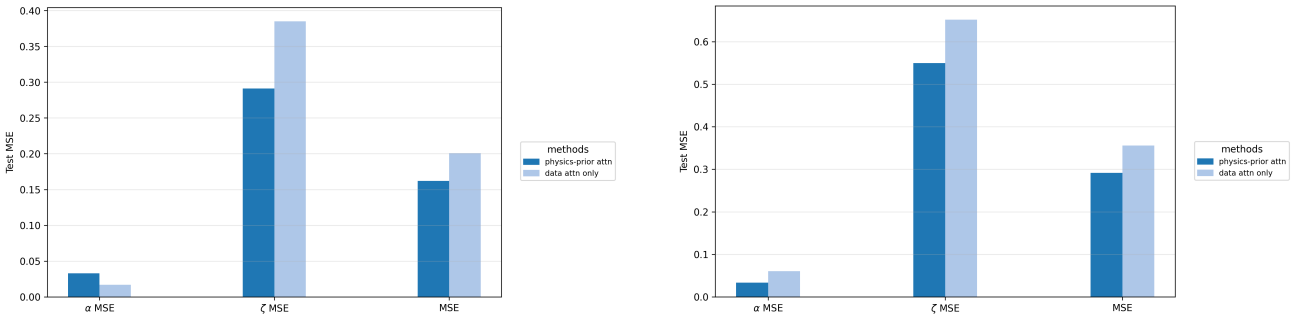
$$\hat{z}_1 = g_{\phi_1}(\tilde{z}_1), \qquad z_1^{\text{tgt}} = f_{\theta_1'}(x), \tag{7}$$

$$\hat{z}_2 = g_{\phi_2}(\tilde{z}_1, \tilde{z}_2), \qquad z_2^{\text{tgt}} = f_{\theta_2'}\big(f_{\theta_1'}(y)\big), \tag{8}$$

where $\tilde{x} = x + \sigma\varepsilon$, $\sigma = 1$, $\varepsilon \sim \mathcal{N}(0, I)$, y is the next T frames of the same trajectory at sliding stride 1 (or $y \equiv x$ for the outer-denoising ablation). EMA targets receive no gradient and update as $\theta_k' \leftarrow \tau\theta_k' + (1-\tau)\theta_k$ with τ=0.996.

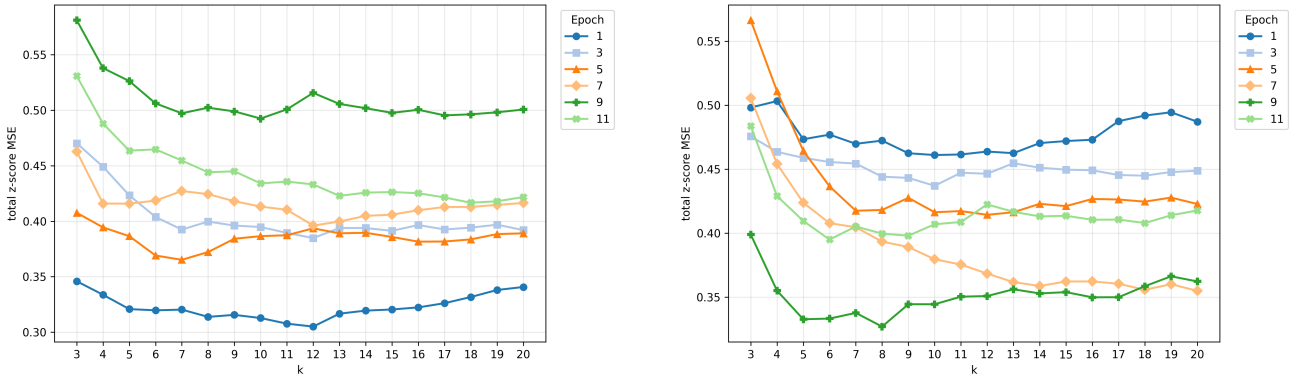
C.2. Encoders and predictors

The split-encoder configuration shared across all rows is `dims` $= [16, 32, 64, 128, 128]$ with `num_res_blocks` $= [3, 3, 3, 9, 3]$ ConvNeXt-style residual blocks (7×7 depth-wise + GELU, layer scale 10^{-6}); we cut the stack after stage 3, so that the shallow encoder f_{θ_1} produces $\tilde{z}_1 \in \mathbb{R}^{64\times4\times56\times56}$ and the deep encoder f_{θ_2} produces $\tilde{z}_2 \in \mathbb{R}^{128\times14\times14}$ (the deep encoder collapses time). The inner predictor is a 3D residual stack of width 128 with 3 residual blocks; the outer predictor projects $\tilde{z}_1$ to $\mathbb{R}^{16\times14\times14}$ via a strided 4×4 Conv2d (stride 4), concatenates with $\tilde{z}_2$, and runs a 2D residual stack of width 256 with 1 residual block. Total trainable parameters: 3.97 M (enc$_1$ 0.27 M, enc$_2$ 2.20 M, inner 0.66 M, outer 0.84 M). The EMA target encoders add 2.47 M non-trainable parameters; the predictors and corruption module have no EMA copy.



(a) Linear probing *(b)* kNN probing

Figure 12. z-score MSE on the test dataset. Panel (a) corresponds to linear probing while panel (b) to kNN probing. Two methods are being compared: FA-JEPA with data attention only (light blue) and with also the physics-prior attention (dark blue).



(a) With dropout *(b)* Without dropout

Figure 13. Total z-score MSE of kNN probing over the validation dataset (with stride 16) as a function of k.

C.3. MSE + SIGReg loss variant

The default HJEPA objective (Eq. 5) uses VICReg, whose variance and covariance terms supply anti-collapse pressure on the predicted latents directly; with VICReg the per-level pair loss decomposes into similarity, variance and covariance terms with coefficients $(\text{sim}, \text{std}, \text{cov}) = (2, 40, 2)$, applied in 5 chunks for memory, matching the JEPA baseline of (Qu et al., 2026).

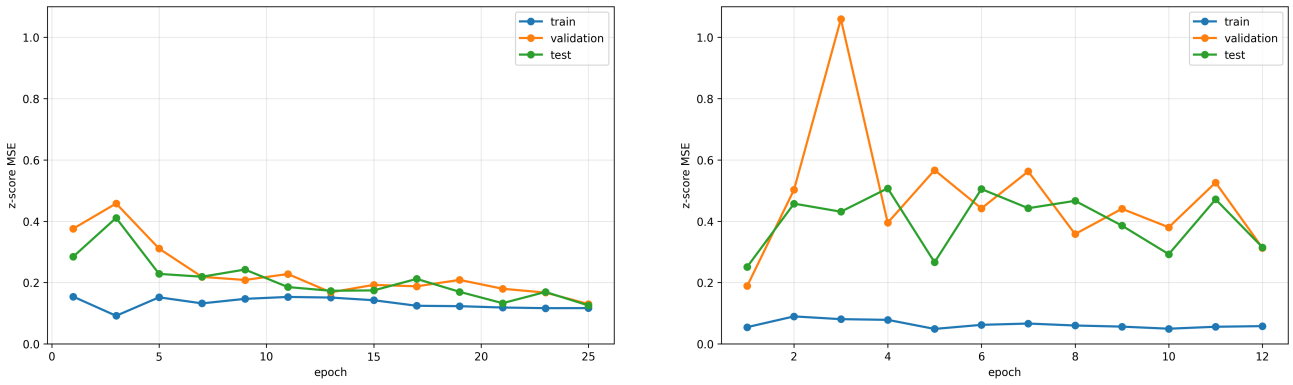
For an ablation we replace ℓ_{V} with an MSE pair loss and re-introduce an explicit encoder-side SIGReg regularizer (Balestriero & LeCun, 2025; Maes et al., 2026) on each online latent:

$$\mathcal{L}^{\text{S}} = \sum_{l \in \{1,2\}} w_l \Big(\ell_{\text{MSE}}(\hat{z}_l, z_l^{\text{tgt}}) + \rho\, R(\tilde{z}_l) \Big), \tag{9}$$

where R is the SIGReg test of isotropic-Gaussian projections with $K = 17$ knots and $P = 1024$ projections (identical to the Field-aware choice; SIGReg is applied only to the online latents, with no target), and $\rho = 0.09$ is the LeWorldModel-attenuated SIGReg/MSE ratio of (Maes et al., 2026). The ratio-matched scaling $\lambda_{z_k} = \rho \cdot w_k$ keeps the per-layer SIGReg-to-MSE balance constant as w_1 varies, so that the inner-weight sweep cleanly isolates the structural hyperparameter from the within-layer pair-loss / regularizer balance. This ablation tests whether an explicit encoder-side anti-collapse term improves the learned representation beyond the implicit one provided by VICReg.

C.4. Training schedule

All ablation rows share a single training schedule: 10 epochs of AdamW (cosine learning rate $10^{-3} \to 10^{-6}$, 2-epoch warmup, weight decay 0.05, betas $(0.9, 0.95)$), batch size 12 in bf16-mixed precision on one A100 40 GB, gradient clipping at norm 1, seed 42, EMA momentum $\tau = 0.996$. Frozen probing follows the Field-aware protocol: closed-form linear least



(a) With dropout *(b)* Without dropout

Figure 14. Total z-score MSE of linear probing over the validation dataset (with stride 16) as a function of the epoch. Each color (and its color) refer to different split of the data: training, validation and test.

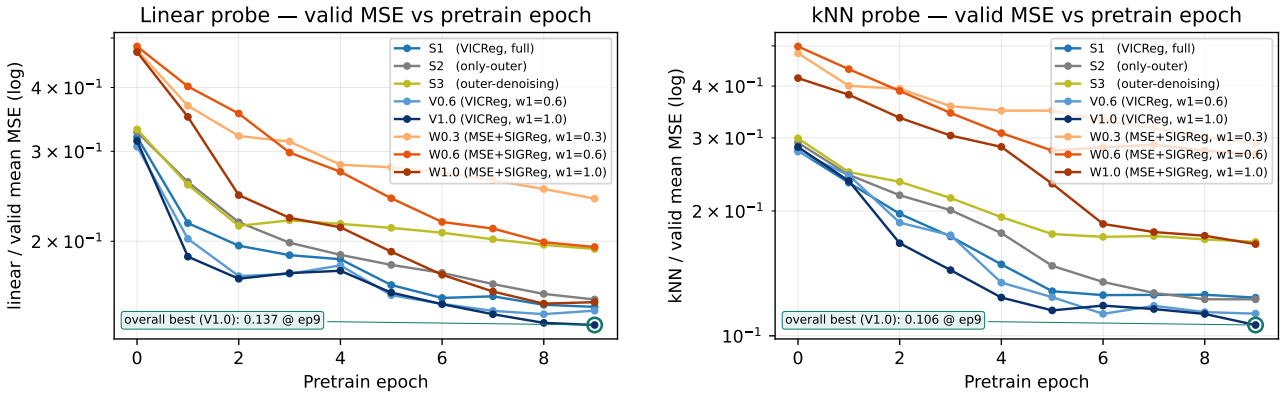
squares and inverse-distance-weighted kNN with k swept over $\{3, \dots, 20\}$ and selected by validation MSE per run, on deep-encoder features $\phi(x) = \text{mean}_{H_2,W_2}(\tilde{z}_2) \in \mathbb{R}^{128}$ (spatial mean pool of $\tilde{z}_2$ over (H_2, W_2)). We show the MSE as a function of the epoch in Fig. 15 for the different variants considered.

ID	w_1	Linear test MSE ↓			kNN test MSE ↓			
		α	ζ	Total	α	ζ	Total	best k
Structural axis (VICReg)								
S1 (full HJEPA)	0.3	0.021	0.201	0.111	0.011	0.223	0.117	17
S2 (only-outer)	–	0.025	0.244	0.135	0.013	0.281	0.147	20
S3 (outer-denoising)	–	0.102	0.387	0.244	0.062	0.337	0.199	13
VICReg sweep								
V0.3	0.3	0.021	0.201	0.111	0.011	0.223	0.117	17
V0.6	0.6	0.015	0.197	0.106	**0.009**	0.201	0.105	20
V1.0	1.0	0.014	**0.173**	**0.093**	0.009	**0.189**	**0.099**	20
MSE + SIGReg sweep								
W0.3	0.3	0.024	0.479	0.251	0.036	0.574	0.305	20
W0.6	0.6	0.014	0.375	0.195	0.030	0.451	0.240	18
W1.0	1.0	**0.013**	0.273	0.143	0.013	0.308	0.161	20

Table 5. Per-row HJEPA test MSE (z-score; lower is better, best in bold). All rows share 10 epochs, batch 12, bf16-mixed, A100 1×, seed 42, σ=1, stride 1, τ=0.996, w_2=1, and the encoder / predictor dimensions of the main text. The kNN column reports the test MSE at the $k \in \{3, \dots, 20\}$ selected by validation MSE per run.

C.5. Compute accounting

A single ablation row trains in roughly 10 GPU-h on one A100 within $\sim$ 19 GB VRAM; the 8 training rows (all trained from scratch) plus the 8 final-probe runs ($\sim$ 2 GPU-h each) total $\leq$ 100 GPU-h, within the project's per-student quota. S1 doubles as the w_1=0.3 point of the VICReg inner-weight sweep, so the VICReg sweep is reported with three distinct checkpoints (S1, V0.6, V1.0).



(a) Linear probe. *(b)* kNN probe.

Figure 15. Frozen-probe validation MSE vs. pretraining epoch for each ablation row. Both recipes are monotone in w_1 within their family (VICReg in cool blues, MSE + SIGReg in warm oranges); structural ablations S2/S3 (greys) sit above the VICReg sweep, confirming that the inner pathway and paired outer supervision both contribute. The "overall best" annotation circles the lowest validation point across all runs.