TAKENS-JEPA: SCIENTIFIC INFERENCE ACROSS SAMPLING INTERVALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Physical processes are often recorded at different time intervals, and a model trained at one interval can become unreliable when that interval changes. We ask whether a self-supervised representation can support scientific inference when the sampling interval at deployment differs from that used for downstream training. Takens' embedding theorem suggests an answer: histories that end at the same moment describe the same underlying state even when their samples are spaced differently. Takens-JEPA turns this correspondence into an auxiliary ranking objective for a joint-embedding predictive architecture, drawing together histories that share an endpoint and separating them from histories with other endpoints. It needs no new data, since the extra histories are subsampled from stored trajectories, and adds no inference cost. On Active Matter, Rayleigh-Bénard convection, and Shear Flow from The Well, we fit parameter-estimation and endpoint-state reconstruction heads to frozen representations at one interval and evaluate them unchanged at coarser, partly unseen intervals. On Active Matter, Takens-JEPA roughly halves the predictive baseline's parameter-estimation error at every shifted interval and cuts endpoint-state error by a third; on Shear Flow it lowers parameter-estimation error under shift by up to 62% at the cost of a small loss at the training interval; gains on Rayleigh-Bénard are smaller. A control that applies the predictive objective at several intervals recovers part of this gain: Takens-JEPA is more accurate at the training interval and the nearest unseen interval, matches the control at the coarsest interval, and uses about one-third less training compute.

1 INTRODUCTION

Scientific inference often begins with a limited recording of a physical process. From that recording, we may wish to estimate governing parameters or recover spatial information about the current state. Recent work shows that pretrained representations can support such tasks (Qu et al., 2026). Their usefulness also depends on how the process is observed. Sensors and simulations may record the same dynamics at different time intervals, so a model must contend with changes in the observation schedule as well as changes in the physical system.

We ask whether temporal relationships within existing recordings can make physical information more accessible when the sampling interval changes. An inference head is trained on representations of a fixed-length observation history collected at one interval and evaluated unchanged on histories collected at other intervals. The number of observations remains fixed, but their spacing and total temporal span change. This protocol tests whether information learned at one observation schedule remains usable at another. Figure 2 shows the consequence of this mismatch for a spatiotemporal joint-embedding predictive architecture (ST-JEPA) trained at one interval: its parameter-estimation error grows steadily as the interval widens. Figure 1 illustrates how a cadence-shifted parameter estimate changes the resulting physical evolution.

Takens' embedding theorem provides an intuition for relating these histories (Takens, 1981). Under suitable assumptions, a sufficiently long measurement history can provide coordinates for the underlying state. If histories sampled at two different intervals both yield valid embeddings, they describe the same state through different coordinate systems; their coordinates need not be numerically equal. Histories that look backward from a common endpoint therefore suggest a natural correspondence.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

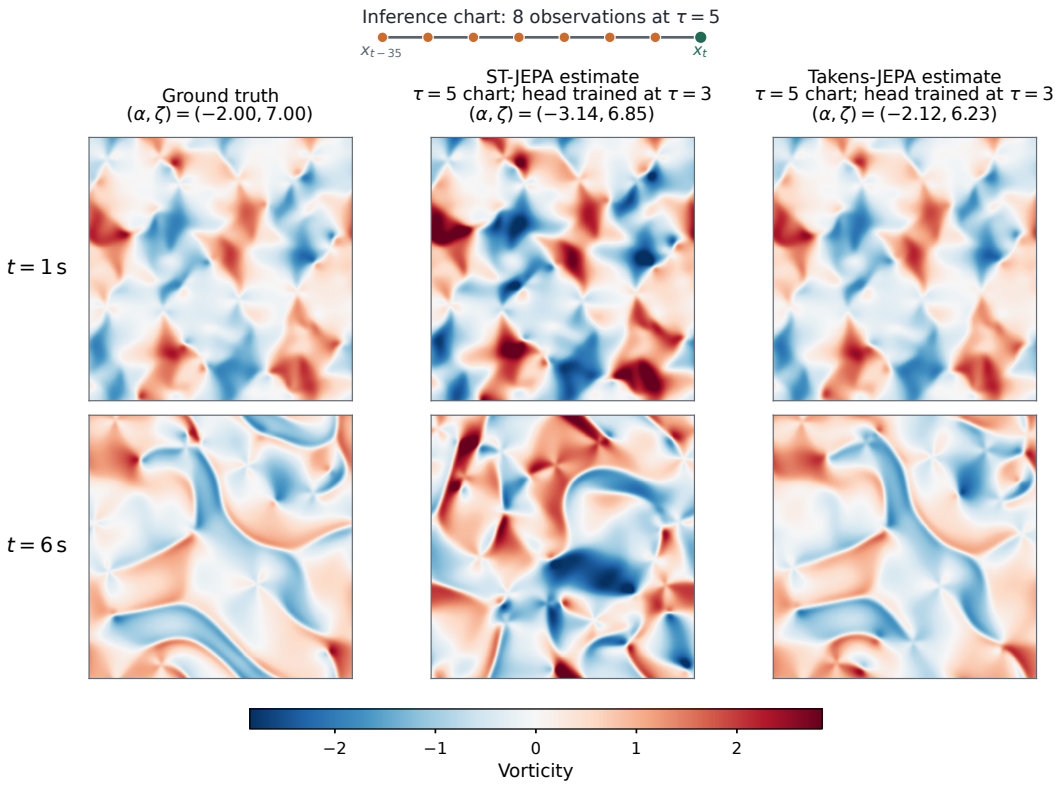

Figure 1: Active Matter simulations conditioned on parameters inferred from a test delay chart at $\tau = 5$ by heads trained at $\tau = 3$. Columns show the stored ground-truth trajectory and simulations using the ST-JEPA and Takens-JEPA estimates. The inset shows the eight observations in the cadence-shifted input chart. Rows show vorticity at $t = 1$ and $t = 6$ seconds on a shared color scale. The chart ends at saved-frame index 65 ($t = 16.25$ s); each simulation is integrated from the trajectory's initial state with the inferred parameters, so the displayed times are measured from that initial state. All simulations use the same matched initialization; the representation models estimate parameters rather than directly generating the fields. Additional intermediate snapshots appear in Appendix G.

For example, $[x_t, x_{t-3}, \ldots, x_{t-21}]$ and $[x_t, x_{t-6}, \ldots, x_{t-42}]$ contain different observations of the trajectory but refer to the same endpoint state.

Takens-JEPA turns this correspondence into an auxiliary training objective for ST-JEPA. A cadence-conditioned projector and a margin-ranking loss associate histories sharing an endpoint relative to non-corresponding histories drawn from other batch entries. The objective neither requires identical encoder outputs nor learns an invertible transformation between histories. Takens' theorem motivates the construction, but does not guarantee that our finite physical-field observations or learned representations are embeddings. Additional histories are subsampled from stored observations, and the projector is discarded after pretraining.

We evaluate physical-parameter estimation and endpoint-state reconstruction on three systems from The Well (Ohana et al., 2024): Active Matter, Rayleigh-Bénard convection, and Shear Flow. The encoder is frozen, and each downstream head is trained at one cadence and transferred without adaptation. The strongest improvements over ST-JEPA occur on Active Matter across both tasks and on Shear Flow parameter estimation under cadence shift.

We also compare against a multi-cadence control that applies the complete predictive objective independently at each cadence. It improves substantially over single-cadence training, but Takens-JEPA makes more efficient use of the same histories: on Active Matter it is more accurate at the training cadence and the nearest unseen cadence, matches the control at the coarser cadences, and needs

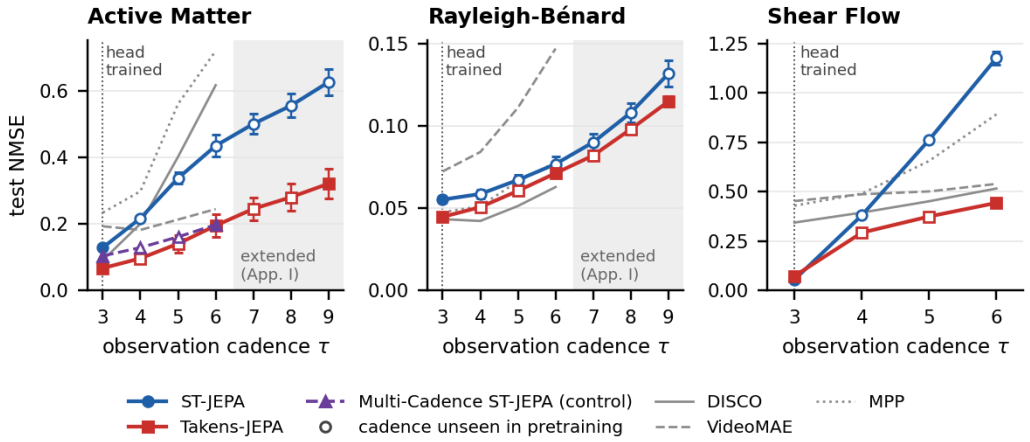

Figure 2: Physical-parameter estimation across observation cadences. Every head is trained at $\tau = 3$ (dotted line) and evaluated unchanged; hollow markers mark cadences a method never saw during pretraining, and the shaded band is the extended evaluation of Appendix I. Standard errors are shown for ST-JEPA and Takens-JEPA; Multi-Cadence ST-JEPA is also averaged over three runs. Lower NMSE is better.

about one-third fewer training FLOPs. These results show that explicitly relating same-endpoint histories can be more effective than treating each cadence as an independent data augmentation with its own predictive objective.

Beyond these benchmarks, cadence-robust inference would let one trained model serve instruments with different sampling limits and simulation outputs stored at different intervals, reducing the need for cadence-specific models or re-recorded data. Extending the principle to irregular timestamps and missing measurements would broaden its relevance further.

2 Method

2.1 Temporal Observation Charts

Let s_i be the physical state at saved-frame index i, and let $x_i = h(s_i) \in \mathbb{R}^S$ be the observation, with S observed scalar degrees of freedom. A delay chart contains d observations separated by an integer stride $\tau > 0$:

$$X_i^\tau = [x_i, x_{i-\tau}, \ldots, x_{i-(d-1)\tau}] \in \mathbb{R}^{S\times d}. \tag{1}$$

Here, cadence refers to the integer stride between saved observations. Different cadences can therefore be constructed by reindexing an existing trajectory whenever the required frames are available; no additional simulation output is required. Under the backward-delay convention in Equation (1), each chart represents the state at its endpoint. Consequently, $X_i^{\tau_r}$ and $X_i^{\tau_s}$ contain different histories but refer to the same endpoint state s_i. This endpoint alignment is important: aligning their starting or intermediate frames would generally produce charts representing different endpoint states. Appendix F examines this choice empirically.

2.2 Predictive Pretraining

We use the encoder and predictor of Qu et al. (2026). Given the primary chart $z_i^{\tau_0} = E_\theta(X_i^{\tau_0})$, the predictor estimates the representation of a later chart:

$$\widehat{z}_{i+\Delta}^{\tau_0} = P_\psi(z_i^{\tau_0}), \qquad z_{i+\Delta}^{\tau_0} = E_\theta(X_{i+\Delta}^{\tau_0}). \tag{2}$$

The displacement Δ specifies the predictive horizon in saved-frame indices. In Takens-JEPA, this predictive objective is evaluated only at the primary cadence τ_0. The additional cadences are used by the endpoint-aligned objective described below and do not receive separate predictive losses. The

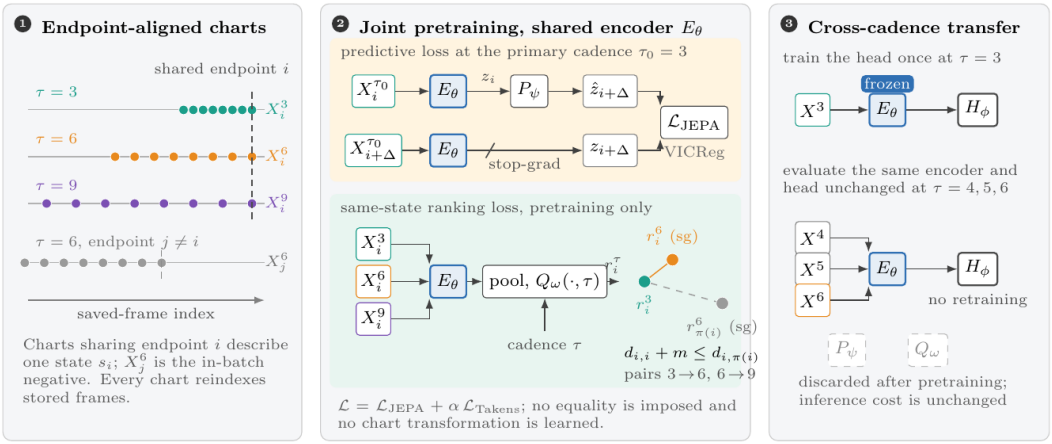

Figure 3: Overview of Takens-JEPA. Endpoint-aligned charts provide the auxiliary ranking loss during pretraining. The predictor and projector are then discarded, and the frozen encoder and downstream head are reused unchanged across cadences.

VICReg-style objective is

$$\mathcal{L}_{\text{JEPA},b} = \lambda_{\text{repr}} \frac{1}{B} \sum_{i \in b} \|\widehat{z}_{i+\Delta}^{\tau_0} - z_{i+\Delta}^{\tau_0}\|_2^2 + \lambda_{\text{var}} \mathcal{L}_{\text{var}} + \lambda_{\text{cov}} \mathcal{L}_{\text{cov}}, \tag{3}$$

where b is a minibatch of B examples. The representation term matches predicted and target features, while the variance and covariance terms discourage collapsed and redundant representations (Bardes et al., 2022). This objective learns predictive structure but does not explicitly relate histories that share an endpoint.

2.3 Endpoint-Aligned Ranking Objective

Takens motivation. Consider a compact n-dimensional manifold M, a C^2 diffeomorphism $F : M \to M$, and a C^2 scalar observation $g : M \to \mathbb{R}$. At stride τ, define

$$\Phi_{\tau,d}(s) = [g(s), g(F^{-\tau}s), \ldots, g(F^{-(d-1)\tau}s)]. \tag{4}$$

For a generic pair of map and observation satisfying the classical assumptions, $d \geq 2n + 1$ suffices for an embedding (Takens, 1981). Applying this conclusion at a particular stride requires the corresponding stride map and observation to satisfy those assumptions. If $\Phi_{\tau_r,d}$ and $\Phi_{\tau_s,d}$ are both embeddings of the same state manifold, then

$$\varphi_{r\to s} : \Phi_{\tau_r,d}(M) \to \Phi_{\tau_s,d}(M), \qquad \varphi_{r\to s} = \Phi_{\tau_s,d} \circ \Phi_{\tau_r,d}^{-1}. \tag{5}$$

Thus, valid delay charts provide different coordinate descriptions of the same underlying state.

Applying this result directly would require learning an invertible nonlinear transformation between chart spaces, potentially for every cadence pair. This is difficult to scale and introduces assumptions that cannot be verified from finite physical trajectories. Several tractable approximations are possible. An equality objective directly aligns same-endpoint chart representations, but imposes a stronger relation than the theorem requires. A Lie formulation models chart changes through a shared continuous latent generator, while a Koopman formulation uses a shared linear operator. Both replace the unknown diffeomorphism with a more restrictive form of latent transport.

We adopt a weaker relational surrogate. Valid embeddings imply that different charts of the same state correspond across coordinate systems, while distinct states remain distinguishable because each chart map is injective. Our ranking objective preserves these correspondence and separation relations: charts sharing an endpoint should be more closely associated than charts drawn from different endpoint entries. It does not estimate $\varphi_{r\to s}$ or enforce its smoothness or invertibility. Appendix F compares this objective with the equality, Lie, and Koopman alternatives.

Our physical observations are finite, high-dimensional, and potentially noisy, and we do not verify the genericity or embedding-dimension assumptions of the classical theorem. We therefore use Takens' theorem as a design principle rather than claim that the learned representations satisfy its guarantees.

Comparison space. We pool encoder features as needed and apply a shared cadence-conditioned projector:

$$r_i^\tau = \mathrm{norm}\big(Q_\omega(\mathrm{pool}(z_i^\tau), \tau)\big), \tag{6}$$

where norm denotes ℓ_2 normalization. The comparison is imposed in this auxiliary space, allowing the encoder to retain cadence-specific information. Within a minibatch, indices i and k denote sampled trajectory and endpoint entries. For a directed cadence relation (τ_r, τ_s), define

$$d_{i,k}^{(r,s)} = \|r_i^{\tau_r} - \mathrm{sg}(r_k^{\tau_s})\|_2^2, \tag{7}$$

where sg stops gradients through the comparison representation. A one-position cyclic permutation π supplies a non-corresponding batch entry, with $\pi(i) \neq i$. The pairwise objective is

$$\mathcal{L}_{\mathrm{Takens}}^{(r,s)} = \frac{1}{B}\sum_{i=1}^{B}\Big[m + d_{i,i}^{(r,s)} - d_{i,\pi(i)}^{(r,s)}\Big]_+, \qquad [u]_+ = \max(0,u), \tag{8}$$

where $m > 0$ is the ranking margin. The first distance compares charts with the same endpoint, whereas the second compares the anchor with a permuted batch entry. The objective asks the same-endpoint pair to be closer by at least m without requiring the two representations to be numerically identical. Because the comparison representation is stop-gradient, each (τ_r, τ_s) is directed: τ_r receives the anchor-side gradient, while τ_s is stopped for this relation.

Role of negatives. A constant projected representation gives equal same-endpoint and permuted distances and therefore incurs loss m. It cannot achieve zero ranking loss, although this alone does not guarantee that optimization avoids collapse. The VICReg terms regularize predictive features rather than the auxiliary projector directly. Moreover, batch permutation supplies non-corresponding endpoint entries rather than certifying distinct physical states: two entries may share a trajectory or parameter setting, represent nearby states, or, in principle, revisit the same state.

Combined objective. Let $\mathcal{T}$ be the selected cadences and $\mathcal{P} \subseteq \{(\tau_r, \tau_s) \in \mathcal{T}^2 : \tau_r \neq \tau_s\}$ be the selected directed relations. The auxiliary and complete pretraining objectives are

$$\mathcal{L}_{\mathrm{Takens},b} = \frac{1}{|\mathcal{P}|}\sum_{(\tau_r,\tau_s)\in\mathcal{P}} \mathcal{L}_{\mathrm{Takens}}^{(r,s)}, \tag{9}$$

$$\mathcal{L}_b = \mathcal{L}_{\mathrm{JEPA},b} + \alpha\mathcal{L}_{\mathrm{Takens},b}. \tag{10}$$

Here $\alpha \geq 0$ is the chart-loss weight. Relations need not form a cycle. After pretraining, the encoder supplies the representation; the predictor and projector are discarded. Figure 3 summarizes the configuration, and Appendix H gives pseudocode.

3 Experimental Protocol

Datasets and baselines. We use Active Matter, Rayleigh-Bénard convection, and Shear Flow from The Well (Ohana et al., 2024). We compare Takens-JEPA with ST-JEPA (Qu et al., 2026), VideoMAE (Tong et al., 2022), DISCO (Morel et al., 2025a), and Multiple Physics Pretraining (MPP) (McCabe et al., 2024). Dataset descriptions and model details appear in Appendices B and C.

Shared data and downstream training. Within each dataset and task, all methods use the same trajectory splits, observed fields, 224×224 spatial resolution, primary eight-frame charts, and eligible endpoints. The earliest eligible endpoint is determined by the widest pretraining chart:

$$(d-1)\max \mathcal{T}_{\mathrm{pre}} = (8-1)\times 9 = 63. \tag{11}$$

Restricting every method to this common endpoint range ensures that differences cannot be attributed to the number or temporal location of available training windows. Each method constructs the masks, targets, or additional views required by its pretraining objective from these stored trajectories. Takens-JEPA additionally encodes endpoint-aligned auxiliary charts. Data availability is therefore controlled, although pretraining compute differs across objectives.

ST-JEPA and Takens-JEPA share the same encoder, predictor, primary predictive objective, and downstream heads; their only intended methodological difference is the endpoint-aligned auxiliary objective. For every method, the pretrained representation-producing backbone is frozen, and a separate task-specific head is fitted for each downstream task. The same pretrained backbone is used for both tasks.

Cadence transfer. Downstream training uses $d = 8$ observations with stride $\tau = 3$. The encoder and fitted head are then evaluated unchanged at

$$\mathcal{T}_{\text{test}} = \{3, 4, 5, 6\}. \tag{12}$$

Takens-JEPA constructs pretraining charts at

$$\mathcal{T}_{\text{pre}} = \{3, 6, 9\}, \qquad \mathcal{P} = \{(3, 6), (6, 9)\}. \tag{13}$$

The relation set $\mathcal{P}$ was chosen on Active Matter parameter estimation using the evaluation cadences (Appendix D) and then held fixed for the other two systems, the endpoint-state task, and the extended cadences of Appendix I, none of which informed the choice. ST-JEPA applies its predictive objective only at the primary cadence $\tau = 3$ and does not encode the auxiliary charts. Takens-JEPA uses the same primary predictive objective and additionally evaluates $\mathcal{L}_{\text{Takens}}$ over the directed relations in $\mathcal{P}$. As a direct data-augmentation control, Multi-Cadence ST-JEPA applies the complete predictive objective independently at every cadence in $\mathcal{T}_{\text{pre}}$, without the endpoint-aligned ranking loss. This comparison separates the effect of relating same-endpoint charts from the effect of adding predictive training examples at multiple cadences.

For each endpoint, the auxiliary strides 6 and 9 can be obtained by subsampling a longer history on the same stride-3 grid; they are not contained within the eight-frame primary chart alone. Evaluations at $\tau = 4, 5$ use the benchmark's stored frames to test unseen intermediate cadences. Evaluation at $\tau = 6$ shifts the downstream input cadence but uses a cadence seen by the auxiliary pretraining objective. Appendix I extends the evaluation to $\tau = 7, 8, 9$; we do not evaluate beyond the widest pretraining cadence.

Observation scope. At the dataset and downstream-task level, every retained timestamp contains all physical field channels, and we do not remove sensor locations or physical variables. Model-specific corruption used internally by a pretraining objective, such as VideoMAE masking, does not change this observation protocol. Governing parameters are withheld from every encoder. Partial observability in our experiments therefore arises from temporal subsampling, while parameter estimation targets an unobserved system property.

Tasks. Let n index a trajectory, let i denote an eligible endpoint, and define the frozen representation

$$z^{\tau}_{n,i} = E_\theta(X^{\tau}_{n,i}). \tag{14}$$

For *physical-parameter estimation*, let p_n be the vector of governing parameters associated with trajectory n. Every chart sampled from that trajectory inherits the same target, and an attentive regression head predicts

$$\widehat{p}^{\tau}_{n,i} = H^{\text{param}}_{\phi}(z^{\tau}_{n,i}), \qquad y_{n,i} = p_n. \tag{15}$$

For *endpoint-state reconstruction*, a spatial decoder maps the frozen representation back to the physical field at the chart endpoint:

$$\widehat{x}^{\tau}_{n,i} = H^{\text{state}}_{\phi}(z^{\tau}_{n,i}), \qquad y_{n,i} = x_{n,i}. \tag{16}$$

Although $x_{n,i}$ is included in the input history, the decoder receives only the frozen representation. This task therefore measures how much spatial information about the endpoint survives the representation bottleneck and remains accessible when the observation cadence changes; it does not measure recovery of an unobserved physical field. Each downstream head is fitted using only $\tau = 3$ charts. The encoder and fitted head are then evaluated unchanged at every cadence in $\mathcal{T}_{\text{test}}$.

Metrics and variation. Target channels are standardized using statistics computed from the downstream training set. We report mean-squared error in this normalized space (NMSE), averaged over target parameters or over the channels and spatial positions of the reconstructed field; lower is better. We report absolute error at every cadence rather than degradation relative to $\tau = 3$ alone, because limited degradation is useful only when the representation is also accurate at the shifted cadence. ST-JEPA, Takens-JEPA, and Multi-Cadence ST-JEPA report means over three independent pretraining and downstream runs. Standard errors for ST-JEPA and Takens-JEPA appear in Appendix A; the remaining baseline and appendix-ablation entries are single runs.

Table 1: Physical-parameter estimation under cadence transfer. Each head is trained at $\tau = 3$ and evaluated unchanged. Entries are test NMSE; bold marks the lowest value in each dataset and cadence column. ST-JEPA and Takens-JEPA entries are three-run means, with standard errors in Appendix A.

Method	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$
Active Matter				
ST-JEPA	0.1287	0.2166	0.3371	0.4344
Takens-JEPA	**0.0669**	**0.0963**	**0.1403**	**0.1947**
DISCO	0.0903	0.2004	0.4013	0.6162
VideoMAE	0.1926	0.1817	0.2128	0.2442
MPP	0.2336	0.2978	0.5598	0.7192
Rayleigh-Bénard				
ST-JEPA	0.0552	0.0586	0.0671	0.0768
Takens-JEPA	0.0447	0.0506	0.0605	0.0712
DISCO	**0.0433**	**0.0422**	**0.0513**	**0.0628**
VideoMAE	0.0723	0.0841	0.1110	0.1469
MPP	0.0475	0.0512	0.0663	0.0778
Shear Flow				
ST-JEPA	**0.0548**	0.3800	0.7621	1.1762
Takens-JEPA	0.0728	**0.2924**	**0.3743**	**0.4430**
DISCO	0.3435	0.3939	0.4513	0.5159
VideoMAE	0.4524	0.4867	0.5011	0.5393
MPP	0.4288	0.4899	0.6554	0.8913

Table 2: Endpoint-state reconstruction from frozen representations. Targets are the spatial fields at the observed endpoint; decoders are trained at $\tau = 3$. Entries are test NMSE averaged over field channels and spatial positions. Bold marks the lowest value, including ties at the displayed precision. ST-JEPA and Takens-JEPA entries are three-run means.

	Active Matter				Rayleigh-Bénard				Shear Flow			
Method	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$
ST-JEPA	0.2215	0.2437	0.2730	0.3050	0.0252	0.0274	0.0304	0.0335	**0.0028**	0.0041	0.0050	0.0059
Takens-JEPA	**0.1500**	**0.1630**	**0.1826**	**0.2059**	0.0204	0.0224	0.0250	0.0279	**0.0028**	**0.0040**	**0.0049**	**0.0057**
VideoMAE	0.3027	0.3389	0.3724	0.4047	**0.0181**	**0.0203**	**0.0231**	**0.0258**	0.0069	0.0099	0.0131	0.0157
DISCO	0.4925	0.5784	0.6702	0.7654	0.0791	0.0905	0.1030	0.1151	0.0079	0.0190	0.0257	0.0303
MPP	0.4212	0.5313	0.6194	0.6999	0.0682	0.0785	0.0933	0.1082	0.0102	0.0177	0.0203	0.0245

4 RESULTS

4.1 PHYSICAL-PARAMETER ESTIMATION

Table 1 shows lower mean NMSE for Takens-JEPA than ST-JEPA at every shifted cadence. On Active Matter, NMSE falls by approximately 48% at $\tau = 3$ and by 55 to 58% at $\tau = 4, 5, 6$. The differences in means are large relative to the reported standard errors (Table 4).

On Shear Flow, Takens-JEPA has higher error at $\tau = 3$ (0.0728 versus 0.0548), but reduces mean error by approximately 23%, 51%, and 62% at $\tau = 4, 5, 6$. Better shifted-cadence accuracy therefore need not coincide with the best accuracy at the training cadence. Rayleigh-Bénard shows smaller improvements over ST-JEPA, and DISCO has the lowest error at each cadence. The strongest representation method therefore varies by system.

4.2 ENDPOINT-STATE RECONSTRUCTION

Table 2 evaluates spatial information preserved by the frozen representation. The strongest result is again Active Matter: Takens-JEPA reduces mean NMSE by approximately 32% at both $\tau = 3$ and $\tau = 6$. Its error rises from 0.1500 to 0.2059, compared with 0.2215 to 0.3050 for ST-JEPA. These improvements concern decodability through the bottleneck, since the target frame is observed.

Table 3: Active Matter parameter estimation with multi-cadence pretraining. The head is trained at $\tau = 3$. Entries are three-run mean test NMSE.

Method	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$
ST-JEPA	0.1287	0.2166	0.3371	0.4344
Multi-Cadence ST-JEPA	0.1038	0.1286	0.1608	0.1973
Takens-JEPA	**0.0669**	**0.0963**	**0.1403**	**0.1947**

On Rayleigh-Bénard, Takens-JEPA improves over ST-JEPA, but VideoMAE has the lowest absolute error. On Shear Flow, ST-JEPA and Takens-JEPA are tied at the displayed precision at $\tau = 3$, and their differences under shift are small. The seed summaries in Table 5 support the large Active Matter improvement but do not establish a uniformly substantial advantage across all tasks and systems.

4.3 Ablation: Multi-Cadence Predictive Training

Takens-JEPA differs from ST-JEPA in two related ways: it exposes the encoder to auxiliary cadence charts and relates charts sharing the same endpoint. Consequently, its improvement might arise simply from training on multiple cadences rather than from endpoint-aligned supervision. We test this possibility with Multi-Cadence ST-JEPA, a direct data-augmentation control that receives the same Active Matter trajectories, eligible endpoints, and pretraining cadences as Takens-JEPA.

Instead of applying the chart-ranking objective, Multi-Cadence ST-JEPA evaluates the complete predictive loss independently at every cadence in $\mathcal{T}_{\mathrm{pre}} = \{3, 6, 9\}$:

$$\mathcal{L}_{\text{MC-JEPA},b} = \frac{1}{|\mathcal{T}_{\mathrm{pre}}|} \sum_{\tau \in \mathcal{T}_{\mathrm{pre}}} \mathcal{L}^{(\tau)}_{\text{JEPA},b}. \tag{17}$$

Its encoder is frozen after pretraining, and its downstream head is trained only at $\tau = 3$ and evaluated using the same cadence-transfer protocol. This ablation tests whether relating same-endpoint charts is more effective than treating each auxiliary cadence as an independent predictive training example.

Multi-Cadence ST-JEPA improves substantially over ST-JEPA at every evaluated cadence (Table 3), confirming that direct predictive training at multiple cadences is itself a strong baseline. At $\tau = 6$, it obtains an NMSE of 0.1973, close to the 0.1947 achieved by Takens-JEPA. This similarity is expected because $\tau = 6$ receives direct predictive supervision during multi-cadence pretraining.

Takens-JEPA has lower mean error at the training cadence and at both unseen intermediate cadences: 0.0669, 0.0963, and 0.1403 at $\tau = 3, 4, 5$ against 0.1038, 0.1286, and 0.1608. The margins at $\tau = 3$ and $\tau = 4$ exceed the run-to-run variation of both methods, whereas the margin at $\tau = 5$ lies within it. Relating endpoint-aligned histories therefore adds to cadence exposure, most clearly at and near the training cadence; this control covers Active Matter parameter estimation only.

The two approaches also allocate training computation differently. Takens-JEPA applies one primary predictive objective and compares auxiliary charts through the lighter ranking objective, whereas Multi-Cadence ST-JEPA performs complete predictive training at all three cadences. In the measured Active Matter configuration, Takens-JEPA uses approximately $2\times$ the training FLOPs of ST-JEPA, while Multi-Cadence ST-JEPA uses approximately $3\times$ (Appendix E). All three methods retain the same inference architecture and cost.

5 Related Work

Physical representations. Physical learning systems often optimize forward prediction through autoregressive pretraining, solution operators, or trajectory-conditioned evolution operators (McCabe et al., 2024; Herde et al., 2024; Morel et al., 2025a). Self-supervised approaches learn through masked reconstruction or latent prediction (Tong et al., 2022; Assran et al., 2023; Bardes et al., 2024). Qu et al. (2026) evaluate these representations for parameter estimation and provide our ST-JEPA backbone. We extend this evaluation to temporal observation shift by freezing the pretrained encoder, fitting an inference head at one cadence, and evaluating both components unchanged at other cadences.

Temporal and multi-view learning. Our auxiliary objective belongs to margin-based metric learning (Schroff et al., 2015). Time-Contrastive Networks associate simultaneous observations from different viewpoints (Sermanet et al., 2018), and TS2Vec contrasts augmented temporal contexts to learn timestamp representations (Yue et al., 2022). Sampling temporally separated clips as positive views is also common in self-supervised video learning (Qian et al., 2021; Feichtenhofer et al., 2021); here, by contrast, positives share an endpoint across different strides, the projector is conditioned on the stride, and the head is transferred across strides. Our contribution concerns endpoint-aligned histories with different sampling intervals, their combination with predictive pretraining, and the evaluation of scientific inference across cadences unavailable during downstream training.

Memory and delay reconstruction. Temporal history has been used to compensate for information unavailable in an instantaneous observation. Memory-based physical models primarily use this history to improve future prediction from unresolved or partially observed states, including settings involving spatial coarse-graining (Buitrago Ruiz et al., 2025; Morel et al., 2025b). Delay-coordinate reconstruction instead establishes conditions under which an observation history can identify the underlying current state (Takens, 1981; Sauer et al., 1991).

These directions do not directly address whether a pretrained physical representation remains usable by the same downstream model when the temporal sampling cadence changes. We bridge this gap by incorporating same-endpoint correspondence across delay charts into predictive representation learning and evaluating the resulting frozen representations across both seen and unseen cadences. Takens' theorem motivates this correspondence, but we do not claim that the ranking objective establishes an embedding or recovers the coordinate transformations between valid delay maps.

6 Conclusion and Future Work

This work asked whether physical representations learned at one temporal cadence remain useful when the same dynamics are sampled at another. Takens-JEPA relates histories that contain different observations but end at the same physical state, and the clearest improvements occur on Active Matter across both tasks and on Shear Flow parameter estimation under cadence shift: endpoint-aligned supervision reduces reliance on cadence-specific temporal patterns while preserving information needed for scientific inference.

The multi-cadence control confirms that predictive training at several cadences is itself beneficial. Takens-JEPA relates the same auxiliary histories through a theory-motivated objective instead, and on Active Matter it is more accurate at the training cadence and the nearest unseen cadence at about one-third less training cost, so how the views are related during pretraining matters beyond exposure to them.

Several limitations define the scope of this conclusion. We do not verify the smoothness, genericity, and embedding-dimension assumptions of Takens' theorem for finite, high-dimensional trajectories, and the ranking objective preserves only the correspondence and separation relations the theorem motivates; it neither estimates nor guarantees an invertible transformation between charts. The experiments use regular integer cadences, synchronized endpoints, complete field channels, and cadences within the pretraining range. The relation set was selected on Active Matter test cadences before being applied unchanged to the other systems and tasks, the multi-cadence control covers Active Matter parameter estimation only, and no single-cadence contrastive control with the same loss and compute is included, so the contribution of endpoint alignment relative to a generic contrastive term remains to be isolated. Appendix J shows that the same-state ordering is learned in the discarded projector and only weakly in the retained encoder; the demonstrated result is downstream transfer, not a metric on encoder features.

Future work should extend same-state chart learning to irregular timestamps, asynchronous sensors, missing channels, and cadences beyond the pretraining range, and should examine how the objective interacts with other backbones and when relational regularization, direct multi-cadence prediction, or their combination gives the best accuracy for the compute spent.

AI USE STATEMENT

AI tools assisted with manuscript editing, figure layout, and feedback on clarity, consistency, and experimental controls. Responsibility for the scientific claims and the final manuscript rests with the authors.

REPRODUCIBILITY STATEMENT

Appendix B describes the datasets and splits; Appendix C gives the architectures, pretraining and downstream hyperparameters, and the data protocol shared by all methods; Appendix A reports variation across three runs; and Appendix H specifies the ranking-loss computation.

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*, 2023.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024. URL https://arxiv.org/abs/2404.08471.

Ricardo Buitrago Ruiz, Tanya Marwah, Albert Gu, and Andrej Risteski. On the benefits of memory for modeling time-dependent PDEs. In *International Conference on Learning Representations*, 2025.

Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

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*, 2024.

Suryanarayana Maddu, Scott Weady, and Michael J. Shelley. Learning fast, accurate, and stable closures of a kinetic theory of an active fluid. *Journal of Computational Physics*, 504:112869, 2024. doi: 10.1016/j.jcp.2024.112869.

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*, 2024.

Rudy Morel, Jiequn Han, and Edouard Oyallon. DISCO: Learning to discover an evolution operator for multi-physics-agnostic prediction. In *Proceedings of the International Conference on Machine Learning*, 2025a.

Rudy Morel, Francesco Pio Ramunno, Jeff Shen, Alberto Bietti, Kyunghyun Cho, Miles Cranmer, Siavash Golkar, Olexandr Gugnin, Geraud Krawezik, Tanya Marwah, et al. Predicting partially observable dynamical systems via diffusion models with a multiscale inference scheme. In *Advances in Neural Information Processing Systems*, 2025b.

Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina J. Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Stuart B. Dalziel, Drummond B. Fielding, et al. The well: A large-scale collection of diverse physics simulations for machine learning. In *Advances in Neural Information Processing Systems*, 2024.

Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Helen Qu, Rudy Morel, Michael McCabe, Alberto Bietti, François Lanusse, Shirley Ho, and Yann LeCun. Representation learning for spatiotemporal physical systems. *arXiv preprint arXiv:2603.13227*, 2026. URL https://arxiv.org/abs/2603.13227.

Tim Sauer, James A. Yorke, and Martin Casdagli. Embedology. *Journal of Statistical Physics*, 65 (3–4):579–616, 1991. doi: 10.1007/BF01053745.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, and Sergey Levine. Time-contrastive networks: Self-supervised learning from video. In *IEEE International Conference on Robotics and Automation*, 2018.

Floris Takens. Detecting strange attractors in turbulence. In David Rand and Lai-Sang Young (eds.), *Dynamical Systems and Turbulence, Warwick 1980*, volume 898 of *Lecture Notes in Mathematics*, pp. 366–381. Springer, 1981. doi: 10.1007/BFb0091924.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. TS2Vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Table 4: Physical parameter estimation across three runs. Entries are mean test NMSE ± standard error; lower is better.

Dataset	Method	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$
Active Matter	ST-JEPA	0.1287 ± 0.0068	0.2166 ± 0.0068	0.3371 ± 0.0172	0.4344 ± 0.0331
	Takens-JEPA	0.0669 ± 0.0031	0.0963 ± 0.0126	0.1403 ± 0.0256	0.1947 ± 0.0350
Rayleigh-Bénard	ST-JEPA	0.0552 ± 0.0021	0.0586 ± 0.0027	0.0671 ± 0.0032	0.0768 ± 0.0044
	Takens-JEPA	0.0447 ± 0.0018	0.0506 ± 0.0010	0.0605 ± 0.0007	0.0712 ± 0.0005
Shear Flow	ST-JEPA	0.0548 ± 0.0016	0.3800 ± 0.0065	0.7621 ± 0.0098	1.1762 ± 0.0334
	Takens-JEPA	0.0728 ± 0.0006	0.2924 ± 0.0056	0.3743 ± 0.0034	0.4430 ± 0.0060

A Variation Across Random Seeds

Tables 4 and 5 report means and standard errors across three independent pretraining and downstream runs for ST-JEPA and Takens-JEPA. With sample standard deviation s and $n = 3$, the standard error is $s/\sqrt{3}$.

For parameter estimation, the clearest separations occur on Active Matter and under cadence shift on Shear Flow. ST-JEPA has lower mean error at the Shear Flow training cadence. Rayleigh-Bénard differences are smaller, especially at the wider strides. The largest standard error in this table is 0.0350 for Takens-JEPA on Active Matter at $\tau = 6$.

Table 5: Endpoint-state reconstruction across three aligned runs. Entries are mean test NMSE ± standard error; lower is better.

Dataset	Method	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$
Active Matter	ST-JEPA	0.2215 ± 0.0053	0.2437 ± 0.0076	0.2730 ± 0.0092	0.3050 ± 0.0111
	Takens-JEPA	0.1500 ± 0.0012	0.1630 ± 0.0019	0.1826 ± 0.0023	0.2059 ± 0.0024
Rayleigh-Bénard	ST-JEPA	0.0252 ± 0.0005	0.0274 ± 0.0005	0.0304 ± 0.0006	0.0335 ± 0.0006
	Takens-JEPA	0.0204 ± 0.0007	0.0224 ± 0.0009	0.0250 ± 0.0013	0.0279 ± 0.0016
Shear Flow	ST-JEPA	0.0028 ± 0.0001	0.0041 ± 0.0001	0.0050 ± 0.0001	0.0059 ± 0.0001
	Takens-JEPA	0.0028 ± 0.0001	0.0040 ± 0.0001	0.0049 ± 0.0001	0.0057 ± 0.0001

For endpoint reconstruction, Active Matter shows the largest absolute improvement. On Rayleigh-Bénard, the differences remain favorable to Takens-JEPA but are smaller; on Shear Flow the methods are close, including a tie at $\tau = 3$ at the displayed precision. Three-run mean differences alone should not be interpreted as a universal robustness guarantee.

B Datasets

The three datasets from The Well (Ohana et al., 2024) provide nonlinear physical trajectories, multiple observed fields, and trajectory-level governing parameters. Table 6 lists the stored resolutions and targets; model inputs use the resolution reported in Appendix C.

Table 6: Summary of the three datasets. The listed parameters are the targets of the physical-parameter inference task.

Dataset	Stored trajectory	Physical fields	Inferred parameters
Active Matter	81 frames, 256×256	Concentration, velocity, orientation, strain rate	α, ζ
Rayleigh-Bénard	200 frames, 512×128	Buoyancy, pressure, velocity	Rayleigh, Prandtl
Shear Flow	200 frames, 256×512	Tracer, pressure, velocity	Reynolds, Schmidt

B.1 Active Matter

Active Matter models a continuum of rod-like active particles in a two-dimensional Stokes fluid (Maddu et al., 2024). Observed fields include concentration, two velocity components, an orientation

tensor, and a strain-rate tensor. The dynamics couple particle alignment and fluid motion, including transitions between isotropic and nematic order. Each trajectory has 81 saved frames on a uniform 256×256 grid over the periodic domain $[0, 10] \times [0, 10]$, stored every 0.25 seconds over $t \in [0, 20]$.

The simulation parameters are

$$\alpha \in \{-1, -2, -3, -4, -5\}, \quad \beta = 0.8, \quad \zeta \in \{1, 3, 5, 7, 9, 11, 13, 15, 17\}. \tag{18}$$

Here α is active dipole strength and ζ controls alignment. This physical α is distinct from the chart-loss weight in the training objective. Since β is fixed, inference targets only (α, ζ). Five initializations for each of 45 parameter combinations give 225 trajectories.

B.2 Rayleigh-Bénard Convection

Rayleigh-Bénard convection describes a two-dimensional fluid heated from below and cooled from above. Buoyancy, pressure, and two velocity components evolve according to

$$\partial_t b - \kappa \Delta b = -\mathbf{u} \cdot \nabla b, \tag{19}$$

$$\partial_t \mathbf{u} - \nu \Delta \mathbf{u} + \nabla p - b\mathbf{e}_y = -\mathbf{u} \cdot \nabla \mathbf{u}, \qquad \nabla \cdot \mathbf{u} = 0. \tag{20}$$

The thermal diffusivity and viscosity are

$$\kappa = (\mathrm{Ra}\,\mathrm{Pr})^{-1/2}, \qquad \nu = (\mathrm{Ra}/\mathrm{Pr})^{-1/2}. \tag{21}$$

Each trajectory has 200 saved frames on a 512×128 grid over $t \in [0, 50]$. The horizontal direction is periodic and uniformly sampled; the vertical direction uses Chebyshev nodes and Dirichlet conditions. The parameter sets are

$$\mathrm{Ra} \in \{10^6, 10^7, 10^8, 10^9, 10^{10}\}, \qquad \mathrm{Pr} \in \{0.1, 0.2, 0.5, 1, 2, 5, 10\}. \tag{22}$$

Initial conditions vary the buoyancy amplitude $\delta b_0 \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ and random seed. There are 50 initial conditions for each of 35 parameter combinations, giving 1,750 trajectories. Inference targets $(\mathrm{Ra}, \mathrm{Pr})$.

B.3 Shear Flow

Shear Flow models a periodic, incompressible two-dimensional fluid with an advected tracer. Its dynamics are

$$\partial_t \mathbf{u} + \nabla p - \nu \Delta \mathbf{u} = -\mathbf{u} \cdot \nabla \mathbf{u}, \qquad \nabla \cdot \mathbf{u} = 0, \tag{23}$$

$$\partial_t s - D \Delta s = -\mathbf{u} \cdot \nabla s, \tag{24}$$

where s is the tracer and

$$\nu = 1/\mathrm{Re}, \qquad D = \nu/\mathrm{Sc}. \tag{25}$$

The observed fields are tracer, pressure, and two velocity components. Each trajectory has 200 saved frames at spacing 0.1 on a uniform 256×512 grid over the periodic spatial domain $[0, 1] \times [-1, 1]$. The parameter sets are

$$\mathrm{Re} \in \{10^4, 5 \times 10^4, 10^5, 5 \times 10^5\}, \qquad \mathrm{Sc} \in \{0.1, 0.2, 0.5, 1, 2, 5, 10\}. \tag{26}$$

Initial conditions vary the number of shear layers $n_{\mathrm{shear}} \in \{2, 4\}$, perturbation modes $n_{\mathrm{blobs}} \in \{2, 3, 4, 5\}$, and shear width $w \in \{0.25, 0.5, 1, 2, 4\}$. The 40 initial conditions for each of 28 parameter combinations give 1,120 trajectories. Inference targets $(\mathrm{Re}, \mathrm{Sc})$.

C Baseline Architectures and Implementation Details

C.1 Architectures and Objectives

Table 7 summarizes the representation-producing backbones. Only these backbones are retained for downstream evaluation.

ST-JEPA. Following Qu et al. (2026), the encoder is a downsampling spatiotemporal convolutional network based on ConvNeXt blocks. A convolutional predictor with an inverse channel bottleneck

Table 7: Summary of the evaluated models. Only the representation-producing backbone is retained for downstream evaluation.

Method	Representation backbone	Pretraining objective
ST-JEPA	ConvNeXt-style spatiotemporal CNN	Predictive VICReg loss in latent space
Takens-JEPA	Same backbone as ST-JEPA	ST-JEPA loss plus same-state chart loss
VideoMAE	VideoMAE-Small/16	Masked pixel reconstruction
DISCO	CNN and axial-transformer hypernetwork	Normalized next-state prediction
MPP	Axial Vision Transformer	Multi-physics autoregressive prediction

estimates the representation of a later chart. Pretraining uses representation matching, variance, and covariance terms (Bardes et al., 2022).

Takens-JEPA. The encoder, predictor, and primary predictive objective are the same as ST-JEPA. A cadence-conditioned projector computes the auxiliary ranking loss in Section 2. The projector is discarded after pretraining, so the downstream encoder architecture is unchanged.

VideoMAE. We use the VideoMAE configuration in Qu et al. (2026), following Tong et al. (2022). A small Vision Transformer uses 16×16 spatial patches and temporal tube masking. A lightweight decoder reconstructs masked patches with a pixel-space mean-squared error loss. The reconstruction decoder is removed for downstream inference.

DISCO. A hypernetwork combines a convolutional tokenizer with axial attention over temporal and spatial dimensions. Its representation parameterizes a compact convolutional evolution operator that is integrated to predict the next physical state (Morel et al., 2025a). The model is trained end to end with its normalized next-state prediction objective. Downstream heads receive the hypernetwork representation.

MPP. Multiple Physics Pretraining uses an Axial Vision Transformer with field-specific embeddings and reversible normalization (McCabe et al., 2024). Attention is applied separately along temporal and spatial axes. The model is jointly pretrained across Active Matter, Rayleigh-Bénard, and Shear Flow with its multi-physics autoregressive objective; downstream heads use the transformer features.

C.2 Data Construction

Methods retain their respective architectures and pretraining objectives, while sharing benchmark data availability. Data construction and dataset-specific window strides follow the released ST-JEPA benchmark configuration.[1] Primary windows contain eight observations at $\tau = 3$. The widest auxiliary chart, at $\tau = 9$, requires an initial lag of $(8 - 1) \times 9 = 63$ saved frames. A common eligible endpoint range prevents baselines from receiving additional examples near the start of a trajectory. Auxiliary charts use the same trajectory and endpoint, but include historical frames outside the primary eight-frame window (Table 8). All methods use the benchmark's shared spatial preprocessing, normalization, and trajectory splits.

Table 8: Shared data construction used by every evaluated model. Window stride denotes the displacement between consecutive examples, not the cadence within an eight-observation chart.

Dataset	Resolution	Frames	Window stride	Test cadences
Active Matter	224×224	8	1	$3, 4, 5, 6$
Rayleigh-Bénard	224×224	8	4	$3, 4, 5, 6$
Shear Flow	224×224	8	4	$3, 4, 5, 6$

C.3 Pretraining Configuration

ST-JEPA and Takens-JEPA follow the released ST-JEPA configurations of Qu et al. (2026): 30 pretraining epochs, AdamW with a cosine schedule, VICReg coefficients $(\lambda_{\text{repr}}, \lambda_{\text{var}}, \lambda_{\text{cov}}) =$

[1] https://github.com/helenqu/physical-representation-learning

$(2, 40, 2)$, predictive horizon $\Delta = 3$ saved frames, and batch size 8. The encoder output is pooled by global average pooling over its 14×14 spatial grid to a 128-dimensional feature. Takens-JEPA adds the cadence-conditioned projector Q_ω, a two-layer MLP that concatenates the pooled feature with the scalar cadence and maps it to a 32-dimensional ℓ_2-normalized comparison space, with margin $m = 0.2$ and weight $\alpha = 0.1$ on all datasets. Pretraining uses $\mathcal{T}_{\text{pre}} = \{3, 6, 9\}$ and $\mathcal{P} = \{(3, 6), (6, 9)\}$, selected as described in Appendix D.

VideoMAE follows the benchmark configuration of Qu et al. (2026). DISCO is retrained separately for each dataset under the shared sparse-window protocol. MPP is jointly pretrained across all three datasets under the same data-availability protocol. Schedules are not tuned using cross-cadence test performance. Equal data availability does not imply equal architecture size or total training cost.

Relation to the benchmark of Qu et al. (2026). Our absolute values differ from theirs, and the ordering of DISCO and ST-JEPA on Shear Flow is reversed, because the protocols differ: their encoders receive 16 consecutive frames whereas ours receive eight frames at stride 3; every method here shares the common eligible endpoint range described above; MPP is evaluated with a frozen backbone rather than end-to-end fine-tuning; and DISCO is retrained on the sparse charts rather than evaluated from a released dense-cadence checkpoint.

C.4 Downstream Training

Each backbone is frozen. Parameter estimation uses an attentive pooling head; endpoint reconstruction uses a lightweight spatial-decoder family. Input projections accommodate each backbone's feature dimensions and token layout. Both head types are trained for 100 epochs using AdamW, learning rate 10^{-3}, weight decay 10^{-2}, two warm-up epochs, and cosine learning-rate decay. Training uses only $\tau = 3$ charts, followed by unchanged evaluation at $\tau \in \{3, 4, 5, 6\}$. Within each task, methods share trajectory splits, eligible endpoints, window strides, target normalization, and the model-selection criterion. The Rayleigh and Reynolds numbers are log-transformed before standardization. Model selection uses the best validation NMSE at $\tau = 3$.

D Training Chart-Pair Ablation

We compare four sets of directed relations on Active Matter:

$$\begin{aligned} &\mathcal{P}_1 = \{(3, 6)\}, && \mathcal{P}_2 = \{(3, 6), (3, 9)\}, \\ &\mathcal{P}_3 = \{(3, 6), (6, 9)\}, && \mathcal{P}_4 = \{(3, 6), (3, 9), (6, 9)\}. \end{aligned} \tag{27}$$

All configurations use eight-frame histories, the same encoder architecture, and chart-loss weight $\alpha = 0.1$. Heads are trained at $\tau = 3$ and evaluated unchanged at $\tau = 3, 4, 5, 6$. The final set includes all three increasing-stride relations; it does not include their reverse directions.

Each row of Table 9 is a single pretraining run. The $\mathcal{P}_3$ row is one run of the configuration whose three-run mean appears in Table 1, which is why its shifted-cadence values differ from the main table.

Table 9: Active Matter parameter estimation over directed relation sets, single run per configuration. The head is trained at $\tau = 3$; entries are test NMSE. The three-run mean for $\mathcal{P}_3$ is in Table 1.

Chart relations $\mathcal{P}$	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$
$\{(3, 6)\}$	0.0980	0.1334	0.2098	0.2910
$\{(3, 6), (3, 9)\}$	0.0849	0.1252	0.1777	0.2175
$\{(3, 6), (6, 9)\}$	**0.0669**	**0.0826**	**0.1198**	**0.1603**
$\{(3, 6), (3, 9), (6, 9)\}$	0.0674	0.1068	0.1512	0.2004

Adding a second relation improves over $\mathcal{P}_1$ at every cadence, and the chain $\mathcal{P}_3$ gives the lowest error at every cadence; adding the remaining relation does not improve further. We selected $\mathcal{P}_3$ for the main experiments on the basis of this ablation, which uses the evaluation cadences, and we report the single-pair configuration so the effect of that choice is visible.

E COMPUTE

Costs are measured for Active Matter with batch size 8, 11 channels, eight temporal observations, and 224×224 resolution on a single NVIDIA H100 GPU. Training time and FLOPs are per minibatch step; inference time and FLOPs are per example.

Table 10: Training and inference cost on Active Matter. Times are averages; lower is better.

Method	Train time (s/step)	Train FLOPs (TFLOPs/step)	Inference time (ms/example)	Inference FLOPs (GFLOPs/example)
ST-JEPA	0.7322	7.315	9.854	31.09
Takens-JEPA	1.4435	14.610	9.854	31.09
Multi-Cadence ST-JEPA	2.1913	21.945	9.854	31.09

Takens-JEPA uses approximately $2\times$ the training FLOPs and $1.97\times$ the training time of ST-JEPA. Multi-Cadence ST-JEPA applies the full predictive objective at three cadences, using approximately $3\times$ the ST-JEPA training cost. In this measurement, Takens-JEPA uses about one-third less training compute than Multi-Cadence ST-JEPA. These are per-step comparisons; total cost additionally depends on the number of optimization steps.

All three have the same measured inference cost because the auxiliary training components are discarded. With K current charts and one future target chart, the dominant encoder work scales approximately as $(K+1)/2$ relative to ST-JEPA's two charts. For $K = 3$, this predicts a factor of two. Pairwise distances add little work after the features have been computed.

At batch size 8, the measured training FLOPs correspond to approximately $1{,}826$ GFLOPs per example for Takens-JEPA and 914 for ST-JEPA. Their measured minibatch-step times are 1.44 and 0.73 seconds, respectively. Peak memory rises from 22.9 to 45.6 GiB. Inference remains approximately 9.85 ms and 31.1 GFLOPs per example for both.

F TAKENS MOTIVATION AND DESIGN ABLATIONS

F.1 ALTERNATIVE CHART OBJECTIVES

Equation (5) motivates correspondence between valid delay embeddings but does not determine a training loss. We compare four constructions in a learned comparison space. All variants share the ST-JEPA backbone, the predictive loss, and the projector input; they differ only in the relation imposed on projected charts.

Equality. Direct attraction minimizes

$$\mathcal{L}_{\mathrm{Eq}}^{(r,s)} = \|Q(z_i^{\tau_r}) - \mathrm{sg}(Q(z_i^{\tau_s}))\|_2^2. \tag{28}$$

This encourages agreement in projected coordinates. Its behavior depends on the projector, normalization, and other regularization terms; equality of projected features does not necessarily force identical encoder outputs.

Lie formulation. A learned generator A defines

$$Q(z_i^{\tau_s}) \approx \exp((\tau_s - \tau_r)A)Q(z_i^{\tau_r}). \tag{29}$$

The matrix exponential provides an invertible, compositional map, but restricts cadence changes to one linear one-parameter group in this space.

Koopman-style formulation. A learned linear map K defines

$$Q(z_i^{\tau_s}) \approx K^{|\tau_s - \tau_r|}Q(z_i^{\tau_r}). \tag{30}$$

This restricts cadence relations to powers of a single operator. The label describes this linear construction; the experiment does not establish a Koopman operator for the underlying physical time evolution.

Ranking formulation. Takens-JEPA uses the cadence-conditioned, normalized projector in Equation (6) and the ranking term in Equation (8).

Table 11 reports single-run Active Matter results in two settings: the single relation $\{(3, 6)\}$ for all four objectives, and the chained relations $\{(3, 6), (6, 9)\}$ of the main configuration for the three objectives we trained in that setting. The ST-JEPA row has no auxiliary relation; it is the single run corresponding to the three-run mean in Table 1.

Table 11: Alternative chart objectives on Active Matter parameter estimation, single run per row. Heads are trained at $\tau = 3$. Entries are test NMSE; bold marks the lowest value within each block.

Relations	Objective	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$
none	ST-JEPA	0.1228	0.2273	0.3715	0.5005
$\{(3, 6)\}$	Equality	0.1416	0.2426	0.3925	0.5227
	Lie	0.1195	0.2054	0.3521	0.4814
	Koopman	0.1171	0.2208	0.3949	0.5289
	Ranking (ours)	**0.0980**	**0.1334**	**0.2098**	**0.2910**
$\{(3, 6), (6, 9)\}$	Lie	0.0839	0.1434	0.2498	0.3611
	Koopman	**0.0631**	0.1293	0.2428	0.3269
	Ranking (ours)	0.0669	**0.0826**	**0.1198**	**0.1603**

With a single relation, the ranking objective has the lowest error at every cadence, and Equality is worse than ST-JEPA at every cadence. With the chained relations, the Koopman variant is marginally better at the training cadence, but the ranking objective degrades least under shift: at $\tau = 6$ its error is roughly half that of the Lie and Koopman variants. Direct equality of projected coordinates is therefore too strong a relation, and linear chart translators buy in-domain accuracy at the cost of transfer. These single-run comparisons motivate the ranking formulation; they do not establish that it recovers a chart diffeomorphism.

F.2 Endpoint and Start Alignment

The backward-delay chart in Equation (1) uses observations available at its endpoint. Histories sharing that endpoint therefore refer to the same state at the latest available observation. If fixed-length histories instead share their earliest frame and have different strides, they generally end at different times.

Takens' theorem does not privilege the endpoint. Under a forward-delay convention,

$$\widetilde{X}_i^\tau = [x_i, x_{i+\tau}, \ldots, x_{i+(d-1)\tau}], \tag{31}$$

the shared starting point can serve as the state anchor, but its representation uses observations after that time. Endpoint alignment is therefore motivated by the intended causal use of a history available at the current time.

Our same-start control uses forward charts at $\tau \in \{3, 6, 9\}$ with a common first observation and the relations $\{(3, 6), (6, 9)\}$. Because forward charts extend 63 frames past the anchor, the predictive context is placed after the full relation-chart span and predicts a later state, which keeps relation-chart observations out of the prediction target; this placement is the one difference from the endpoint-aligned model beyond the anchor itself. Both variants use the same number of training windows, backbone, and downstream parameter-estimation protocol.

Table 12: Start- versus endpoint-aligned histories on Active Matter parameter estimation, single run per row. Both use relations $\{(3, 6), (6, 9)\}$ and a head trained at $\tau = 3$; entries are test NMSE.

Chart anchor	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$
Shared start	0.1052	0.1516	0.2492	0.3565
Shared endpoint	**0.0669**	**0.0826**	**0.1198**	**0.1603**

Endpoint alignment is better at every cadence and the gap widens under shift. Because the control also moves the predictive context, the comparison does not isolate the anchor choice from the

horizon change. It shows that the endpoint convention, with the standard predictive setup, outperforms the forward-chart alternative a start-aligned design would require, and that the causal endpoint convention aligns the relation with the state the downstream model must describe.

G Illustrative Parameter-Conditioned Simulation

Figure 4 shows what the parameter error of a cadence-shifted estimate means physically. Parameters inferred from a single test chart at the unseen cadence $\tau = 5$ are supplied to the Active Matter simulator, starting from the same matched initial state as the reference trajectory. The representation models estimate parameters; the simulator generates the fields. This is one illustrative example, not a rollout benchmark.

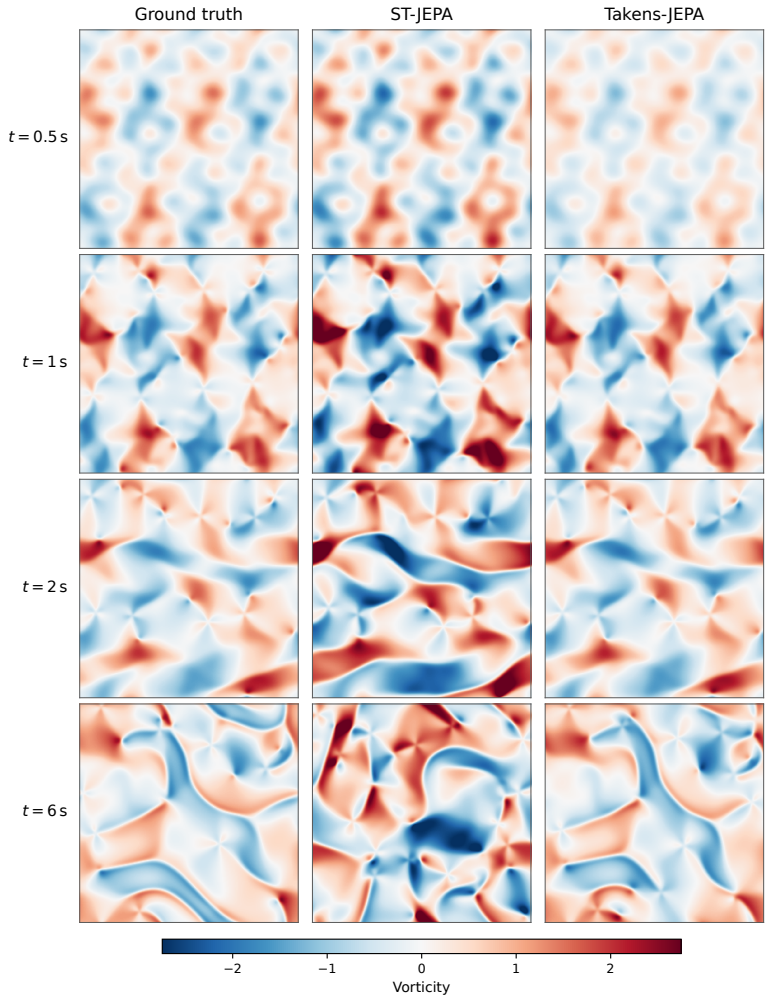

Figure 4: Active Matter simulations conditioned on parameters inferred from a test delay chart at $\tau = 5$. Columns: ground truth with $(\alpha, \zeta) = (-2.00, 7.00)$, the ST-JEPA estimate $(-3.14, 6.85)$, and the Takens-JEPA estimate $(-2.12, 6.23)$. Rows show vorticity at $t = 0.5$, 1, 2, 6 seconds on one shared color scale.

H Pseudocode

The following PyTorch-style pseudocode specifies a training step. Pooling produces a vector per example when the encoder returns spatial features; the projector outputs have shape $[B, D]$. The pre-

dictive loss includes its variance and covariance terms. For the reported configuration, `cadences` is $\{3, 6, 9\}$, `primary_tau` is 3, and `pair_set` is $\{(3, 6), (6, 9)\}$.

```
optimizer.zero_grad(set_to_none=True)

# Each batch entry identifies one trajectory and endpoint.
z = {tau: encoder(charts[tau]) for tau in cadences}
z_future = encoder(future_chart)
loss = jepa_loss(predictor(z[primary_tau]), z_future)

r = {}
for tau in cadences:
    pooled = pool_features(z[tau])
    r[tau] = normalize(projector(pooled, tau), dim=1)

takens_loss = 0
for tau_r, tau_s in pair_set:
    anchor = r[tau_r]
    same = r[tau_s].detach()
    different = same.roll(shifts=1, dims=0)
    d_same = ((anchor - same) ** 2).sum(dim=1)
    d_diff = ((anchor - different) ** 2).sum(dim=1)
    pair_loss = relu(margin + d_same - d_diff).mean()
    takens_loss += pair_loss

loss += alpha * takens_loss / len(pair_set)
loss.backward()
optimizer.step()
```

The encoder and projector are shared across cadences. Detaching the comparison branch makes each relation directed. A cyclic permutation changes batch entries; it does not certify distinct physical states. Only the encoder is retained for downstream representation extraction.

I Extended Cadence-Transfer Results

We extend physical-parameter estimation on Active Matter and Rayleigh-Bénard convection to larger observation intervals, $\tau \in \{7, 8, 9\}$. We reuse the frozen encoders and downstream heads trained at $\tau = 3$, without adaptation. Takens-JEPA uses the pretraining relations $\mathcal{P} = \{(3, 6), (6, 9)\}$. Each input still contains eight observations; increasing τ changes their spacing and temporal span, not their number. The extended evaluations use the same eligible test endpoints across cadences and methods.

Figure 2 shows that both representations become less accurate as the observation interval grows, but Takens-JEPA retains lower mean NMSE at every extended cadence. On Active Matter, its error is approximately 51%, 50%, and 49% lower than ST-JEPA at $\tau = 7, 8, 9$, respectively. At $\tau = 9$, the comparison is 0.3206 versus 0.6255. Rayleigh-Bénard shows a smaller gap, with corresponding reductions of approximately 9%, 9%, and 13%, and NMSE of 0.1148 versus 0.1317 at $\tau = 9$.

The advantage therefore persists when the downstream observation interval is increased beyond the main evaluation range, rather than disappearing immediately after $\tau = 6$. These evaluations distinguish cadences unseen during pretraining ($\tau = 7, 8$) from one used by the auxiliary objective ($\tau = 9$); all three are absent from downstream training. They support transfer to wider temporal histories within the pretraining cadence range, not extrapolation beyond that range or improved autoregressive rollout.

J Representation-Geometry Diagnostics

We examine whether the same-endpoint ranking objective learns the intended relation between temporal charts and whether either the encoder or projector exhibits representational collapse. The analysis uses three independently pretrained Active Matter checkpoints and clean validation inputs. For each checkpoint, we sample 128 held-out windows and encode charts at $\tau \in \{3, 6, 9\}$. We

Table 13: Chart-distance diagnostics on Active Matter. Each entry is the mean ± standard error across three checkpoints, where each checkpoint-level mean is computed over held-out windows. "Margin acc." is the fraction of random comparisons satisfying $d_{\text{same}} + 0.2 < d_{\text{random}}$. Lower distances are desirable for the same endpoint, while larger distances are desirable for the comparison categories.

Space	Relation	Same endpoint	Same trajectory	Same parameters	Different parameters	Random endpoint	Margin acc.
Encoder	$3 \to 6$	0.0853 ± 0.0147	0.0883 ± 0.0139	0.1192 ± 0.0193	0.1255 ± 0.0265	0.1279 ± 0.0184	$4.4\% \pm 1.4\%$
	$6 \to 9$	0.1909 ± 0.0401	0.1842 ± 0.0428	0.1966 ± 0.0405	0.2074 ± 0.0346	0.2204 ± 0.0407	$17.4\% \pm 3.7\%$
Projector	$3 \to 6$	0.1654 ± 0.0175	0.2512 ± 0.0307	0.3605 ± 0.0622	0.7112 ± 0.0169	0.6644 ± 0.0239	$69.3\% \pm 2.9\%$
	$6 \to 9$	0.1830 ± 0.0176	0.2224 ± 0.0171	0.3064 ± 0.0322	0.6202 ± 0.0355	0.5722 ± 0.0399	$57.8\% \pm 2.0\%$

evaluate the trained relations $(3, 6)$ and $(6, 9)$ in both the pooled encoder space and the cadence-conditioned projector space. All reported distances are squared Euclidean distances after ℓ_2 normalization, matching the distance used by the Takens objective. For two unit-normalized vectors, this distance equals $2 - 2\cos(\theta)$ and lies in $[0, 4]$: smaller values indicate more closely aligned representations. Because the encoder and projector have different dimensions and learned geometries, their absolute distance magnitudes are not directly comparable; the relevant comparison is between endpoint categories within each space.

For each anchor chart, we compare its same-endpoint chart with four types of comparison endpoint: another endpoint from the same trajectory, an endpoint from a different trajectory with the same governing parameters, an endpoint with different parameters, and a randomly selected endpoint. These categories separate easy negatives, which may differ in their global physical regime, from harder negatives that share a trajectory or parameter setting. The distributions are taken over validation windows; the three checkpoints are used to assess replication of the resulting statistics.

The cadence-conditioned projector produces substantially clearer separation than the pooled encoder. For $(3, 6)$, the mean same-endpoint distance is 0.1654, compared with 0.6644 for a random endpoint and 0.7112 for an endpoint with different governing parameters. The corresponding random-negative margin accuracy is 69.3%. The same pattern holds for $(6, 9)$, although the separation is weaker, with a random-negative margin accuracy of 57.8%. This confirms that the learned comparison space does not merely map every input to the same point.

The controlled negative categories reveal a more qualified picture. Endpoints from the same trajectory remain much closer to the anchor than endpoints with different parameters, and only approximately 15% of these hard negatives satisfy the full margin in projector space (same-trajectory margin accuracy of 14.8% for $(3, 6)$ and 14.6% for $(6, 9)$). This is plausible because nearby states on one trajectory may be physically similar, but it also shows that the current randomly shifted minibatch negative does not enforce uniform separation among all distinct endpoints. In encoder space, the distance ordering is considerably weaker, particularly for $(6, 9)$. This difference is expected because the ranking loss is imposed through the projector, which is discarded after pretraining, rather than directly on the pooled encoder features.

To assess collapse beyond pairwise distances, we also measure coordinate-wise variance and effective rank within each cadence. If p_j denotes the jth covariance eigenvalue divided by the sum of all eigenvalues, effective rank is $\exp(-\sum_j p_j \log p_j)$. A value near one indicates that variation is concentrated in one direction, whereas a value near the ambient dimension indicates that variation is distributed approximately evenly across the available directions. Effective rank measures spectral concentration, not the literal number of nonzero coordinates.

All encoder and projector coordinates retain nonzero variance at the stated threshold, but their covariance spectra are concentrated. The effective rank of the normalized encoder is approximately 20 at $\tau = 3$ and $\tau = 6$ and drops to approximately 10 at $\tau = 9$; a $\tau = 9$ chart covers 63 of the 80 saved-frame intervals, so the eligible endpoints (63 to 80) yield histories that cover almost the same time span, shifted by at most 17 frames, and their contents are highly correlated at that cadence. The normalized projector has an effective rank near 4 at every cadence. Together with the negative-distance results, this rules out constant collapse but indicates that the learned metric is carried by a relatively small number of dominant directions.

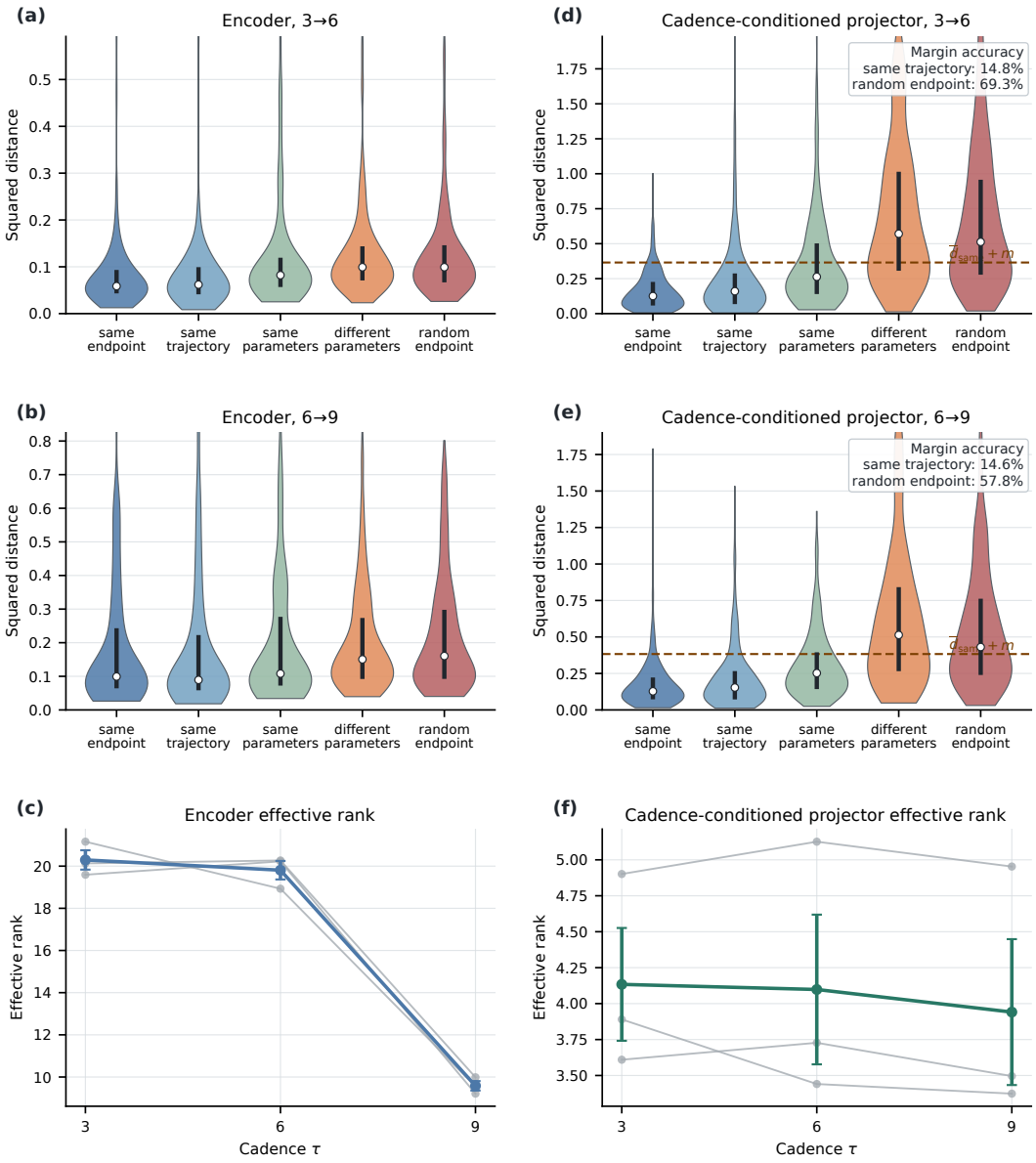

Figure 5: Representation geometry for Active Matter. Panels (a), (b), (d), and (e) show per-window distance distributions pooled across three independently pretrained checkpoints. Vertical bars and white points inside each violin denote the interquartile range and median. In the projector panels, the dashed line marks the pooled mean same-endpoint distance plus the ranking margin $m = 0.2$; annotations report the exact per-window margin accuracy using each anchor's own same-endpoint distance. Panels (c) and (f) show effective rank at each cadence; gray lines are individual checkpoints and the colored line is the mean $\pm$ standard error. Distances are computed after ℓ_2 normalization.

Table 14: Variance and effective rank of ℓ_2-normalized representations. Entries are means $\pm$ standard errors across three checkpoints. Effective rank is computed from the normalized covariance spectrum; its maximum is the ambient dimension. No feature dimension has variance below 10^{-6} in any reported setting.

Space	Cadence	Ambient dim.	Mean feature variance	Effective rank
Encoder	$\tau = 3$	128	$(4.04 \pm 0.80) \times 10^{-4}$	20.29 ± 0.46
	$\tau = 6$	128	$(5.13 \pm 0.91) \times 10^{-4}$	19.81 ± 0.44
	$\tau = 9$	128	$(10.51 \pm 2.18) \times 10^{-4}$	9.58 ± 0.22
Projector	$\tau = 3$	32	0.0108 ± 0.0004	4.13 ± 0.39
	$\tau = 6$	32	0.0084 ± 0.0004	4.10 ± 0.52
	$\tau = 9$	32	0.0086 ± 0.0003	3.94 ± 0.51