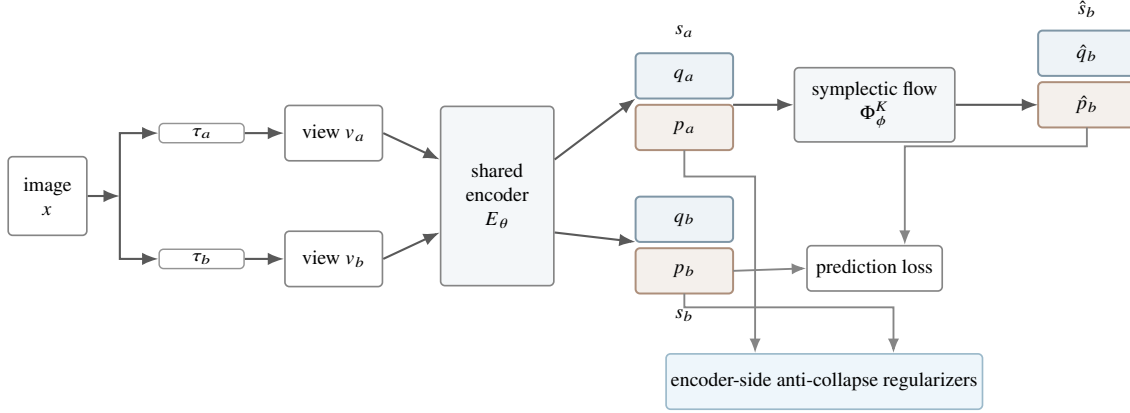


Beyond Isotropy in JEPAs: Hamiltonian Geometry and Symplectic Prediction

Technical summary

Core idea. The paper’s central reframing is that Euclidean isotropy is *not* a neutral regularizer. It is already a metric choice. If downstream tasks live in a structured geometry encoded by an SPD operator $H \succ 0$, then the canonical representation is not “round in ambient coordinates” but round only *after whitening* by $H^{1/2}$. Under an H -energy budget, both the minimax covariance and the maximum-entropy law are proportional to H^{-1} , and forcing $\Sigma \propto I$ incurs a closed-form price of isotropy.



Reported downstream results (all $q/p/(q, p)$ readouts that the paper reports explicitly).

CIFAR-100, 30 epochs		
Method	kNN@20	Linear
SIGReg	26.56 ± 0.18	30.43 ± 0.30
HamJEPA q	31.45 ± 0.27	33.95 ± 0.24
HamJEPA p	29.71 ± 0.25	33.07 ± 0.36
HamJEPA (q, p)	30.88 ± 0.39	34.18 ± 0.17

ImageNet-100, 45 epochs		
Method	kNN@20	Linear
SIGReg q	20.10	24.40
SIGReg p	18.78	23.86
SIGReg (q, p)	20.94	25.54
HamJEPA q	24.92	31.92
HamJEPA p	24.72	32.08
HamJEPA (q, p)	24.64	32.04

CIFAR-100, 80 epochs: predictor-family ablation

Method	Variant	Linear	kNN@20	kNN best
SIGReg	single	33.95	27.98	28.71 @ $k = 1$
MLP-HJEPA	q	42.26	28.10	28.10 @ $k = 20$
MLP-HJEPA	p	41.90	28.30	28.30 @ $k = 20$
MLP-HJEPA	(q, p)	42.77	28.15	28.15 @ $k = 20$
HamJEPA	q	44.59	34.43	34.43 @ $k = 20$
HamJEPA	p	44.44	32.96	32.96 @ $k = 20$
HamJEPA	(q, p)	44.52	33.66	33.66 @ $k = 20$

Note. The paper’s main 30-epoch CIFAR table reports SIGReg only as a single baseline; the explicit $q/p/(q, p)$ decomposition is reported there for HamJEPA, for both methods on ImageNet-100, and for both HamJEPA and the non-symplectic MLP-HJEPA ablation at 80 epochs.

1. The main theoretical move. A lot of JEPA work regularizes toward an isotropic Gaussian because isotropy feels like a generic anti-collapse prior. This paper says that instinct is already doing something much stronger: it is declaring the ambient Euclidean metric to be the correct notion of symmetry. If the downstream task family has structure, that is the wrong reference frame. The right question is not “should the latent be spherical?” It is “spherical with respect to *which* metric?”

The theory formalizes that point by introducing a structured SPD operator $H \succ 0$, modeling downstream linear tasks by

$$y = w^\top \tilde{z}, \quad \mathcal{W}_H = \{w : w^\top H^{-1} w \leq 1\},$$

and imposing an H -energy budget

$$\text{tr}(H\Sigma) = c, \quad \Sigma = \text{Cov}(\tilde{z}).$$

Under that geometry, the unique minimax covariance is

$$\Sigma^\star = \frac{c}{d} H^{-1},$$

and the unique maximum-entropy law under the same constraint is

$$\tilde{z} \sim \mathcal{N}\left(0, \frac{c}{d} H^{-1}\right).$$

So the canonical object is anisotropic in the original coordinates and isotropic only after whitening by $H^{1/2}$.

2. The price of using the wrong sphere. If one nevertheless insists on Euclidean isotropy, the only feasible covariance under the same budget is

$$\Sigma_{\text{iso}} = \frac{c}{\text{tr}(H)} I,$$

and the worst-case task variance degrades by

$$\rho(H) = \frac{V(\Sigma_{\text{iso}})}{V(\Sigma^\star)} = \frac{d \lambda_{\max}(H)}{\text{tr}(H)}, \quad 1 \leq \rho(H) \leq d.$$

That formula is the paper’s cleanest insight. “Isotropic” is only optimal when the task geometry itself is Euclidean, i.e. when $H \propto I$. Otherwise Euclidean isotropy is not unbiased regularization; it is a mismatch penalty.

A useful bridge to the architecture is the phase-space lift of this structured Gaussian. With

$$\mathcal{H}_H(q, p) = \frac{1}{2}q^\top Hq + \frac{1}{2}\|p\|_2^2, \quad \pi_H(q, p) \propto \exp\left(-\frac{d}{c}\mathcal{H}_H(q, p)\right),$$

the q -marginal is exactly $\mathcal{N}(0, (c/d)H^{-1})$ while $p \sim \mathcal{N}(0, (c/d)I)$. The paper’s practical move is then: if the right geometry is phase-space-like, put structure in the transport map rather than forcing a Euclidean sphere at the encoder output.

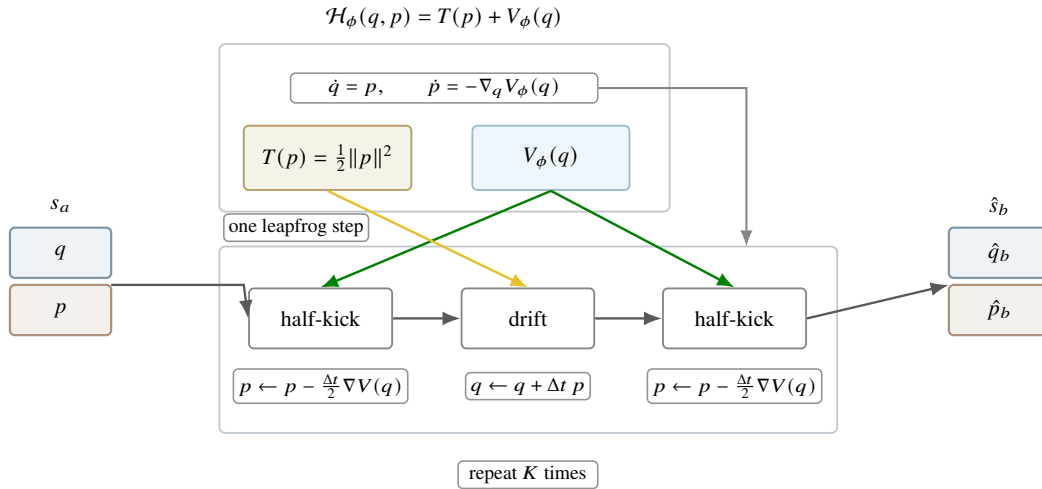
3. What HamJEPA actually changes. Because the relevant H is unknown in practice, the method does not hard-code a structured encoder marginal. Instead it places Hamiltonian structure in the *predictor*. Each view is encoded as a phase-space state $s = (q, p)$, where q is the intended content coordinate and p is an auxiliary momentum-like state. Cross-view prediction is then done by a K -step leapfrog rollout of a separable

Hamiltonian

$$\mathcal{H}_\phi(q, p) = \frac{1}{2}\|p\|_2^2 + V_\phi(q).$$

The resulting predictor map $\hat{s}_b = \Phi_\phi^K(s_a)$ is symplectic, volume-preserving, and exactly reversible under time-step sign reversal. The important claim is not that the encoder itself becomes Hamiltonian in any literal physical sense; it is that the view-to-view predictor can no longer be an arbitrary geometry-destroying regressor.

4. Symplecticity is not the same thing as anti-collapse. The paper is careful about this. A symplectic predictor cannot locally crush phase-space volume, but the encoder could still emit a degenerate state cloud. HamJEPA therefore adds encoder-side non-collapse terms: fixed-unit energy budgets for q and p , a projected log-determinant floor, and a participation-ratio floor. These are deliberately weaker and more geometry-aware than forcing $\Sigma \propto I$: they preserve scale and rank without declaring all directions equivalent.



5. What is held fixed, and why that matters. The comparisons are deliberately controlled. Across methods, the paper keeps the JEPA backbone, augmentations, optimizer, batch size, and training schedule fixed; both datasets use a headless token regime based on ResNet stage-3 features, and the projector is the identity map so that the q/p split is not remixed away by an MLP head. This is why the results should be read as a geometry argument rather than a recipe argument.

6. How to read the q/p results. On CIFAR-100 at 30 epochs, q is the strongest HamJEPA readout for kNN, while (q, p) gives the best linear probe. The intended interpretation is that q carries the cleanest metric-consistent content geometry, whereas p carries complementary linearly decodable signal that helps a learned probe more than raw nearest-neighbor structure. On ImageNet-100 the picture changes because the stable regime matches the full state (q, p) during pretraining. There, q , p , and (q, p) are all strong and nearly tied for HamJEPA, while for SIGReg the concatenated (q, p) readout remains better than either half alone. The paper reads that difference as complementarity for SIGReg versus redundancy for HamJEPA: HamJEPA makes both coordinates individually informative.

7. What the numbers imply. The main short-budget results are already decisive. On CIFAR-100 after 30 epochs, HamJEPA- q improves kNN@20 from 26.56 ± 0.18 to 31.45 ± 0.27 and linear probe from 30.43 ± 0.30 to

33.95 ± 0.24 ; (q, p) reaches the best linear probe at 34.18 ± 0.17 . On ImageNet-100 after 45 epochs, HamJEPA improves every reported readout over SIGReg, with the strict 1024-D comparison being q vs. $+4.82$ points at kNN@20 and $+7.52$ in linear probing. Even the dimension-matched 2048-D comparison remains favorable: SIGReg- (q, p) reaches 20.94 / 25.54, while HamJEPA- (q, p) reaches 24.64 / 32.04.

8. Why the 80-epoch ablation matters. The appendix makes the mechanism sharper. Replacing only the symplectic leapfrog predictor with a non-symplectic residual MLP leaves linear probe much improved over SIGReg but barely changes neighborhood quality: MLP-HJEPA peaks at just 28.30 kNN@20, while symplectic HamJEPA jumps to 34.43. That is the paper’s strongest evidence that the symplectic rollout is not decorative. Phase-space parameterization plus non-isotropic spectral control already explain much of the linear-probe gain; the large jump in local semantic geometry is what the symplectic predictor contributes.

The geometry diagnostics support that reading. On CIFAR-100 at 30 epochs, HamJEPA- q raises the top-256 effective-rank proxy from about 109.7 for SIGReg to about 125.6, while HamJEPA- p is more concentrated at about 79.5. On ImageNet-100, SIGReg- q/p sit around effective rank 38–39, whereas HamJEPA- q/p are around 95–98. So the gains track broader mean-centered covariance rather than low-rank collapse.

Bottom line. The lasting idea is not merely “use Hamiltonians in SSL.” It is more precise: *isotropy should be defined in the metric induced by the problem, not automatically in ambient Euclidean coordinates.* Once that is said clearly, the method becomes natural. The oracle geometry is H^{-1} ; because H is unknown, HamJEPA puts Hamiltonian structure in the view-to-view transport map, keeps that map symplectic, and prevents encoder degeneracy with non-isotropic spectral floors. The reported empirical pattern is consistent with exactly that thesis: broader useful covariance, stronger frozen features, and—in the clean 80-epoch predictor-family ablation—a specifically symplectic gain in neighborhood quality.