LocaLeJEPA: Maximum Entropy Representations under Geometric Priors

Anonymous authors
Paper under double-blind review

Abstract

Joint-Embedding Predictive Architectures (JEPA) offer a compelling approach to self-supervised learning, but are typically prone to representation collapse. LeJEPA resolves this by introducing a provably robust distribution-matching objective that drives the embedding towards an isotropic Gaussian. However, this approach is fundamentally tied to Euclidean assumptions. Therefore, it is incompatible with the incorporation of domain-specific geometric priors, such as the use of non-Euclidean embedding geometries, e.g., equivariant CNNs or hyperbolic ViTs. We thus introduce LocaLeJEPA, a framework motivated by a manifold-theoretic reformulation of LeJEPA's entropy-maximization property. The accompanying LocaLeReg, a novel locally Euclidean distribution-matching regularizer, enables a robust, practical implementation of the LocaLeJEPA framework. We then perform extensive experiments on models trained on ImageNet-1k. LocaLeJEPA outperforms LeJEPA not only when paired with hyperbolic ViT-S/8 and equivariant ResNet-50 backbones, but even with an ordinary Euclidean ViT-S/8, both in terms of linear and kNN classification performance on ImageNet and on a series of downstream tasks tested. Overall, we view the LocaLeJEPA formalism as a principled approach to the JEPA paradigm that better respects intrinsic data geometry while also accommodating generalizations to exotic embedding geometries.

1 Introduction

The Joint-Embedding Predictive Architecture (JEPA) paradigm (Bromley et al., 1993; LeCun, 2022; Assran et al., 2023) provides a tantalizing alternative to both feature-space masked modeling and contrastive sample-based objectives for Self-Supervised Learning (SSL). Allowing the model to instead learn relevant features by understanding the relationship between various semantically overlapping 'views' of the same object in latent space, it ideally represents a more streamlined approach to SSL. However, in practice, this learning objective falls prey to representation collapse, requiring the incorporation of techniques such as stop-gradient (Chen et al., 2020), EMA schedulers (Caron et al., 2021), and various whitening-based regularizers (Ermolov et al., 2021; Tian et al., 2021). The recent Latent-Euclidean JEPA (LeJEPA) (Balestriero & LeCun, 2025) bypasses these ad-hoc methodologies and instead identifies the isotropic Gaussian as the optimal embedding distribution for downstream tasks. Optimality here is minimax, in that no other embedding distribution yields a lower worst-case risk for linear regression or kNN classification on an unknown downstream task. LeJEPA consequently adds a Gaussian distribution matching objective to the JEPA loss, preventing collapse while being stable and scalable.

Consequent to the introduction of LeJEPA, there has been a flurry of activity in adapting this methodology for use in video models (Kuhn et al., 2026a), world models (Maes et al., 2026), and vision-language models (Kuhn et al., 2026b), as well as diverse domains such as time series (Chemeris et al., 2026), molecular representations (Rottach et al., 2026) and DXA representations.

Motivation. Concurrently, however, another line of recent work challenges the claimed optimality of LeJEPA's isotropic Gaussian, proposing instead objectives that match to rectified Gaussians (Kuang et al., 2026), subspace Gaussians (Zhao et al., 2026), or the uniform distribution on the sphere (Nicollier et al., 2026a;b). A throughline of these works is that they operate in regimes that violate the fundamental Euclidean-space assumptions of LeJEPA, thus voiding the theoretical optimality of the Gaussian distribution. For instance, Nicollier et al. (2026a) considers embeddings constrained to

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.

a hypersphere $\mathbb{S}^d$ while Kuang et al. (2026) considers l_p-normed spaces. As the JEPA framework is increasingly adapted to diverse, and often explicitly non-Euclidean domains, this naturally raises the question whether there is a principled generalization of LeJEPA to such domains that still retains its theoretical grounding and empirical robustness.

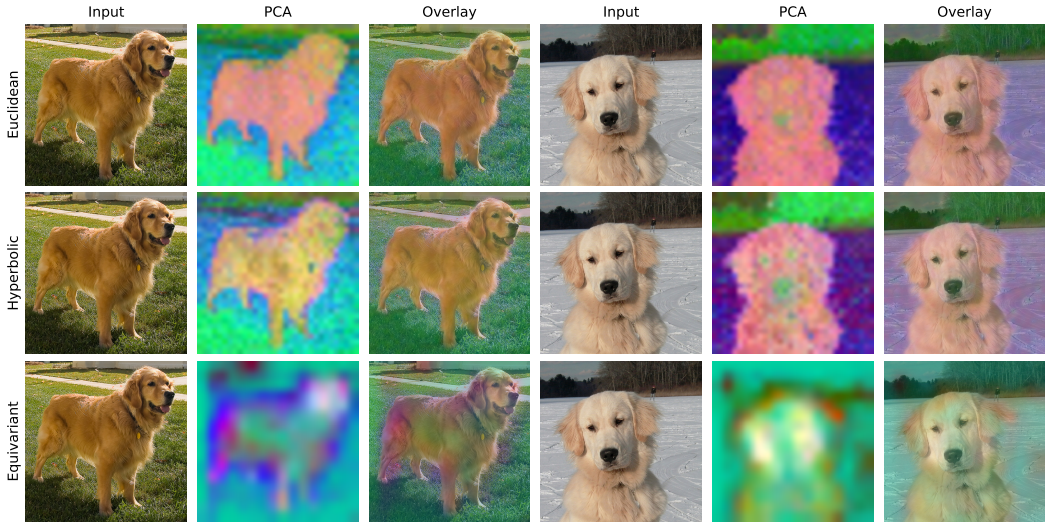


Figure 1: *LocaLeJEPA learns semantically meaningful representations, across geometries.* PCA visualizations of frozen LocaLeJEPA patch features for two ImageNet-1K validation images. Rows correspond to an ordinary (i.e. Euclidean) ViT-S/8, hyperbolic ViT-S/8, and D_4-equivariant ResNet-50, while columns show the input, the first three principal components mapped to RGB, and their overlay. Across all three geometries, the self-supervised features learned by LocaLeJEPA exhibit coherent object–background structure despite their substantially different representation spaces. Full visualization details and comparisons with LeJEPA are provided in Section D.2.

Contribution. In this work, we introduce **LocaLeJEPA**, a manifold-theoretic JEPA objective designed to allow for a rich interplay between the geometry of the data manifold $\mathcal{M}$ and of an exotic embedding space $\mathcal{N}$. The accompanying regularizer, ***Loca**lly **L**atent **e**uclidean **Reg**ularization* (LocaLeReg), drives the embeddings to be locally uniformly distributed on a Euclidean ball. Since every manifold is locally Euclidean and the uniform distribution is entropy maximization on a compact space, we view this as a local adaptation of the global entropy maximization property of LeJEPA's Gaussian distribution. We implement this by first partitioning each batch of embeddings into K neighbourhoods by K-means in latent space, and then applying an Epps-Pulley-based distribution-matching objective to each neighbourhood. This retains LeJEPA's sketched projections approach and thus inherits its linear cost and stability, while allowing for a far more expressive embedding distribution due to locality. Notably, while locality allows us to consider exotic $\mathcal{N}$, even the global $K = 1$ case leads to better preservation of the intrinsic structure of $\mathcal{M}$, as our target distribution is compact, theoretically allowing the encoder to learn a continuous function with respect to $\mathcal{M}$.

We experimentally instantiate LocaLeJEPA in the vision setting, though we note the framework is suitably general to support diverse modalities such as graph data. Across Euclidean, hyperbolic, and D_4-equivariant backbones pretrained on Imagenette and ImageNet-1K, we show LocaLeJEPA produces qualitatively better embeddings that inherently preserve the intrinsic topology of the data manifold $\mathcal{M}$, such as preserving known circular or toroidal features, while also better accommodating non-Euclidean embedding spaces $\mathcal{N}$. We then quantitatively establish the downstream performance of LocaLeJEPA's embeddings, showing it outperforms LeJEPA under linear regression and kNN classification using Euclidean, hyperbolic, and equivariant backbones, for all datasets tested. Altogether, LocaLeJEPA expands the scope of the JEPA paradigm as a whole while retaining the stability and scalability afforded by the entropy-maximizing objective of LeJEPA.

2 Related Works

Self-Supervised Learning The goal of self-supervised learning (SSL) is to learn data representations that capture intrinsic features that are useful for a wide-range of downstream tasks. Initial approaches

to SSL relied largely on positive-negative sample pairs combined with a contrastive loss, as in CPC (van den Oord et al., 2018), MoCo (He et al., 2020), and SimCLR (Chen et al., 2020). However, requiring such pairs can be costly and inefficient, leading to the development of non-contrastive approaches such as BYOL (Grill et al., 2020), SiamSiam (Chen & He, 2021), Barlow Twins (Zbontar et al., 2021), VICReg (Bardes et al., 2022), which instead involve objectives based on learning invariant features with respect to certain transformations. The JEPA paradigm extends this non-contrastive approach and may be understood as instead predicting a latent image representation from semantically overlapping views (LeCun, 2022; Assran et al., 2023).

SSL with Geometric Priors The adaptation of principles of geometric deep learning and its incorporation of geometric priors is still a burgeoning subfield of the broader SSL landscape. Initial works in this field have explored applications of equivariance to SSL, such as in the contrastive learning setting (Dangovski et al., 2022) or as alternatives to invariance-based learning objectives (Wang et al., 2024). In another direction, self-supervised and unsupervised approaches have also seen applications to 3D and point-cloud data, as in Xie et al. (2020) or Zhang et al. (2021). Further, other works have leveraged geometric architectures towards representation learning for molecular data, such as Stärk et al. (2022) and Liu et al. (2022a), as well as for graph-valued data more generally (Veličković et al., 2019; You et al., 2020; Hassani & Khasahmadi, 2020; Thakoor et al., 2021; Xu et al., 2026).

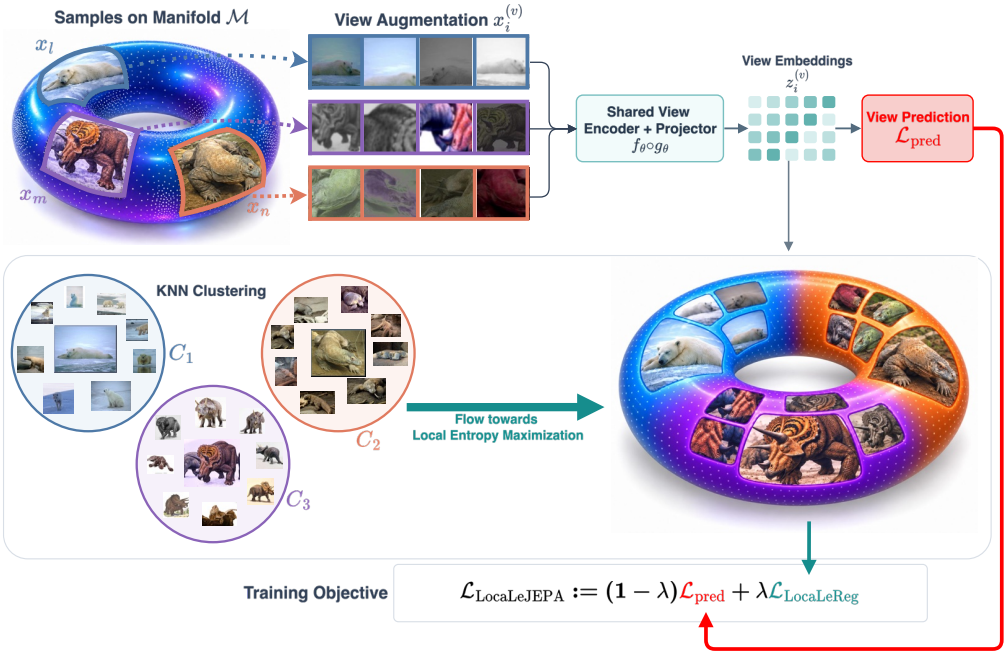


Figure 2: *LocaLeJEPA Pipeline.* A shared encoder and projector feed two losses: $\mathcal{L}_{\text{pred}}$ for view prediction and $\mathcal{L}_{\text{LocaLeReg}}$ for distribution matching. $\mathcal{L}_{\text{pred}}$ drives the model to predict an image in latent space from semantically synecdochic views. Concurrently, $\mathcal{L}_{\text{LocaLeReg}}$ counteracts representation collapse, without driving to a fixed global distribution. This is achieved by clustering batches of embedded views by K-Means, and then driving each cluster towards the uniform distribution on the Euclidean ball $\mathbb{B}^d$. Locality allows for the expressivity necessary to learn the intrinsic geometry of the data manifold, while uniformity within clusters encourages entropy maximization.

3 Theoretical Motivation

By the manifold hypothesis (Fefferman et al., 2016), we may suppose that our data lies on a compact Riemannian manifold $\mathcal{M}$, though embedded in a high-dimensional ambient space, i.e., $\mathcal{M} \subset \mathbb{R}^N$. The general objective of self-supervised learning may be understood as learning a $f : \mathbb{R}^N \to \mathcal{N}$, for $\mathcal{N}$ an appropriate embedding space for the downstream tasks under consideration. For simplicity, we assume f restricts to a smooth embedding on $\mathcal{M}$, so that the image $\mathcal{W} = f(\mathcal{M})$ is itself a closed Riemannian manifold, carrying the metric induced from $\mathcal{N}$. Here, taking $\mathcal{N}$ to be the Euclidean $\mathbb{R}^d$ recovers the setting of LeJEPA (Balestriero & LeCun, 2025) while letting $\mathcal{N}$ be the hypersphere $\mathbb{S}^d$

recovers the setting of Sphere-JEPA (Nicollier et al., 2026a). However, we may consider more general geometric priors. For instance, we may consider $\mathcal{N}$ to be hyperbolic space or a space equivariant under actions of a given Lie group.

LeJEPA (Balestriero & LeCun, 2025) then ensures non-collapse of the embedding by a distribution matching objective that drives the final embeddings towards an isotropic Gaussian, which they identify as the optimal distribution for downstream linear regression and kNN classification. A natural first attempt to generalize LeJEPA to geometric settings would be to consider a Gaussian distribution on $\mathcal{N}$. While there is no canonical analog of a Gaussian distribution such that it is 'intrinsic' to an arbitrary geometry, perhaps the most natural extension is given by considering heat flow on a manifold. Recall that the heat equation is given by $\frac{\partial \mathbf{U}(x,t)}{\partial t} = -\Delta_{\mathcal{W}}\mathbf{U}(x,t)$, where $\Delta_{\mathcal{W}}$ is the Laplace-Beltrami operator on the Riemannian manifold $\mathcal{W}$ with metric g. We then have the following.

Proposition 3.1 (Gaussian as the Solution of Heat Flow)**.** *Let $(\mathcal{W}, g)$ be a closed, connected Riemannian manifold. The solution to the heat equation on $\mathcal{W}$ with initial condition $\mathbf{U}$ is given by:*

$$\mathbf{U}(x,t) = \int_{\mathcal{W}} K_{\mathcal{W}}(x,y,t)\mathbf{U}(y,t)d\,\mathrm{Vol}_g(x) \tag{1}$$

where $K_{\mathcal{W}}(x,y,t)$ is the heat kernel and $d\,\mathrm{Vol}_g$ is the canonical volume form. Explicitly, the heat kernel on $\mathbb{R}^n$ is given by $K_{\mathbb{R}^n}(x,y,t) = \frac{1}{4\pi t^{n/2}}\exp\left(-\frac{|x-y|^2}{4t}\right)$ and on arbitrary closed Riemannian $\mathcal{W}$, as $t \to 0$, a first order approximation is given by $K_{\mathcal{W}}(x,y,t) \approx \frac{1}{4\pi t^{n/2}}\exp\left(-\frac{d(x,y)^2}{4t}\right)$, where $d(x,y)$ is the geodesic distance on $\mathcal{W}$. (See Berline et al. (1992) for details).

Proposition 3.1 therefore lets us define a 'Gaussian' intrinsically on a manifold, as the heat kernel $K_{\mathcal{W}}(x,\cdot,t)$ centered at x with t in the role of a variance. That identification holds only as $t \to 0$, however, and in exactly that regime the kernel is concentrated near x. An intrinsic Gaussian is therefore a *strictly local* object on $\mathcal{W}$. As such, in practical terms, we may consider a local version of LeJEPA's SIGReg as an appropriate analog for the manifold setting. Unfortunately, this naive approach does not preserve LeJEPA's optimality guarantees. For instance, in Sphere-JEPA (Nicollier et al., 2026a), it is shown that for the case $\mathcal{N} = \mathbb{S}^d$, the uniform distribution is optimal for downstream linear regression and kNN classification tasks, and they propose SUSReg, a distribution matching objective that instead drives towards the uniform distribution. Taking on a heat flow perspective allows us to interpolate between these objectives, not just on a sphere, but on an arbitrary $\mathcal{N}$. In particular, we then have an entire family of distributions $K_{\mathcal{W}}(\cdot,\cdot,t)$, parameterized by time, that may reasonably be understood as targets for a distribution-matching objective that 'lifts' LeJEPA to manifolds. However, this would not consider the entropy maximization property of the Gaussian distribution that has been tied to the success of LeJEPA, in works such as Rectified LpJEPA (Kuang et al., 2026). More generally, an entropy-maximization objective has been shown to be desirable in SSL (Liu et al., 2022b; Zhang et al., 2024; Chakraborty et al., 2025). We recall that the usual Gaussian distribution, under a given variance, is the unique entropy-maximizing distribution in Euclidean space, while the uniform distribution is the unique entropy-maximizing distribution on an arbitrary compact Riemannian manifold. The optimal transport perspective allows us to understand heat flow as gradient flow towards entropy maximization on an arbitrary manifold.

Proposition 3.2 (Heat Flow as Wasserstein Gradient Flow)**.** *Let $(\mathcal{W}, g)$ be a closed, connected Riemannian manifold. Recall that the differential entropy of a distribution ρ on $\mathcal{W}$ is given by :*

$$h(\rho) = -\int_{\mathcal{W}} \rho(x)\log\rho(x)\,d\,\mathrm{Vol}_g(x) \tag{2}$$

Then in the 2-Wasserstein space of measures, $\mathcal{P}_2(\mathcal{W})$, the sequence of distributions arising as solutions of $\frac{\partial \rho_t}{\partial t} = -\Delta_{\mathcal{W}}\rho$ constitute gradient flow of the entropy functional in $\mathcal{P}_2(\mathcal{W})$. (See Ambrosio et al. (2013) or Jordan et al. (1998) for details).

This sequence is bracketed by its two asymptotic ends: the short-time regime of Proposition 3.1, and the following long-time regime.

Proposition 3.3 (Uniform Distribution as the Solution of Heat Flow)**.** *Let $(\mathcal{W}, g)$ be a closed, connected Riemannian manifold. Then in the infinite-time regime, heat flow satisfies:*

$$\mathbf{U}(\cdot,t) \xrightarrow[t\to\infty]{} \frac{1}{\mathrm{Vol}_\infty(\mathcal{W})}\int_{\mathcal{W}} \mathbf{U}(\cdot,0)d\,\mathrm{Vol}_g(x) \tag{3}$$

in $L^2(\mathcal{W})$. Equivalently, if $\boldsymbol{U}(\cdot, 0)$ is an initial probability distribution, $\boldsymbol{U}(x,t) \xrightarrow[t\to\infty]{} \frac{1}{\mathrm{Vol}(\mathcal{W})}$ uniformly in x, i.e. converges to the uniform Riemannian volume measure. (See Berline et al. (1992)).

As such, under this additional constraint of entropy maximization, the optimal embedding distribution sits at the $t \to \infty$ end of the heat-flow family. That is, we want $f(\mathcal{M}) \subset \mathcal{N}$ to be 'maximally spread out' while preserving the relevant geometry of $\mathcal{M}$. Notably, the simple 'geometry preservation' assumption already precludes the final embedding from being Gaussian even in the Euclidean setting $\mathcal{N} = \mathbb{R}^d$. Indeed, a continuous f carries the compact $\mathcal{M}$ to a compact image, and thus driving towards the infinitely-supported Gaussian distributions necessarily causes us to learn an f that 'tears' $\mathcal{M}$. Thus, we wish to modify the LeJEPA paradigm such that **i.** we can preserve the structure of $\mathcal{M}$ and **ii.** without conditioning on a predetermined $\mathcal{N}$, while **iii.** still allowing for a practical implementation that is stable and scalable. To drive towards a true entropy-maximizing distribution would typically require either a static but globally formulated objective, conditioned on a fixed choice of $\mathcal{N}$ and violating **ii**, or a local-to-global heat flow style objective exploiting Proposition 3.2, which would then require dynamic re-computations of the distribution matching objective during training, likely violating **iii**. As most reasonable notions of 'geometry preserving' for **i** would imply that $f(\mathcal{M})$ should be locally Euclidean, we propose a compromise *locally entropy maximizing* objective, by driving towards the uniform distribution on a Euclidean ball $\mathbb{B}^d$ on neighborhood clusters.

4 LocaLeReg

We propose *locally latent Euclidean Regularization* (LocaLeReg), a novel regularizer that realises the local target of Section 3 in practice. To keep the implementation stable and scalable, we obtain it through minor modifications to the sketched isotropic Gaussian regularization (SIGReg) of LeJEPA.

A batch consists of N samples, each presented in V augmented views $x_i^{(v)}$. As is standard in LeJEPA (Balestriero & LeCun, 2025), an encoder is followed by a projector, and it is on the embeddings $z_i^{(v)} \in \mathbb{R}^d$ it produces that the pretraining loss is computed, while downstream probes read the encoder outputs. In the notation of Section 3, the map f is this composition and $\mathcal{N}$ the space in which the embeddings live. We write $\bar{z}_i = \frac{1}{V}\sum_{v=1}^{V} z_i^{(v)}$ for the mean of a sample's embeddings across views. Recall that LeJEPA's objective is given as $(1-\lambda)\mathcal{L}_{\text{pred}} + \lambda\mathcal{L}_{\text{SIGReg}}$, where $\lambda \in (0,1)$ trades the two terms off and $\mathcal{L}_{\text{pred}}$ is the usual JEPA prediction term,

$$\mathcal{L}_{\text{pred}} = \frac{1}{NVd}\sum_{i=1}^{N}\sum_{v=1}^{V}\left\|z_i^{(v)} - \bar{z}_i\right\|_2^2. \tag{4}$$

Further, LeJEPA's SIGReg leverages a hyperspherical version of the Cramér–Wold theorem (Cramér & Wold, 1936), which implies that, for $X, Y, \mathbb{R}^d$-valued random vectors,

$$\langle u, X\rangle \overset{d}{=} \langle u, Y\rangle, \forall u \in \mathbb{S}^{d-1} \iff X \overset{d}{=} Y. \tag{5}$$

where $\overset{d}{=}$ denotes equality in distribution. This allows us to bring to bear the univariate Epps-Pulley test (Epps & Pulley, 1983), which measures L^2 discrepancy between characteristic functions under a weight w, which we take throughout to be the Gaussian window $w(t) = e^{-t^2/2}$:

$$\mathrm{EP}(X,Y) = N\int_{\mathbb{R}} \left|\hat{\varphi}_X(t) - \varphi_Y(t)\right|^2 w(t)dt. \tag{6}$$

Here, X is now a scalar random variable with empirical characteristic function $\hat{\varphi}_X(t) = \frac{1}{N}\sum_{j=1}^{N} e^{itx_j}$ while Y is a fixed reference distribution with characteristic function $\varphi_Y(t) = \mathbb{E}\left[e^{itY}\right]$. SIGReg thus consists of drawing directions $u \sim \mathrm{Unif}\left(\mathbb{S}^{d-1}\right)$, and for each view v, matching the empirical distribution of the $\left\{u^\top z_i^{(v)}\right\}_{i=1}^{N}$ against the standard normal $Y = \mathcal{N}(0,1)$ via Epps-Pulley.

In the LocaLeJEPA pipeline, we begin by constructing local neighborhoods of our embeddings through cosine-similarity K-means on the view means $\bar{z}_i$, held fixed so that no gradient passes through the assignment, giving a clustering $\{C_q\}_{q=1}^{K}$ that is shared across views. Neighborhoods falling below a fixed minimum size are discarded as too sparse to test, leaving the valid set $\mathcal{Q}$ of retained neighborhoods. If it is empty, the objective reverts to the global test on the whole batch. Within a cluster, we consider the residuals of our embeddings, given by

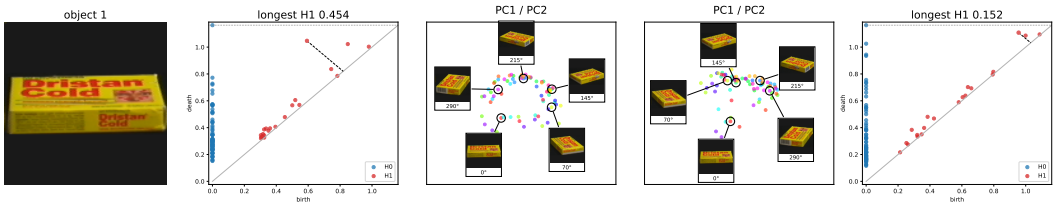


Figure 3: *LocaLeJEPA embeddings better preserve intrinsic topology.* The COIL-100 consists of images of objects being rotated. As such, these images should trace out the circle, $\mathbb{S}^1$, in latent space. As measured by the persistent homology class H_1 and PCA visualization, the embeddings more naturally lie on $\mathbb{S}^1$ when using LocaLeReg as opposed to LeJEPA's SIGReg. Additional experimental details may be found in Section D.1.

$$r_i^{(v)} = z_i^{(v)} - \mu_q^{(v)}, \qquad \mu_q^{(v)} = \frac{1}{|C_q|} \sum_{j \in C_q} z_j^{(v)}. \tag{7}$$

We then drive each cluster towards the uniform distribution on a Euclidean ball $\mathbb{B}^d$. Note that, like the isotropic Gaussian distribution or the uniform distribution on the sphere, this target distribution is radially symmetric and thus amenable to being precomputed for use in the Epps-Pulley test. We may explicitly derive (see Section A.1 for details) that the characteristic function of the uniform distribution Y on the ball $\mathbb{B}_R^d$ of radius R, which we set to $R = \sqrt{d+2}$ so that the target has identity covariance and is normalised exactly as SIGReg's isotropic Gaussian, is given by

$$\varphi_Y(t) = \mathbb{E}[e^{\mathrm{i}t\langle Y,u\rangle}] = 2^{d/2}\Gamma\left(\frac{d}{2}+1\right)\frac{J_{d/2}(R|t|)}{(R|t|)^{d/2}}, \qquad \varphi_Y(0) = 1, \tag{8}$$

where J_ν is the Bessel function of the first kind. In this case, our empirical characteristic function is

$$\widehat{\varphi}_{q,v,u}(t) = |C_q|^{-1} \sum_{i \in C_q} e^{\mathrm{i}t\langle r_i^{(v)},u\rangle}. \tag{9}$$

LocaLeJEPA Objective

LocaLeReg aggregates the local discrepancies over the valid neighborhoods $\mathcal{Q}$ and the V augmented views:

$$\mathcal{L}_{\text{LocaLeReg}} = \frac{1}{|\mathcal{Q}|V} \sum_{q \in \mathcal{Q}} \sum_{v=1}^{V} D_{q,v}, \tag{10}$$

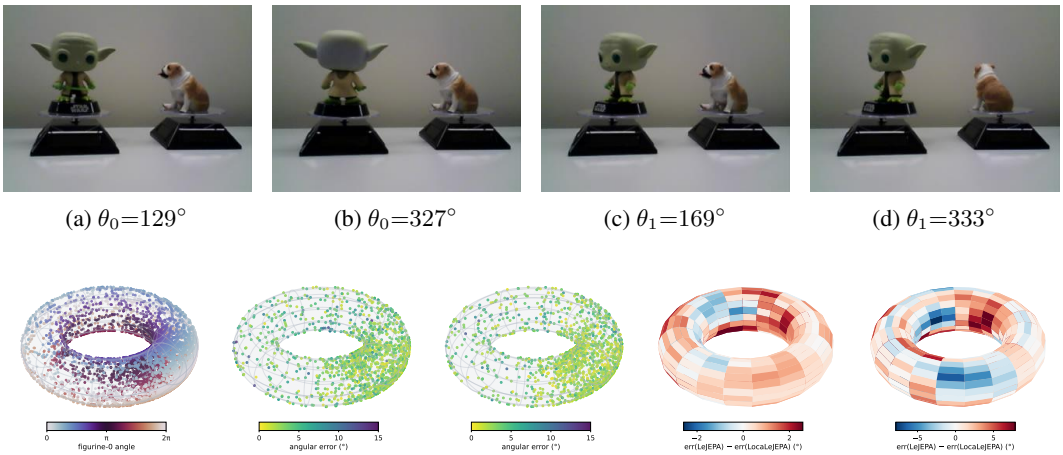
where:

$$D_{q,v} = |C_q|\, \mathbb{E}_{u \sim \text{Unif}(\mathbb{S}^{d-1})} \int_{\mathbb{R}} |\widehat{\varphi}_{q,v,u}(t) - \varphi_Y(t)|^2\, e^{-t^2/2}\, dt. \tag{11}$$

Combining local distribution matching with the prediction objective gives:

$$\boxed{\mathcal{L}_{\text{LocaLeJEPA}} = (1-\lambda)\mathcal{L}_{\text{pred}} + \lambda \mathcal{L}_{\text{LocaLeReg}}} \tag{12}$$

The number of neighborhoods K controls the discretization of the local regime of Section 3, as it sets the scale at which the embedded manifold is treated as Euclidean, and at $K = 1$ the objective becomes global, differing from SIGReg in its target distribution and in that centering the residuals leaves the location free. The construction does require each neighborhood to retain enough residuals for its statistic to be informative, since centering leaves only $|C_q| - 1$ free directions and the neighborhoods are formed within the batch.



(a) θ_0=129° (b) θ_0=327° (c) θ_1=169° (d) θ_1=333°

(e) Ground-truth (f) LeJEPA (g) LocaLeJEPA (h) Full pool (i) 100 frames

Figure 4: *Test-time Topology Preservation.* We perform a controlled experiment to test whether LocaLeJEPA's topology-preservation extends to datasets unseen during training. After pretraining on Imagenette (Howard, 2019), we expose it to a dataset consisting of two-independently rotating figurines *(Top)*, which should thus lie on a latent torus $\mathbb{T}^2$. We consider performance by testing angle prediction of each of the figurines. *(Bottom)*(a) Ground-truth torus defined by the two rotation angles. (b,c) Reconstructions obtained from frozen LeJEPA (MAE 3.28°) and LocaLeJEPA (2.68°) features, colored by angular error. (d,e) Regional difference $\mathrm{err}(\mathrm{SIGReg}) - \mathrm{err}(\mathrm{LocaLeJEPA})$ using the full probe pool (+0.60°, 61% wins) and 100 training frames (+1.31°, 57% wins), with red favouring LocaLeJEPA. Full grids are in Section C.3.3.

5 Experiments

LocaLeJEPA is designed to better preserve the intrinsic manifold structure of data, while also being adaptable to exotic embedding geometries. While LocaLeJEPA's embeddings can be qualitatively shown to better preserve topology (see Figure 3 and Figure 4) of data while also adapting to the geometry of the embedding space (see Figure 1 and Figure 5), our main experiments focus on the performance of the LocaLeJEPA framework at scale. In particular, we perform pretraining on ImageNet-1K using Euclidean, hyperbolic, and equivariant backbones under the LocaLeJEPA objective. We then evaluate performance through frozen-feature classification on ImageNet-1K in Section 5.1, using both linear and kNN classification, where it consistently outperforms LeJEPA in all regimes tested. We then consider transferability of the resulting representations across seven downstream datasets in Section 5.2, where it again outperforms LeJEPA in all regimes tested. Complete experimental details, including training protocols, checkpoint-selection procedures, evaluation settings, and additional experiments and ablations over K, λ, and batch size are provided in Section C, with the ImageNet-specific analysis reported in Section C.4.3.

5.1 ImageNet-1K Frozen-Feature Classification

Task. We employ ImageNet-1K (Deng et al. (2009)) to test whether the advantages of LocaLeJEPA persist at scale and remain consistent across distinct representation geometries. Following (Balestriero & LeCun (2025)), we pretrain three encoder families for 100 epochs: a standard ViT-S/8 with Euclidean outputs, a hyperbolic ViT-S/8 whose representations lie in a Poincaré ball of curvature -0.1 (Ermolov et al. (2022)), and a D_4-equivariant ResNet-50 retaining the regular representation at its output (Klee et al. (2023)). All models use

Table 2: *LocaLeJEPA better preserves toroidal structure across probe-label budgets.* Continuing with the setting of the toroidal angle prediction task of Figure 4, we consider performance against a range of global distribution matching objectives: Balestriero & LeCun (2025)'s SIGReg, Nicollier et al. (2026a)'s SUSReg, Zhao et al. (2026)'s SubReg, and Kuang et al. (2026)'s RDMReg. LocaLeJEPA features prove superior for this task across label budgets. Additional details are provided in Section C.3.

Method	10	50	100	500	Full
nViT SUSReg	44.6±4.3	20.2±2.0	11.0±0.7	4.91±.09	3.64±.04
ViT SIGReg	39.2±3.6	17.8±1.1	11.0±0.4	4.70±.09	3.30±.05
ViT SubReg	37.9±4.4	18.0±1.7	11.2±0.5	4.70±.10	3.40±.04
ViT RDMReg	39.8±7.0	19.8±2.0	10.5±0.6	5.47±.08	3.79±.05
LocaLeJEPA (K=2)	**36.0±3.8**	**15.5±1.2**	**9.0±0.3**	**3.80±.05**	**2.67±.05**
LocaLeJEPA (Dyn-K)	37.5±4.3	17.1±1.1	10.9±0.5	4.10±.10	2.89±.04

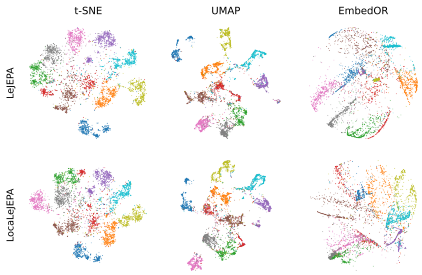


Figure 5: *LocaLeJEPA clusterings are curvature-aware.* Two-dimensional projections of frozen hyperbolic ViT-S/8 representations trained with LeJEPA (top) and LocaLeJEPA (λ=0.1, K=4; bottom). Columns show t-SNE, UMAP, and EmbedOR (Saidi et al., 2026); colors indicate ground-truth classes. While Euclidean visualizers show similar layouts, curvature-aware EmbedOR reveals substantially different cluster structure.

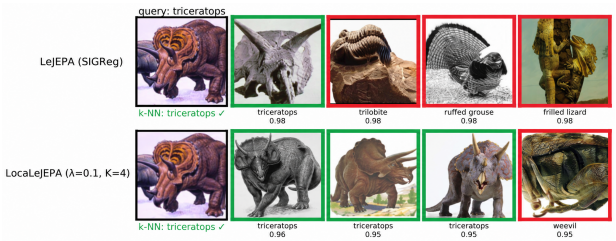


Figure 6: *LocaLeJEPA learns more semantically consistent local neighborhoods.* For representative ImageNet-1K queries, we retrieve the nearest training images in the frozen ViT-S/8 representation space. Green borders indicate neighbors sharing the query class, and red borders indicate different classes. Compared with LeJEPA's SIGReg, LocaLeJEPA retrieves substantially more same-class examples for the illustrated queries, consistent with a more semantically organized local feature geometry. The corresponding analysis over all 1,000 classes and additional experimental details are provided in Section D.3.

Table 1: *ImageNet-1K Frozen-feature Evaluation across Representation Geometries.* ImageNet classification, seven-dataset transfer, and bobblehead rotation recovery for the principal Euclidean, hyperbolic, and D_4-equivariant configurations. Full sweeps and additional baselines are reported in Table 18. **Bold** denotes the best result within each encoder family and metric.

	ImageNet-1K linear (%)		k-NN (%)	Transfer (%)
Method	top-1(↑)	top-5(↑)	k=20(↑)	7-dataset avg(↑)
Plain ViT-S/8 (Euclidean, 21.7M)				
LeJEPA	40.7 ± 0.23	66.9 ± 0.21	24.3 ± 0.19	56.7 ± 0.29
LocaLeJEPA (λ=0.1, K=4)	$\mathbf{49.6 \pm 0.23}$	$\mathbf{75.0 \pm 0.20}$	$\mathbf{34.1 \pm 0.21}$	$\mathbf{63.0 \pm 0.32}$
Hyperbolic ViT-S/8 (Poincaré ball, 21.7M)				
LeJEPA	41.1 ± 0.22	67.1 ± 0.21	25.1 ± 0.19	57.6 ± 0.60
Rectified LpJEPA	29.2 ± 0.21	53.4 ± 0.23	15.1 ± 0.16	47.5 ± 0.28
LocaLeJEPA (λ=0.1, K=4)	$\mathbf{50.5 \pm 0.23}$	$\mathbf{75.8 \pm 0.19}$	$\mathbf{34.9 \pm 0.21}$	$\mathbf{63.9 \pm 0.48}$
LocaLeJEPA (λ=0.1, K=1→10)	47.4 ± 0.22	73.1 ± 0.20	29.9 ± 0.20	61.3 ± 0.29
LocaLeJEPA-SIGReg (λ=0.1, K=4)	40.7 ± 0.22	66.5 ± 0.21	26.0 ± 0.20	57.2 ± 0.27
D_4-equivariant ResNet-50 (24.4M)				
LeJEPA	35.7 ± 0.22	63.0 ± 0.22	23.3 ± 0.19	48.2 ± 0.32
Rectified LpJEPA	27.9 ± 0.20	54.1 ± 0.22	19.7 ± 0.18	42.6 ± 0.44
LocaLeJEPA (λ=0.1, K=4)	$\mathbf{42.3 \pm 0.22}$	$\mathbf{69.8 \pm 0.21}$	$\mathbf{33.0 \pm 0.21}$	54.0 ± 0.31
LocaLeJEPA (λ=0.1, K=1→10)	42.2 ± 0.22	69.5 ± 0.21	32.5 ± 0.21	$\mathbf{54.3 \pm 0.40}$
LocaLeJEPA-SIGReg (λ=0.1, K=4)	35.9 ± 0.21	62.7 ± 0.22	27.0 ± 0.20	50.3 ± 0.29

four 128×128 views, a 2048-2048-16 projector, and four-step gradient accumulation over batches of 64 images. After pretraining, the encoder is frozen and evaluated with cosine-weighted k-NN (k=20) and linear probing on its 512-dimensional representation. In the comparison, we also include *LocaLeJEPA-SIGReg*, which keeps the same local partition but replaces centered matching to $\mathrm{Unif}\left(\mathbb{B}^{d-1}_{\sqrt{d+2}}\right)$ with Gaussian matching on uncentered clusters. This tests whether locality alone explains the gains, without disentangling centering from reference-distribution choice. Complete optimization details, k-NN sweeps, top-5 results, and ablations over K, λ, and the local reference distribution are reported in Section C.4.3.

Results. We consider LocaLeJEPA with both fixed numbers of local clusters, as well as a dynamic approach, K=1→10, where we steadily increase the number of clusters every ten training epochs, aligned with our heat-flow heuristic. LocaLeJEPA in both regimes improves frozen-feature recogni-

tion across all three representation geometries (Table 1), with gains of 6.6–9.3 points in linear top-1 accuracy and 9.7–9.8 points in k-NN over LeJEPA's SIGReg. The strikingly stable $\sim$ 10-point k-NN gain is particularly informative because this readout operates directly on the frozen feature metric, and is consistent with the stronger local semantic organization quantified in Section D.3.

5.2 Transfer to Downstream Datasets

Task. We next test whether the ImageNet-1K gains persist beyond the pretraining distribution. Keeping each encoder frozen, we fit a single linear classifier on standardized features for seven downstream datasets spanning textures, fine-grained recognition, and generic object classification: DTD (Cimpoi et al. (2014)), FGVC-Aircraft (Maji et al. (2013)), CIFAR-10 and CIFAR-100 (Krizhevsky (2009)), Flowers-102 (Nilsback & Zisserman (2008)), Food-101 (Bossard et al. (2014)), and Oxford-IIIT Pets (Parkhi et al. (2012)). The probe learning rate is selected on a 10% holdout from the training split, and top-1 accuracy is reported on the official test set. Full per-configuration results and protocol details are provided in Section C.4.3 and Section C.4.

Results. The ImageNet-1K gains transfer consistently beyond the pretraining distribution (Table 3). Across the Euclidean, hyperbolic, and equivariant encoder families, LocaLeJEPA improves the seven-dataset average by 6.0–6.4 percentage points over LeJEPA's SIGReg. Both ViT encoders improve on all seven datasets, with the largest gains on fine-grained recognition tasks such as Pets and Flowers. The equivariant ResNet encoder further highlights the role of dynamic locality: fixed $K{=}4$ improves five of seven datasets, whereas the $K{=}1{\to}10$ schedule improves all seven and attains the highest transfer average in this family (54.6%). This means that the benefit of local distribution matching is not confined to ImageNet-1K classification, but persists across substantially different representation geometries and downstream domains.

Table 3: *Linear Transfer across Seven Downstream Datasets.* Frozen-feature top-1 accuracy (%); *Avg* is the unweighted mean. Each entry is the mean over five probe seeds. **Bold** denotes the best mean within each encoder family and column.

Method	DTD	Aircraft	CIFAR-10	CIFAR-100	Flowers	Food	Pets	Avg
Plain ViT-S/8 (Euclidean, 21.7M)								
LeJEPA (SIGReg)	56.4±1.5	26.1±0.8	84.4±0.4	61.2±0.6	68.4±0.6	48.5±0.3	52.0±0.8	56.7±0.3
LocaLeJEPA ($\lambda{=}0.1$, $K{=}4$)	**63.0±1.2**	**31.1±0.9**	**86.0±0.4**	**63.8±0.5**	**77.4±0.9**	**55.7±0.3**	**64.2±1.6**	**63.0±0.3**
Hyperbolic ViT-S/8 (Poincaré ball, 21.7M)								
LeJEPA (SIGReg)	57.4±2.0	27.3±0.9	85.3±0.4	60.3±0.5	70.9±2.6	48.6±0.3	53.4±0.8	57.6±0.6
LocaLeJEPA ($\lambda{=}0.1$, $K{=}4$)	**62.1±1.2**	**32.1±0.8**	**87.5±0.3**	**65.0±0.5**	**77.9±3.3**	**56.4±0.3**	**65.9±0.9**	**63.9±0.5**
LocaLeJEPA ($\lambda{=}0.1$, $K{=}1{\to}10$)	59.5±1.4	28.6±1.0	86.5±0.3	64.0±0.6	74.7±0.6	54.0±0.3	62.0±1.0	61.3±0.3
LocaLeJEPA-SIGReg ($\lambda{=}0.1$, $K{=}4$)	55.7±1.3	27.0±0.9	83.1±0.4	59.0±0.6	70.5±0.6	49.0±0.3	56.5±0.8	57.2±0.3
D_4-equivariant ResNet-50 (24.4M)								
LeJEPA (SIGReg)	51.3±1.4	16.4±0.7	83.8±0.4	55.8±0.5	52.9±0.7	34.3±0.3	42.7±1.1	48.2±0.3
LocaLeJEPA ($\lambda{=}0.1$, $K{=}4$)	55.1±1.2	20.7±1.4	82.3±0.5	55.3±0.5	**66.4±0.6**	**41.8±0.3**	**56.6±0.8**	54.0±0.3
LocaLeJEPA ($\lambda{=}0.1$, $K{=}1{\to}10$)	**56.7±1.2**	**21.7±1.5**	**84.8±0.4**	**58.5±0.5**	63.5±1.7	40.5±0.3	54.6±1.0	**54.3±0.4**
LocaLeJEPA-SIGReg ($\lambda{=}0.1$, $K{=}4$)	50.0±1.2	20.4±0.9	82.8±0.4	54.9±0.5	59.0±0.7	32.9±0.3	51.8±0.9	50.3±0.3

The LocaLeJEPA-SIGReg controls (Tables 1 and 3) indicate that the centered uniform-ball target contributes to these gains: within the same neighborhood LocaLeJEPA-SIGReg performs worse than LocaLeJEPA for both hyperbolic and equivariant encoders. This further validates that uniformity, i.e. entropy maximization, plays a key role in the downstream performance of LocaLeJEPA's embeddings and not just locality alone. See also Table 4b for further discussion and comparison to the local Gaussian objective.

6 Conclusion

LocaLeJEPA brings a manifold-theoretic perspective to the JEPA paradigm, introducing the heuristic that self-supervised training should behave like heat flow. As a practical implementation, we develop the novel LocaLeReg regularizer, based upon a locally Euclidean entropy-maximizing objective. We find that both qualitatively and quantitatively, LocaLeJEPA produces embeddings that better preserve intrinsic semantic structure and can adapt to the geometry of the representation space, while being stable and scalable. While we instantiate our framework on image data in the current work, we believe LocaLeJEPA represents a general-use framework that allows for a principled extension of the JEPA paradigm to settings that require handling non-Euclidean data or harnessing non-Euclidean representations.

AI USE STATEMENT

In this work, we used generative AI tools to assist with both scientific writing and parts of the experimental workflow. For writing, these tools were used to improve clarity, grammar, phrasing, and overall presentation. We also used AI-based coding agents to support the implementation of experimental pipelines, debugging, execution of experiments, and preliminary analysis of experimental outputs. All AI-assisted code, analyses, and experimental results were independently inspected and validated by the authors, including through extensive unit testing, consistency checks, and comparison against independently verified implementations and expected behavior. Generative AI tools were not relied upon to autonomously formulate the scientific contributions or determine the conclusions of the work. All scientific interpretations, claims, and conclusions were made and verified by the authors. We take full responsibility for the methodology, experimental results, analysis, and final content of this work.

ETHICS STATEMENT

This work does not involve human subjects, personally identifiable information, or the collection of sensitive data. The datasets used in our experiments are publicly available research benchmarks and are used in accordance with their respective licenses and intended research purposes. We are not aware of any significant ethical concerns arising from the proposed methodology or its experimental evaluation. As with other machine learning methods, downstream applications should nevertheless consider potential issues related to bias, fairness, privacy, and misuse. We take responsibility for ensuring that the research is conducted and reported in accordance with established standards of research integrity.

REPRODUCIBILITY STATEMENT

We have taken several steps to facilitate reproducibility of both the proposed method and the experimental results. The main paper provides the core methodological formulation, model design, and experimental setting, while the appendices contain the additional information required to reproduce the complete pipeline. In particular, Sections C.1, C.3 and C.4.1 provide implementation and architectural details, report the hyperparameters and training configurations used throughout the experiments, and describe the preprocessing procedures and evaluation protocols. All experiments were run from explicit configuration files, and the experimental code was subjected to systematic consistency checks and unit testing, including tests of the main model components and experimental pipelines. The source code, configuration files, and scripts required to reproduce the reported experiments, baselines, ablations, and evaluation procedures are available at https://anonymous.4open.science/r/LocaLeJEPA.

REFERENCES

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. Heat flow and calculus on metric measure spaces with ricci curvature bounded below—the compact case. In *Analysis and Numerics of Partial Differential Equations*, volume 4, pp. 63–115. Springer, Milan, 2013. doi: 10.1007/978-88-470-2592-9_8.

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

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025. doi: 10.48550/arXiv.2511.08544. URL https://arxiv.org/abs/2511.08544.

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

Ulrich Bauer. Ripser: Efficient computation of Vietoris–Rips persistence barcodes. *Journal of Applied and Computational Topology*, 5:391–423, 2021.

Nicole Berline, Ezra Getzler, and Michèle Vergne. *Heat Kernels and Dirac Operators*, volume 298 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin, 1992. ISBN 3-540-53340-0. doi: 10.1007/978-3-642-58088-8.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Jane Bromley, James W. Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(4):669–688, August 1993. doi: 10.1142/S0218001493000339.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, October 2021.

Deep Chakraborty, Yann LeCun, Tim G. J. Rudner, and Erik Learned-Miller. Improving pre-trained self-supervised embeddings through effective entropy maximization. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 433–441. PMLR, 2025.

Frédéric Chazal, David Cohen-Steiner, Leonidas J. Guibas, Facundo Mémoli, and Steve Y. Oudot. Gromov–hausdorff stable signatures for shapes using persistence. *Computer Graphics Forum*, 28 (5):1393–1403, 2009.

Alexander Chemeris, Ming Jin, and Randall Balestriero. Lenepa: No-augmentation next-latent prediction for time-series representation learning. In *Proceedings of the 12th Mining and Learning from Time Series Workshop*, 2026. URL https://arxiv.org/abs/2607.00958. KDD MILETS 2026.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Xinlei Chen and Kaiming He. Siamsiam: A simple siamese representation learning framework. In *International Conference on Learning Representations*, 2021.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Harald Cramér and Herman Wold. Some theorems on distribution functions. *Journal of the London Mathematical Society*, s1-11(4):290–294, 1936. doi: 10.1112/jlms/s1-11.4.290.

Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljašić. Equivariant contrastive learning. In *International Conference on Learning Representations*, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Persi Diaconis and David Freedman. A dozen de Finetti-style results in search of a theory. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 23(S2):397–423, 1987.

Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 28(4):511–533, 2002.

T. W. Epps and L. B. Pulley. A test for normality based on the empirical characteristic function. *Biometrika*, 70(3):723–726, 1983. doi: 10.1093/biomet/70.3.723.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3015–3024. PMLR, 2021. URL https://proceedings.mlr.press/v139/ermolov21a.html.

Aleksandr Ermolov, Leyla Mirvakhabova, Valentin Khrulkov, Nicu Sebe, and Ivan Oseledets. Hyperbolic vision transformers: Combining improvements in metric learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7399–7409. IEEE, 2022.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016. doi: 10.1090/jams/852.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive multi-view representation learning on graphs. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4116–4126. PMLR, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2020. doi: 10.1109/CVPR42600.2020.00975.

Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from imagenet, March 2019. URL https://github.com/fastai/imagenette.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, January 1998. doi: 10.1137/S0036141096303359.

David Klee, Jung Yeon Park, Robert Platt, and Robin Walters. A comparison of equivariant vision models with ImageNet pre-training. In *NeurIPS 2023 Workshop on Symmetry and Geometry in Neural Representations*, 2023. URL https://neurips.cc/virtual/2023/84252.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yilun Kuang, Yash Dagade, Tim G. J. Rudner, Randall Balestriero, and Yann LeCun. Rectified lpjepa: Joint-embedding predictive architectures with sparse and maximum-entropy representations. *International Conference on Machine Learning (ICML)*, 2026.

Lukas Kuhn, Lucas Maes, Giuseppe Serra, Quentin Le Lidec, Yann LeCun, Randall Balestriero, and Florian Buettner. Levjepa: Efficient & scalable video pretraining without the heuristics, 2026a. URL https://arxiv.org/abs/2608.27395.

Lukas Kuhn, Giuseppe Serra, Randall Balestriero, and Florian Buettner. Levljepa: End-to-end vision-language pretraining without negatives, 2026b. URL https://arxiv.org/abs/2607.00784.

Yann LeCun. A path towards autonomous machine intelligence. OpenReview, June 2022. URL https://openreview.net/forum?id=BZ5a1r-kVsf. Version 0.9.2, June 27, 2022.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=xQUe1pOKPam.

Xin Liu, Zhongdao Wang, Ya-Li Li, and Shengjin Wang. Self-supervised learning via maximum entropy coding. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022b. doi: 10.52202/068431-2471.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworld-model: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026. URL https://arxiv.org/abs/2603.19312.

S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Sameer A. Nene, Shree K. Nayar, and Hiroshi Murase. Columbia object image library (COIL-100). Technical Report CUCS-006-96, Columbia University, 1996.

Léo Nicollier, Max Dunitz, Marc Pic, Pablo Musé, Enric Meinhardt-Llopis, and Gabriele Facciolo. Sphere-jepa: Spherical prediction with homogeneous embeddings. *arXiv preprint arXiv:2605.26900*, 2026a. doi: 10.48550/arXiv.2605.26900. URL https://arxiv.org/abs/2605.26900.

Léo Nicollier, Enric Meinhardt-Llopis, Max Dunitz, Marc Pic, Pablo Musé, and Gabriele Facciolo. Expanding sphere-jepa: A family of statistical regularizers for the hypersphere, 2026b. URL https://arxiv.org/abs/2606.17603.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Florian Rottach, Sebastian Schieferdecker, William Rudman, Randall Balestriero, and Carsten Eickhoff. Mol-jepa: A multimodal joint embedding predictive architecture for molecules, 2026. URL https://arxiv.org/abs/2608.22642.

Tristan Luca Saidi, Abigail Hickok, Bastian Rieck, and Andrew J Blumberg. Provable cluster-preserving visualizations with curvature-based stochastic neighbor embeddings. *Proceedings of the National Academy of Sciences*, 123(28):e2509171123, 2026.

Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Lió. 3d infomax improves GNNs for molecular property prediction. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20479–20502. PMLR, 2022.

Shantanu Thakoor, Corentin Tallec, Mohammad Gheshlaghi Azar, Rémi Munos, Petar Veličković, and Michal Valko. Bootstrapped representation learning on graphs. In *ICLR 2021 Workshop on Geometrical and Topological Representation Learning*, 2021.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10268–10278. PMLR, 2021. URL https://proceedings.mlr.press/v139/tian21a.html.

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

Petar Veličković, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R. Devon Hjelm. Deep graph infomax. In *International Conference on Learning Representations*, 2019.

Yifei Wang, Kaiwen Hu, Sharut Gupta, Ziyu Ye, Yisen Wang, and Stefanie Jegelka. Understanding the role of equivariance in self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Saining Xie, Jiatao Gu, Demi Guo, Charles R. Qi, Leonidas J. Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Computer Vision – ECCV 2020*, volume 12348 of *Lecture Notes in Computer Science*, pp. 574–591. Springer, 2020. doi: 10.1007/978-3-030-58580-8_34.

Ruichen Xu, Jingxiang Qu, Wenhan Gao, Jiaxing Zhang, Linsey Pang, Ravid Shwartz-Ziv, Yann LeCun, and Yuefan Deng. Hp-jepa: Hierarchical partitioning for multi-resolution graph joint-embedding predictive learning, 2026. URL https://arxiv.org/abs/2608.00491.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5812–5823. Curran Associates, Inc., 2020.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Pascal Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

Yifan Zhang, Zhiquan Tan, Jingqin Yang, Weiran Huang, and Yang Yuan. Matrix information theory for self-supervised learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59897–59918. PMLR, 2024.

Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any point-cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10252–10263, 2021.

Kai Zhao, Dongliang Nie, Yuchen Lin, Zhehan Luo, Yixiao Gu, Deng-Ping Fan, and Dan Zeng. Sub-jepa: Subspace gaussian regularization for stable end-to-end world models, 2026. URL https://arxiv.org/abs/2605.09241.

Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, 2005.

LocaLeJEPA: Maximum Entropy Representations under Geometric Priors

Supplementary Material

Contents

A Additional Implementation Details **16**

A.1 Characteristic Function of the Uniform Ball 16

A.1.1 One-Dimensional Projections 18

A.1.2 Reference Scale Used in Training 18

A.2 LocaLeJEPA Algorithm 19

A.3 Scale Equivalence Across Different Ball Radii 21

A.4 Complexity Analysis and Computational Cost 22

B Ablation Analysis **23**

C Additional Experimental Details **25**

C.1 Shared Experimental Protocol 25

C.2 Distribution-Matching Objectives and Reference Distributions 26

C.3 Toroidal Geometry Preservation 27

C.3.1 Complete Torus Results at Full Probe Budget 28

C.3.2 Probe Label-Budget Curves 29

C.3.3 Spatial Structure of Rotation Recovery 29

C.4 ImageNet-1K: Training and Evaluation Protocol 34

C.4.1 ImageNet-1K: Pretraining Details 34

C.4.2 ImageNet-1K Frozen-Feature Readouts 34

C.4.3 ImageNet-1K: Full Results 35

D Qualitative Embedding Analysis **36**

D.1 Topology Preservation on COIL Dataset 36

D.2 Patch-Level Representation Structure 37

D.3 Class Consistency of Frozen k-NN Neighborhoods 37

A Additional Implementation Details

This appendix derives the uniform ball distribution, the reference used by LocaLeReg. Section A.1 computes its characteristic function and one-dimensional marginals, including equivalent hypergeometric forms, and Section A.2 describes how they enter the local distribution-matching objective and summarizes the training procedure.

A.1 Characteristic Function of the Uniform Ball

Throughout, d denotes the output dimension of the projector, i.e. the space in which the regularizer acts. For $c \in \mathbb{R}^d$ and $R > 0$, we denote the ball of radius R centered at c as:

$$\mathbb{B}_R^d(c) = \{x \in \mathbb{R}^d : \|x - c\|_2 \leq R\} \tag{13}$$

and write $\mathbb{B}_R^d = \mathbb{B}_R^d(0)$. The volume of the unit ball is:

$$V_d = \frac{\pi^{d/2}}{\Gamma(d/2+1)} \tag{14}$$

where Γ is the Gamma function. As a consequence, the uniform distribution over the ball has the following density function.

Lemma A.1 (Uniform Ball)**.** *A random variable uniformly distributed over the ball, denoted* $X \sim \mathrm{Unif}(\mathbb{B}_R^d)$*, has density:*

$$f_X(x) = \frac{1}{V_d R^d} \mathbf{1}_{\mathbb{B}_R^d}(x) \tag{15}$$

We recall the following definition, which we then apply to the uniform ball distribution.

Definition A.2 (Characteristic Function)**.** Given a random variable X, its characteristic function is defined as

$$\varphi_X(t) = \mathbb{E}[e^{it^\top X}], \qquad t \in \mathbb{R}^d \tag{16}$$

Proposition A.3 (Bessel form of the Characteristic Function)**.** *Let* $X \sim \mathrm{Unif}(\mathbb{B}_R^d)$*. Then:*

$$\varphi_X(t) = 2^{d/2}\Gamma\left(\frac{d}{2}+1\right)\frac{J_{d/2}(R\|t\|_2)}{(R\|t\|_2)^{d/2}} \tag{17}$$

and at $t = 0$ *the expression is defined by continuity as* $\varphi_X(0) = 1$*.*

The proof of Proposition A.3 follows from the next two lemmas. Lemma A.5 is a known fact about the Bessel functions of the first kind, hence we state it here without a proof.

Lemma A.4 (Integral Representation of the Characteristic Function)**.** *Let* $X \sim \mathrm{Unif}(\mathbb{B}_R^d)$ *and* $z = R\|t\|_2$*. Then:*

$$\varphi_X(t) = \frac{\Gamma(d/2+1)}{\sqrt{\pi}\,\Gamma((d+1)/2)} \int_{-1}^{1} e^{izu}(1-u^2)^{(d-1)/2}\,du \tag{18}$$

Proof. Rotational invariance implies that:

$$\varphi_X(t) = \int_{\mathbb{B}_R^d} f_X(x) e^{it^\top x}\,dx \tag{19}$$

depends only on $\|t\|_2$. We may therefore align t with the first coordinate axis and write $t = \tau e_1$, where $\tau = \|t\|_2$. Hence:

$$\varphi_X(t) = \int_{\mathbb{B}_R^d} f_X(x) e^{i\tau x_1}\,dx \tag{20}$$

At a fixed first coordinate $x_1 = s \in [-R, R]$, the cross-section of $\mathbb{B}_R^d$ is a $(d-1)$-dimensional ball of radius $\sqrt{R^2 - s^2}$, since $s^2 + \|y\|^2 \leq R^2$ implies $\|y\| \leq \sqrt{R^2 - s^2}$. Integrating over these cross-sections gives:

$$\varphi_X(t) = \frac{1}{V_d R^d} \int_{-R}^{R} e^{i\tau s} \left[\int_{\|y\| \leq \sqrt{R^2-s^2}} dy\right] ds$$

$$= \frac{V_{d-1}}{V_d R^d} \int_{-R}^{R} e^{i\tau s} (R^2 - s^2)^{(d-1)/2}\, ds$$
$$= \frac{\Gamma(d/2+1)}{\sqrt{\pi}\,\Gamma((d+1)/2)} \int_{-1}^{1} e^{izu} (1-u^2)^{(d-1)/2}\, du,$$

where the third equality uses $s = Ru$ and $z = R\|t\|_2$. □

Lemma A.5 (Poisson Integral)**.** *For $\nu > -1/2$ and $z > 0$, the Bessel function of the first kind satisfies:*

$$J_\nu(z) = \frac{(z/2)^\nu}{\sqrt{\pi}\,\Gamma(\nu+1/2)} \int_{-1}^{1} e^{izu} (1-u^2)^{\nu-1/2}\, du \tag{21}$$

Proof of Proposition A.3. Substitute $\nu = d/2$ from Lemma A.5 into (18). □

This characteristic function admits two hypergeometric representations as alternative expressions.

Proposition A.6 (Hypergeometric forms)**.** *Let $X \sim \mathrm{Unif}(\mathbb{B}_R^d)$. Then:*

$$\varphi_X(t) = {}_0F_1\left(;\frac{d}{2}+1; -\frac{R^2\|t\|_2^2}{4}\right) \tag{22}$$

and:

$$\varphi_X(t) = e^{-iR\|t\|_2}\, {}_1F_1\left(\frac{d+1}{2}; d+1; 2iR\|t\|_2\right) \tag{23}$$

Proof. Expression in terms of ${}_0F_1$. Using the rising factorial $(a)_k = \Gamma(a+k)/\Gamma(a)$, with $(a)_0 = 1$, the Bessel series can be written as:

$$J_\nu(z) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!\,\Gamma(k+\nu+1)} \left(\frac{z}{2}\right)^{2k+\nu}$$
$$= \frac{(z/2)^\nu}{\Gamma(\nu+1)} \sum_{k=0}^{\infty} \frac{(-z^2/4)^k}{(\nu+1)_k\, k!}$$
$$= \frac{(z/2)^\nu}{\Gamma(\nu+1)}\, {}_0F_1\left(;\nu+1; -\frac{z^2}{4}\right)$$

Substitution into (17) gives (22).

Expression in terms of ${}_1F_1$. Starting from (18), let $u = 2w - 1$ and set $a = (d+1)/2$. Since $1 - u^2 = 4w(1-w)$, we obtain:

$$\varphi_X(t) = \frac{2^d \Gamma(d/2+1)}{\sqrt{\pi}\,\Gamma(a)} e^{-iz} \int_0^1 e^{2izw} w^{a-1} (1-w)^{a-1}\, dw \tag{24}$$

Kummer's integral representation gives:

$$\int_0^1 e^{2izw} w^{a-1} (1-w)^{a-1}\, dw = \frac{\Gamma(a)^2}{\Gamma(2a)}\, {}_1F_1(a; 2a; 2iz) \tag{25}$$

The gamma duplication formula implies:

$$\Gamma(d+1) = \frac{2^d}{\sqrt{\pi}} \Gamma\left(\frac{d+1}{2}\right) \Gamma\left(\frac{d}{2}+1\right) \tag{26}$$

so the prefactors cancel, yielding (23). □

Remark. Unlike the Bessel quotient (17), the expression (22) is directly defined at $t = 0$. Although (23) involves complex quantities, the characteristic function is real-valued because the reference distribution is symmetric about the origin.

A.1.1 One-Dimensional Projections

The sliced distribution-matching objective compares scalar projections of the embeddings with the corresponding projections of the reference distribution.

Proposition A.7 (Projected characteristic function)**.** *Let $X \sim \mathrm{Unif}(\mathbb{B}_R^d(c))$, fix $u \in \mathbb{S}^{d-1}$, and define $Y_u = \langle u, X\rangle$. Then:*

$$\varphi_{Y_u}(s) = e^{is\langle u,c\rangle} 2^{d/2}\Gamma\left(\frac{d}{2}+1\right)\frac{J_{d/2}(R|s|)}{(R|s|)^{d/2}}, \qquad s \in \mathbb{R} \tag{27}$$

with $\varphi_{Y_u}(0) = 1$.

Proof. Writing $X = c + X_0$, with $X_0 \sim \mathrm{Unif}(\mathbb{B}_R^d)$, gives:

$$\begin{aligned}\varphi_{Y_u}(s) &= \mathbb{E}[e^{is\langle X,u\rangle}] \\ &= e^{is\langle c,u\rangle}\,\mathbb{E}[e^{i\langle X_0, su\rangle}] \\ &= e^{is\langle u,c\rangle}\varphi_{X_0}(su)\end{aligned}$$

and (17) with $\|su\|_2 = |s|$ yields (27). □

Corollary A.8 (Common scalar reference)**.** *After centering, (27) is independent of u, so all sketch directions use the same scalar reference characteristic function:*

$$\varphi_{R,d}(s) = 2^{d/2}\Gamma\left(\frac{d}{2}+1\right)\frac{J_{d/2}(R|s|)}{(R|s|)^{d/2}}, \qquad s \in \mathbb{R}.$$

Proposition A.9 (Projected density)**.** *In the setting of Proposition A.7, write $m_u = \langle u, c\rangle$. Then*

$$f_{Y_u}(y) = \frac{\Gamma(d/2+1)}{R\sqrt{\pi}\,\Gamma((d+1)/2)}\left(1 - \frac{(y-m_u)^2}{R^2}\right)^{(d-1)/2}, \qquad |y - m_u| < R \tag{28}$$

and $f_{Y_u}(y) = 0$ outside this interval.

Proof. This follows from the cross-section argument in the proof of Lemma A.4. Integrating this density against e^{isy} recovers (27). □

Remark. For the centered unit ball, the hypergeometric expressions of Proposition A.6 reduce to:

$$\varphi_{Y_u}(s) = {}_0F_1\left(;\frac{d}{2}+1;-\frac{s^2}{4}\right) = e^{-i|s|}\,{}_1F_1\left(\frac{d+1}{2};d+1;2i|s|\right) \tag{29}$$

A.1.2 Reference Scale Used in Training

LocaLeReg centers each cluster and matches the resulting residuals to a common uniform ball distribution reference. The radius is chosen to give identity covariance.

Lemma A.10 (Covariance of the uniform ball)**.** *If $X_0 \sim \mathrm{Unif}(\mathbb{B}_R^d)$, then:*

$$\mathrm{Cov}(X_0) = \frac{R^2}{d+2} I_d \tag{30}$$

Proof. This follows from rotational symmetry and $\mathbb{E}\|X_0\|_2^2 = dR^2/(d+2)$. □

Corollary A.11 (LocaLeReg reference)**.** *The training target is:*

$$P_{\mathrm{ball}} = \mathrm{Unif}\left(B^d_{\sqrt{d+2}}\right) \tag{31}$$

which has identity covariance and whose one-dimensional projections have characteristic function:

$$\varphi_d(s) = 2^{d/2}\Gamma\left(\frac{d}{2}+1\right)\frac{J_{d/2}(\sqrt{d+2}\,|s|)}{(\sqrt{d+2}\,|s|)^{d/2}}, \qquad s \in \mathbb{R} \tag{32}$$

Proof. Set $R = \sqrt{d+2}$ in Lemma A.10 and Corollary A.8. □

For the $d = 16$ projector used in our experiments, the target radius is $\sqrt{18}$ (see Section A.3 for more discussion on the scale equivalence of target radius).

A.2 LocaLeJEPA Algorithm

LocaLeJEPA combines the multi-view prediction objective with distribution matching to a uniform ball within learned neighborhoods. The computation consists of constructing a shared neighborhood partition, centering each neighborhood independently, and averaging a sliced Epps–Pulley discrepancy over sufficiently populated neighborhoods.

Neighborhood Construction and Centering. For each image, we compute a reference embedding $\bar{z}_i$ by averaging its projected reference views. Cosine-similarity K-means is then applied to these reference embeddings, using at most ten iterations. This produces a partition $\{C_q\}_{q=1}^{K}$ of the images, shared across all views: every view of a given image belongs to the same neighborhood.

For a neighborhood C_q containing $n_q = |C_q|$ images, we compute a separate mean within each view and form the residuals:

$$\mu_q^{(v)} = \frac{1}{n_q} \sum_{i \in C_q} z_i^{(v)}, \qquad r_i^{(v)} = z_i^{(v)} - \mu_q^{(v)} \tag{33}$$

Gradients propagate through both the projected embeddings and their cluster mean, while the neighborhood assignments remain detached. The residuals are matched to

$$P_{\text{ball}} = \text{Unif}\left(B^d_{\sqrt{d+2}}\right) \tag{34}$$

where d is the projector dimension. This reference has identity covariance, and centering allows the regularizer to constrain the distribution within each neighborhood without driving all neighborhood means to a common origin.

Sliced Distribution Matching. For a direction $u \in \mathbb{S}^{d-1}$, the empirical characteristic function of the projected residuals is:

$$\widehat{\varphi}_{q,v,u}(s) = \frac{1}{n_q} \sum_{i \in C_q} e^{is\langle u, r_i^{(v)} \rangle} \tag{35}$$

Let $\varphi_d(s)$ denote the characteristic function of a one-dimensional projection of P_{ball}, as in (32). Using S directions $\{u_\ell\}_{\ell=1}^{S}$ sampled uniformly from $\mathbb{S}^{d-1}$, we compute:

$$D_{q,v} = \frac{n_q}{S} \sum_{\ell=1}^{S} \int_{\mathbb{R}} \left|\widehat{\varphi}_{q,v,u_\ell}(s) - \varphi_d(s)\right|^2 w(s)\, ds \tag{36}$$

We use $S = 256$ directions and the Gaussian frequency-weighting function $w(s) = e^{-s^2/2}$. The integral is approximated over $[-3, 3]$ using the trapezoidal rule, exploiting symmetry to evaluate 17 equally spaced frequencies on $[0, 3]$. For a fixed projector dimension, the same reference characteristic function is used for every cluster and view.

Minimum Neighborhood Size. Although the Epps–Pulley discrepancy is defined for any nonempty sample, very small neighborhoods provide limited information about the shape of a distribution. Centering further constrains their empirical geometry. For instance, a singleton becomes identically zero, while a two-point neighborhood reduces to a pair of opposite residuals. Consequently, a discrepancy computed from very few points need not reliably distinguish the uniform ball target from alternative distributions. We therefore exclude neighborhoods containing fewer than four images. This is a practical minimum-size threshold, rather than a theoretical condition for the statistic to be defined or a guarantee of reliable distribution discrimination.

Sensitivity to Sample Size and Dimension. To characterize this finite-sample limitation, we compare the statistic on independent samples drawn either from the uniform ball of radius $\sqrt{d+2}$ or from $\mathcal{N}(0, I_d)$. The two distributions have the same covariance, so the comparison tests sensitivity to differences beyond their first two population moments. Samples are evaluated directly, without clustering. For each setting, we perform 200 trials and summarize the separation between the resulting statistic values as:

$$\text{sep} = \frac{\bar{D}_{\text{gauss}} - \bar{D}_{\text{ball}}}{\sqrt{(s^2_{\text{gauss}} + s^2_{\text{ball}})/2}} \tag{37}$$

where $\bar{D}$ and s denote the empirical mean and standard deviation of the sliced EP discrepancy across trials. As such, a positive value indicates that the ball receives a lower discrepancy on average.

At $d = 16$, Table 4a shows that separation with centering increases from 0.81 at the minimum accepted size $n = 4$ to 1.10 at $n = 8$ and 3.54 at $n = 256$. The corresponding uncentered values are lower, with separation remaining negative in the reported trials for $n \leq 32$ and reaching 0.66 at $n = 256$. Thus, centering improves separation in this experiment, but the minimum accepted neighborhood size still provides only modest discrimination between these covariance-matched targets.

Table 4b additionally shows that discrimination weakens as the projector dimension increases. At $n = 256$, separation decreases from 6.29 at $d = 4$ to 3.54 at $d = 16$, 0.91 at $d = 64$, and 0.16 at $d = 256$. This is expected: the one-dimensional marginal of the covariance-normalized uniform ball on $\mathbb{R}^d$ has density proportional to $\left(1 - t^2/(d+2)\right)^{(d-1)/2}$, which is also the one-dimensional marginal of the uniform distribution on the $(d+1)$-dimensional sphere of radius $\sqrt{d+2}$ in $\mathbb{R}^{d+2}$, and therefore converges to the standard Gaussian as d grows.(Diaconis & Freedman, 1987).

Together, these observations expose a tradeoff between locality and statistical resolution. Increasing K produces finer neighborhoods but reduces their typical size, approximately N/K for a balanced partition of a batch of N images. Larger neighborhoods provide more information for distribution matching, whereas the four-image threshold only excludes the smallest neighborhoods. Our sensitivity study characterizes this tradeoff under independent sampling, but it does not establish an optimal cluster size for learned neighborhoods.

Table 4: *Sample size, Centering and Projector Dimension Sensitivity.* In this table, we show the sensitivity of the sliced Epps–Pulley discrepancy to sample size, centering, and projector dimension, with 200 trials per setting. "Ball" and "Gaussian" report the mean discrepancy $\bar{D}_{\text{ball}}$ and $\bar{D}_{\text{gauss}}$ for samples drawn from $\text{Unif}(B^d_{\sqrt{d+2}})$ and $\mathcal{N}(0, I_d)$, respectively. Both are evaluated against the same uniform ball distribution target. "Centered" denotes subtraction of the empirical sample mean before evaluating the discrepancy. (a) Sample size n at $d = 16$, with and without centering. (b) Projector dimension d at $n = 256$, with centering. "sep." is the standardized difference defined in Equation (37); pooled s is its denominator.

(a) Cluster size and centering, $d = 16$.

	centered			uncentered		
n	ball	Gaussian	sep.	ball	Gaussian	sep.
2	0.505	0.515	0.37	1.067	1.045	−0.29
4	0.533	0.551	0.81	1.071	1.042	−0.33
8	0.543	0.570	1.10	1.071	1.060	−0.11
16	0.551	0.580	1.23	1.076	1.061	−0.15
32	0.551	0.589	1.52	1.084	1.065	−0.20
64	0.553	0.605	2.04	1.072	1.080	0.08
128	0.554	0.629	2.64	1.078	1.098	0.21
256	0.555	0.678	3.54	1.081	1.146	0.66

(b) Embedding dimension, $n = 256$.

d	ball	Gaussian	pooled s	sep.
4	0.532	1.596	0.169	6.29
16	0.555	0.678	0.035	3.54
64	0.564	0.580	0.018	0.91
256	0.569	0.572	0.017	0.16

Aggregation and Training Objective. Let $\mathcal{Q} = \{q : |C_q| \geq 4\}$ index the retained neighborhoods. When $\mathcal{Q}$ is nonempty, the regularizer is averaged uniformly over retained neighborhoods and views:

$$\mathcal{L}_{\text{LocaLeReg}} = \frac{1}{|\mathcal{Q}|V} \sum_{q \in \mathcal{Q}} \sum_{v=1}^{V} D_{q,v} \tag{38}$$

If no neighborhood meets the minimum size, we instead apply the same centered uniform ball distribution regularizer to the complete batch, separately within each view. The training objective is:

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_{\text{pred}} + \lambda \mathcal{L}_{\text{LocaLeReg}} \tag{39}$$

Algorithm 1 summarizes the computation. The neighborhood count K is either fixed or follows the epoch-dependent schedule described in Section C.1.

Algorithm 1 LocaLeJEPA objective with Uniform-ball local regularization

Require: Projected embeddings $\{z_i^{(v)}\}_{i=1,\dots,N;\, v=1,\dots,V} \subset \mathbb{R}^d$
Require: Neighborhood count K, weight λ, sketch count S, minimal neighborhood size $n_{\min}$
Compute each image's reference embedding $\bar{z}_i$
$\mathcal{L}_{\text{pred}} \leftarrow \dfrac{1}{NVd} \displaystyle\sum_{i=1}^{N} \sum_{v=1}^{V} \|z_i^{(v)} - \bar{z}_i\|_2^2$
Compute a shared partition using K-means:

$$\{C_1, \dots, C_K\} \leftarrow \text{KMeans}\left(\{\text{stopgrad}(\bar{z}_i)\}_{i=1}^{N}, K\right)$$

$\mathcal{C} \leftarrow \{C_q : |C_q| \geq n_{\min}\}$
if $\mathcal{C} = \varnothing$ **then**
    $\mathcal{C} \leftarrow \{\{1, \dots, N\}\}$ ▷ Fall back to the complete batch
end if
$P_{\text{ball}} \leftarrow \text{Unif}(B^d_{\sqrt{d+2}})$
$\mathcal{L}_{\text{LocaLeReg}} \leftarrow 0$
for each neighborhood $C \in \mathcal{C}$ **do**
    for each view $v = 1, \dots, V$ **do**
        $\mu_C^{(v)} \leftarrow \dfrac{1}{|C|} \displaystyle\sum_{i \in C} z_i^{(v)}$
        $R_C^{(v)} \leftarrow \{z_i^{(v)} - \mu_C^{(v)} : i \in C\}$
        $\mathcal{L}_{\text{LocaLeReg}} \leftarrow \mathcal{L}_{\text{LocaLeReg}} + D_{q,v}$
    end for
end for
$\mathcal{L}_{\text{LocaLeReg}} \leftarrow \dfrac{\mathcal{L}_{\text{LocaLeReg}}}{V|\mathcal{C}|}$
return $(1-\lambda)\mathcal{L}_{\text{pred}} + \lambda \mathcal{L}_{\text{LocaLeReg}}$

A.3 Scale Equivalence Across Different Ball Radii

As we have already mentioned, in our experiments we set $R = \sqrt{d+2}$, so that the target is whitened, i.e. has identity covariance. In fact, as we will see next, this choice is more a unit of length rather than a hyperparameter.

Lemma A.12. *For a random vector $X \in \mathbb{R}^d$ and a unit vector u, let $\varphi_X(t) = \mathbb{E}\,[e^{\mathrm{i}tu^\top X}]$, $t \in \mathbb{R}$. Then for every $\sigma > 0$ and $t \in \mathbb{R}$:*

$$\varphi_X(t/\sigma) = \varphi_{X/\sigma}(t) \tag{40}$$

In particular, for $\sigma = R/\sqrt{d+2}$, $\varphi_{d,R}(t/\sigma) = \varphi_{d,\sqrt{d+2}}(t)$, and if the characteristic functions in the Epps–Pulley statistic are evaluated at t/σ instead of t, with the Gaussian weights unchanged, the statistic at radius R on residuals r equals the statistic at $\sqrt{d+2}$ on r/σ.

Proof. The identity follows from the definition: $\mathbb{E}\left[e^{\mathrm{i}(t/\sigma)\,u^\top X}\right] = \mathbb{E}\left[e^{\mathrm{i}t\,u^\top (X/\sigma)}\right]$.

For $\sigma = R/\sqrt{d+2}$ and $X \sim \text{Unif}\left(B_R^d\right)$ we have $X/\sigma \sim \text{Unif}\left(\mathbb{B}^d_{\sqrt{d+2}}\right)$, so the identity gives $\varphi_{d,R}(t/\sigma) = \varphi_{d,\sqrt{d+2}}(t)$. Applied to the empirical distribution of the residuals, it gives $\widehat{\varphi}_r(t/\sigma) = \widehat{\varphi}_{r/\sigma}(t)$, so the two statistics have the same integrand at every t and u, and the Gaussian weights depend on neither R nor r. □

The invariance loss also has to be rescaled accordingly by dividing it by σ^2:

Corollary A.13. *Let $\sigma = R/\sqrt{d+2}$, and let $\mathcal{L}^R_{\text{LocaLe}-\text{Reg}}$ evaluate the characteristic functions at t/σ as in Lemma A.12. For every batch of embeddings z:*

$$\mathcal{L}^R_{\text{LocaLe}-\text{Reg}}(z) = \mathcal{L}^{\sqrt{d+2}}_{\text{LocaLe}-\text{Reg}}(z/\sigma), \qquad \mathcal{L}_{\text{pred}}(z) = \sigma^2\, \mathcal{L}_{\text{pred}}(z/\sigma) \tag{41}$$

Consequently, $\mathcal{L}_R = (1-\lambda)\,\sigma^{-2}\,\mathcal{L}_{\text{pred}} + \lambda\,\mathcal{L}^R_{\text{LocaLe}-\text{Reg}}$ satisfies $\mathcal{L}_R(z) = \mathcal{L}_{\sqrt{d+2}}(z/\sigma)$.

Proof. Since k-means assignments do not change when all points are rescaled by the same factor, z and z/σ give the same clusters, with residuals divided by σ; Lemma A.12 applied to every cluster gives the first identity.

For the second, $\bar{z}_i/\sigma$ is the mean of the reference views of z/σ, so:

$$\mathcal{L}_{\mathrm{pred}}(z/\sigma) = \frac{1}{NVd}\sum_{i=1}^{N}\sum_{v=1}^{V}\left\|\frac{z_i^{(v)}}{\sigma} - \frac{\bar{z}_i}{\sigma}\right\|^2 = \frac{\sigma^{-2}}{NVd}\sum_{i=1}^{N}\sum_{v=1}^{V}\|z_i^{(v)} - \bar{z}_i\|^2 = \sigma^{-2}\,\mathcal{L}_{\mathrm{pred}}(z) \quad (42)$$

Multiplying the second identity by $(1-\lambda)\sigma^{-2}$, the first by λ, and adding gives the last statement. □

The effect of the two scale corrections is examined in a radius ablation (Table 5). With both corrections, the run at $R = 1$ solves the same problem as the run at $R = \sqrt{18}$ (Corollary A.13), and the two agree closely. The gap of 4 points in probe accuracy is not explained by the objective, which is identical, and we attribute it to optimization. With only the frequencies rescaled, the objective is the one at $\sqrt{18}$ with λ replaced by $\lambda/(\lambda + (1-\lambda)\sigma^2) = 0.818$. The regularizer is matched better than at $\sqrt{18}$, but the normalized invariance loss is 2.9 times larger and the probe accuracy drops to 20.5. Without either correction, the two errors partly cancel: the invariance loss is weighted $\sigma^{-2} = 18$ times less, but the regularizer evaluated on the compressed frequency range is about 90 times weaker against collapse. The representation does not collapse and its invariance is close to that at $\sqrt{18}$, but the probe accuracy is 16 points lower. All runs use a single seed.

Table 5: *Radius Comparison.* In this table, we show the Radius comparison over Imagenette, $K = 4$, $N = 256$, $\lambda = 0.2$, 100 epochs. Freq.: characteristic functions evaluated at t/σ (Lemma A.12); σ^{-2}: invariance loss divided by σ^2 (Corollary A.13). At $R = \sqrt{18}$, $\sigma = 1$ and both corrections are the identity. *Without frequency rescaling, the discrepancy for standardized embeddings is evaluated at frequencies σt_j instead of t_j, while the Gaussian quadrature weights remain defined on t_j. Thus the regularizer values are not directly comparable.

R	freq.	σ^{-2}	probe	$\mathcal{L}_{\mathrm{pred}}/\sigma^2$	$\mathcal{L}_{\mathrm{LocaLe-Reg}}$	$\overline{\Sigma_{ii}}/\sigma^2$
$\sqrt{18}$	$\sigma = 1$		63.7	0.249	0.654	1.08
1	–	–	47.5	0.294	0.021*	0.95
1	✓	–	20.5	0.722	0.566	1.03
1	✓	✓	59.7	0.283	0.668	1.07

A.4 Complexity Analysis and Computational Cost

LocaLeJEPA evaluates its distribution-matching regularizer independently within each neighborhood, introducing additional computation through clustering and cluster-specific statistics.

Settings. Let N denote the batch size, d the projector dimension, K the number of neighborhoods, I the number of K-means iterations, and S the number of sketch directions. Because the neighborhoods form a disjoint partition of the batch, the distribution statistic itself remains linear in the total number of samples.

Theoretical Time and Memory Complexity. Relative to global SIGReg, the additional cost is dominated by K-means and by the cluster-specific sketch matrices, yielding:

$$O(INKd + KdS) \quad \text{time}, \qquad O(NK + KdS) \quad \text{memory} \quad (43)$$

where the NK term corresponds to the K-means distance matrix and KdS to the sketch matrices retained for backpropagation. The additional cost is therefore linear in K and, for fixed K, linear in N.

Table 6: *End-to-end Training Cost.* Wall-clock time per step on one A100-40GB for ViT-S/8 (128×128, V=4, d=16, N=128, K=8, S=256, I=10), averaged over 300 repeated steps.

	s/step	vs. SIGReg
SIGReg	0.263	—
Local SIGReg	0.277	+5.3%
LocaLeJEPA	0.278	+5.7%

Empirical Time and Memory Complexity. At the representative training configuration in Table 6, local SIGReg and LocaLeJEPA increase end-to-end step time by only 5.3% and 5.7%, respectively,

Table 7: *Regularizer Cost as a Function of Batch and Neighborhood Size.* Forward–backward runtime and peak allocated memory on one A100-40GB in FP32. Values are medians over three seeds with V=4, d=16, S=256, I=10.

	Time (ms)			Memory (MiB)	
	SIGReg	Local SIGReg	LocaLeJEPA	SIGReg	Local
$K = 8$					
$N = 32$	1.24	8.89	9.97	12.8	5.4
$N = 64$	1.33	13.14	14.86	25.5	10.4
$N = 128$	1.27	14.57	16.39	51.0	18.6
$N = 256$	1.24	16.25	18.02	105.0	33.2
$N = 512$	1.54	16.24	18.12	203.9	62.0
$N = 256$					
$K = 1$	1.28	3.18	3.33	108.0	108.0
$K = 2$	1.25	8.63	9.03	108.0	67.2
$K = 4$	1.27	11.22	12.22	108.0	45.6
$K = 8$	1.28	16.63	18.19	108.0	33.2
$K = 16$	1.24	25.77	29.67	108.0	28.8
$K = 32$	1.29	44.41	51.59	108.0	26.9

relative to global SIGReg. Peak reserved GPU memory is 22.3 GiB for all three methods, indicating that the additional regularizer state is small relative to the memory required by the backbone activations. The cost is more visible when isolating the regularizer itself. At N=256 and K=8, LocaLeJEPA requires 18.02 ms for a forward–backward pass, compared with 1.24 ms for global SIGReg (Table 7). Runtime increases with K, consistent with the K-dependent terms in (43), while its dependence on N is comparatively weak over the tested range. Interestingly, increasing locality does not increase the peak memory of the regularizer. At N=256, allocated memory decreases from 108.0 MiB at K=1 to 26.9 MiB at K=32, and is lower than global SIGReg for every $K \geq 2$. This occurs because the largest intermediate tensors scale with the size of an individual cluster rather than with the full batch. In end-to-end training, however, these differences are dominated by the backbone activations and do not affect peak reserved GPU memory.

B Ablation Analysis

In this section, we ablate three hyperparameters of LocaLeJEPA:

- λ, the regularization weight;
- K, the requested number of neighborhoods formed by K-means clustering;
- B, the per-step batch size on which clustering and regularization are performed.

We also report $\widehat{K}$, the effective number of neighborhoods remaining after discarding clusters with fewer than four samples. Unlike the quantities above, $\widehat{K}$ is a measured outcome rather than a hyperparameter. Gold, silver and bronze denote the three highest distinct reported mean accuracies within each ranked table; ties at the reported precision share a color.

Training setup. We train ViT-S/8 on Imagenette at 128×128 resolution for 800 epochs. All LocaLeJEPA runs use a learning rate of 10^{-3}, ten warm-up epochs, cosine decay to 10^{-6}, gradient clipping at 1.0, and an effective optimizer batch size of 256. We compare per-step batch sizes of 64 and 256, with four and one gradient-accumulation steps, respectively, denoted by 64×4 and 256×1. We report final-epoch online linear-probe validation accuracy as mean $\pm$ standard deviation over three seeds.

The $\lambda \times K$ grid at a fixed batch size, $B = 64 \times 4$. Table 8 evaluates regularization weights $\lambda \in \{0.02, 0.05, 0.1, 0.2, 0.3\}$ and requested neighborhood counts $K \in \{1, 2, 4, 8, 16, 32\}$ using the 64×4 batch configuration. The highest mean accuracy is obtained with $K = 2$ at both $\lambda = 0.05$ and $\lambda = 0.1$, followed by $K = 1$ and $K = 4$ at $\lambda = 0.1$. At each fixed λ, the best local configuration

($K > 1$) has a higher reported mean accuracy than the global control ($K = 1$), except at $\lambda = 0.02$, where their rounded means are equal (Table 9). These differences vary in magnitude and do not, by themselves, establish statistical significance.

Table 8: *Grid ablations $\lambda \times K$, on Imagenette, 800 epochs, $B = 64 \times 4$.* Accuracy (%): mean ± SD, $n = 3$. † flags requested mean cluster size below four.

λ	$K = 1$	$K = 2$	$K = 4$	$K = 8$	$K = 16$	$K = 32$
0.02	90.28 ± 0.42	90.28 ± 0.18	52.00 ± 5.98	30.37 ± 11.25	26.72 ± 2.97	18.63 ± 5.87 †
0.05	91.08 ± 0.14	91.52 ± 0.13	91.21 ± 0.14	86.84 ± 2.46	64.20 ± 14.80	61.04 ± 9.43 †
0.1	91.38 ± 0.34	91.52 ± 0.24	91.35 ± 0.17	90.78 ± 0.31	88.56 ± 0.16	31.43 ± 2.78 †
0.2	90.32 ± 0.27	90.43 ± 0.41	90.82 ± 0.36	90.45 ± 0.13	89.09 ± 0.48	80.96 ± 0.97 †
0.3	87.76 ± 0.09	88.32 ± 0.15	89.32 ± 0.11	89.11 ± 0.26	88.16 ± 0.46	81.03 ± 0.11 †

Table 9: *Best fixed local arm versus its $K = 1$ control at each λ, $B = 64 \times 4$.* Accuracy (%): mean ± SD, $n = 3$. Δ denotes local minus control accuracy in percentage points; "Chosen K" identifies the best local setting. Bold highlights the highest mean in each row, including ties at the displayed precision.

λ	$K = 1$	Best $K > 1$	Chosen K	Δ
0.02	$\mathbf{90.28 \pm 0.42}$	$\mathbf{90.28 \pm 0.18}$	2	-0.01
0.05	91.08 ± 0.14	$\mathbf{91.52 \pm 0.13}$	2	$+0.44$
0.1	91.38 ± 0.34	$\mathbf{91.52 \pm 0.24}$	2	$+0.14$
0.2	90.32 ± 0.27	$\mathbf{90.82 \pm 0.36}$	4	$+0.50$
0.3	87.76 ± 0.09	$\mathbf{89.32 \pm 0.11}$	4	$+1.56$

Increasing K sharply reduces accuracy under weak regularization. Larger regularization weights mitigate this decline over part of the grid, although substantial degradation remains at high K. In particular, requesting $K = 32$ from a batch of 64 would induce $64/32 = 2$ samples per cluster for a balanced clustering, below the minimum cluster size of four. Consequently, many clusters may be discarded, leaving only a few neighborhoods contributing to the regularization loss. If no cluster contains at least four samples, the loss falls back to regularizing the entire per-step batch. This filtering provides a plausible contributing mechanism for the poor performance at large requested K, but the reported results do not establish it as the sole cause. Table 10 reports the realized neighborhood counts $\widehat{K}$ averaged over the final ten training epochs.

Table 10: *Reported realized neighbourhood counts at $\lambda = 0.1$.* Fixed-K counts average the last ten epochs; ramp entries describe the terminal part of the schedule.

Batch	$K = 1$	$K = 2$	$K = 4$	$K = 8$	$K = 16$	$K = 32$	fwd	rev
64×4	1.0	2.0	3.9	6.9	8.7	5.7	9.0	2.0
256×1	1.0	2.0	4.0	7.9	15.4	27.8	15.5	2.0

Dynamic neighbourhood scheduling across λ at $B = 64 \times 4$. Table 11 examines how increasing or decreasing the requested number of neighborhoods affects performance across regularization weights λ. The forward schedule increases K from 2 to 16, whereas the reverse schedule decreases it from 16 to 2.

The forward schedule at $\lambda = 0.1$ achieves the highest mean accuracy among the $B = 64 \times 4$ configurations, at $91.57 \pm 0.20\%$. However, it exceeds the best fixed local configuration by only 0.05 percentage points and the best global control by 0.19 percentage points; these summaries alone do not establish a statistically significant improvement. Schedule direction has a much larger effect under weak regularization, as the forward schedule outperforms the reverse schedule by 47.38 percentage points at $\lambda = 0.02$ and 18.88 percentage points at $\lambda = 0.05$. The gap narrows at stronger regularization weights, with the reverse schedule achieving a slightly higher mean at $\lambda = 0.3$. These results favor starting with fewer neighborhoods when regularization is weak, but do not establish a universally preferable schedule direction.

Table 11: *Forward and reverse ramps, $B = 64 \times 4$, 800 epochs.* Accuracy (%): mean ± SD, $n = 3$.

Schedule	$\lambda = 0.02$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$
$2 \to 16$ (fwd)	87.31 ± 0.25	91.20 ± 0.22	91.57 ± 0.20	90.45 ± 0.25	88.76 ± 0.23
$16 \to 2$ (rev)	39.93 ± 6.75	72.32 ± 5.20	89.04 ± 0.43	89.48 ± 0.49	88.84 ± 0.27

Ablation in K for $\lambda = 0.1$ and $B = 256 \times 1$. We further investigate the role of locality by increasing the per-step batch size from $B = 64 \times 4$ to $B = 256 \times 1$. The effective batch size remains unchanged, but clustering now operates on four times as many samples at each step. We fix $\lambda = 0.1$, a competitive regularization strength in the preceding ablations, and vary the number of training clusters $K \in \{2, 4, 8, 16, 32\}$. As shown in Table 12, accuracy increases from $88.93 \pm 0.59\%$ at $K = 2$ to $89.74 \pm 0.38\%$ at $K = 8$, before decreasing slightly to $89.64 \pm 0.19\%$ and $89.26 \pm 0.46\%$ at $K = 16$ and $K = 32$, respectively. Compared with the 64×4 setting, the best observed local mean shifts from $K = 2$ to $K = 8$, although it is lower (89.74% versus 91.52%). The clearest difference is the robustness to fine partitions: at $K = 32$, the 64×4 setting reaches only $31.43 \pm 2.78\%$, whereas the 256×1 setting retains $89.26 \pm 0.46\%$. In sum, with the larger per-step batch, the highest mean accuracies are attained at $K = 8$ and $K = 16$, suggesting that finer local partitions become more effective when each clustering step has access to more samples.

Table 12: *Neighborhood-count ablation on Imagenette, 800 epochs, $B = 256 \times 1$.* Final-epoch online linear-probe accuracy (%): mean ± SD, $n = 3$.

λ	$K = 2$	$K = 4$	$K = 8$	$K = 16$	$K = 32$
0.1	88.93 ± 0.59	89.42 ± 0.46	89.74 ± 0.38	89.64 ± 0.19	89.26 ± 0.46

Reference baselines under their published recipes with $B = 64 \times 4$. Under their respective training recipes, LeJEPA and RDMReg achieve mean final-epoch online linear-probe accuracies of 89.29% and 81.90%, respectively. Although their performance is worse than that of the local models ablated above, their differences from the local grid cannot be attributed to locality or target distribution alone, since optimizer settings, regularization weights and centering choices differ. Moreover, the $K = 1$ LocaLeJEPA control is not equivalent to SIGReg, as we are matching to the uniform distribution in the former, instead of the isotropic Gaussian.

Table 13: *Complete reference runs at $B = 64 \times 4$, 800 epochs.* Final accuracy (%), mean ± SD over three seeds. These use their own published recipes, not the controlled local-grid recipe.

Method	Final-epoch accuracy
LeJEPA (SIGReg)	89.29 ± 0.39
RDMReg	81.90 ± 0.52

C Additional Experimental Details

This appendix complements the Section 5 with the full experimental specification underlying the reported results. We first consolidate the experimental conventions shared across studies in the Section C.1, including the representation interface, training and checkpointing conventions, and frozen-feature evaluation protocol. We then provide some more details on the baselines employed in the Section C.2, the experiment-specific settings and analyses for the torus geometry-preservation study in Section C.3, followed by the ImageNet-1K experiments across Euclidean, hyperbolic, and equivariant representation spaces in the Section C.4.

C.1 Shared Experimental Protocol

Across all experiments, we adopt a common interface between self-supervised pretraining and downstream evaluation. The encoder produces a 512-dimensional representation, which is kept

frozen for all downstream readouts, while the self-supervised regularizer is applied after a 2048–2048–16 MLP projector. This separation allows us to evaluate the information retained by the learned representation independently of the lower-dimensional space in which distribution matching is enforced.

Training Details. We follow the same batching convention throughout. Each image contributes four views, and gradients are accumulated over four forward passes of 64 images, so that 256 images contribute to each optimizer update. Local neighborhood assignments and distribution statistics, however, are computed independently within each 64-image forward pass. Experiment-specific optimization choices, including the number of training epochs, learning rate, and regularization parameters, are reported in the corresponding sections below.

Online Evaluation and Checkpointing. Following Balestriero & LeCun (2025), we train a lightweight classifier on detached encoder features during pretraining as an online diagnostic. Its gradients never propagate through the encoder or projector and therefore do not affect representation learning, meaning that therefore, the online classifier is not used for model selection. Each model is trained for the complete prescribed schedule, and all downstream evaluations use the *final saved checkpoint*, being the one at the *last epoch of training*. Thus, Imagenette experiments use the checkpoint obtained after 800 epochs, while ImageNet-1K experiments use the epoch-100 checkpoint. This rule is applied uniformly across global baselines, LocaLeJEPA configurations, dynamic-K variants, and ablations over K, λ, and the local reference distribution.

Frozen-feature Evaluation. All downstream evaluations keep the pretrained encoder fixed and fit only the specified readout. Reported differences therefore reflect information already present in the frozen representation rather than adaptation of the encoder to a particular downstream task.

Dynamic Neighborhood Schedule. In the neighbouring count, we also consider a *dynamic variant* in which the locality of the regularizer is progressively refined during pretraining. At epoch e, the number of neighborhoods is:

$$K(e) = \min\left(K_{\max}, K_{\text{start}} + \left\lfloor \frac{e}{E_{\text{step}}} \right\rfloor\right) \tag{44}$$

with K_{start}=1 throughout. Training therefore begins with a single neighborhood and introduces progressively finer local partitions as the representation develops. We use the same scheduling rule across experiments, while $K_{\max}$ and E_{step} are adapted to the corresponding pretraining horizon and specified in each experiment-specific section.

C.2 Distribution-Matching Objectives and Reference Distributions

Our comparisons span distribution-matching objectives that differ in both the domain on which regularization is imposed and the reference distribution toward which representations are driven. We make this important distinction in our evaluation because LocaLeJEPA changes the regularization from a single global constraint to a collection of locally centered distribution-matching problems, while preserving the same self-supervised prediction objective.

Global baselines. *SIGReg* (Balestriero & LeCun (2025)), the regularizer underlying LeJEPA, matches projected representations globally to an isotropic Gaussian. *SUSReg* (Nicollier et al. (2026a)) instead promotes uniformity on the hypersphere, providing a reference distribution adapted to spherical embeddings. *SubReg* (Zhao et al. (2026)) applies Gaussian regularization within frozen random feature subspaces rather than directly over the complete projected representation. Finally, *RDMReg*, introduced by Rectified LpJEPA (Kuang et al. (2026)), matches rectified representations to a generalized-Gaussian target; we use p=1, corresponding to the rectified Laplace distribution.

Except for RDMReg, these objectives can be written as:

$$\mathcal{L} = (1-\lambda)\,\mathcal{L}_{\text{pred}} + \lambda\,\mathcal{R} \tag{45}$$

where $\mathcal{L}_{\text{pred}}$ aligns the representations of the different views of each image and $\mathcal{R}$ denotes the corresponding distribution-matching regularizer. RDMReg instead follows the weighting proposed by its authors:

$$\mathcal{L}_{\text{RDM}} = 25\,\mathcal{L}_{\text{pred}} + 125\,\mathcal{R}_{\text{RDM}} \tag{46}$$

whose coefficients are absolute and therefore not directly comparable to the mixing parameter λ used by the other objectives.

LocaLeJEPA and Local Control. LocaLeJEPA replaces the global reference constraint with distribution matching within learned neighborhoods. For each view, projected representations are partitioned by cosine-similarity K-means using at most ten iterations. In order to avoid uninformative clusters, we discard clusters containing fewer than four samples, and the regularizer is averaged uniformly across the remaining clusters. Each cluster is centered independently, and the resulting residuals are matched to the uniform distribution on the d-dimensional Euclidean ball of radius:

$$r = \sqrt{d+2} \tag{47}$$

which has identity covariance. With the d=16 projector used throughout our experiments, this gives $r = \sqrt{18}$.

We additionally evaluate *LocaLeJEPA-SIGReg* as a local control. It uses the same neighborhood partition as LocaLeJEPA but replaces the centered bounded reference with Gaussian matching on the uncentered cluster representations. Consequently, this comparison tests whether local partitioning alone is sufficient to explain the gains of LocaLeJEPA, while not separately isolating the effects of centering and reference distribution.

C.3 Toroidal Geometry Preservation

This section specifies the experimental choices particular to the torus geometry-preservation study of Figure 4. We describe the Imagenette pretraining configuration and circular-regression protocol, followed by the complete ablations over neighborhood granularity, regularization strength, and probe-label budget. Unless stated otherwise, training, checkpointing, and frozen-feature evaluation follow Section C.1, while the compared regularizers are defined in Section C.2.

Imagenette Pretraining. All encoders evaluated in Table 15 are pretrained from scratch for 800 epochs on Imagenette, the ten-class ImageNet subset (Howard (2019)), using ViT-S/8 variants with 384 hidden dimensions. Global SIGReg (Balestriero & LeCun, 2025), SUSReg (Nicollier et al., 2026a), and SubReg (Zhao et al., 2026) use a learning rate of 2×10^{-3} and one warmup epoch, with λ=0.02. RDMReg follows the same global optimization recipe, with the objective weighting defined in the Section C.2. LocaLeJEPA and its local control use a learning rate of 10^{-3}, ten warmup epochs, and gradient clipping at 1.0. We report the corresponding values of λ and K for each specific configuration. For the dynamic-neighborhood variant, we instantiate the schedule of the Section C.1 with $K_{\max} = 8$ and $E_{\text{step}} = 100$, so that the partition evolves from K=1 to K=8, reaching its maximum granularity at epoch 700. All models are evaluated from their final epoch-800 checkpoint. Complete optimization settings are reported in Table 14.

Circular Regression and Data Splits. The bobblehead sequence contains 8,097 RGB frames of size 240×320, each associated with two independent rotation angles. Frames are resized to 128×128 and encoded once using the frozen epoch-800 encoder. All subsequent analyses operate on these fixed representations. For each seed $s \in \{0, \dots, 9\}$, the frames are randomly permuted and split into 5,182 training, 1,296 validation, and 1,619 test examples. The same ten splits are shared across all encoders, and the pretrained representation is held fixed throughout, so seeds vary only the downstream probe partition. For a probe-training budget n, we use the first $\min(n, 5182)$ examples of the corresponding shuffled training pool. Budgets are therefore nested within each seed, while its validation and test sets remain unchanged. Features are standardized using statistics computed exclusively from the selected training subset. For each rotation coordinate θ, an independent ridge regressor predicts $(\cos\theta, \sin\theta)$ together with an unregularized intercept. The ridge coefficient is selected from $\{10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$, using angular MAE on the complete 1,296-frame validation split and is then evaluated once on the corresponding test set.

Given predictions $(\hat{c}, \hat{s})$, we recover $\hat{\theta} = \operatorname{atan2}(\hat{s}, \hat{c})$ and report the wrapped angular MAE:

$$\text{MAE} = \frac{180}{\pi N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \left| \text{Arg}\left(e^{\mathrm{i}(\hat{\theta}_i - \theta_i)}\right) \right| \tag{48}$$

The reported R^2 pools residual sums of squares over the cosine and sine targets. Results are means and sample standard deviations over the ten shared probe splits. Because each configuration uses a fixed pretrained encoder, these deviations quantify probe-split variability rather than variability across independent pretraining runs. The random split holds out individual frames rather than complete

Table 14: *Imagenette Pretraining Settings.* The upper block lists settings shared by all encoders, while the remaining blocks give the optimization parameters and loss weights for each objective family.

Shared across all backbones	
Dataset	Imagenette, 160 px variant (10 classes)
Training horizon	800 epochs
Batch / accumulation	64 images per loss / 4 steps (256 per update)
Encoder	Small ViT variants, patch 8; 384 internal / 512 output dims
Projector	MLP 2048-2048-16 with BatchNorm
Views per image	4
Crop	`RandomResizedCrop` 128×128, scale $[0.08, 1.0]$
Augmentations	h-flip; jitter p=0.8; grayscale p=0.2; blur p=0.5; solarize p=0.2; ImageNet norm
Optimizer	AdamW, weight decay 5×10^{-2}, cosine LR
Online probe	LayerNorm + Linear, LR 10^{-3}, wd 10^{-7}
Global (SIGReg / SUSReg / SubReg)	
LR / warmup / grad-clip	2×10^{-3} / 1 ep / -
Loss balance	$\lambda = 0.02$
Local (LocaLeReg / SIGReg) - stabilized recipe	
LR / warmup / grad-clip	10^{-3} / 10 ep / 1.0
Loss balance	$\lambda \in \{0.1, 0.3\}$ in the reported angle table
Rectified LpJEPA (RDMReg)	
LR / warmup / grad-clip	2×10^{-3} / 1 ep / -
Loss balance	$25\,\mathcal{L}_{\text{inv}} + 125\,\text{RDMReg}$

objects, trajectories, or angular sectors. The experiment therefore measures recovery and interpolation of the two periodic factors on the observed sequence, rather than extrapolation to unseen objects or disjoint regions of $\mathbb{T}^2$.

Table 15: *Rotation recovery after Imagenette pretraining, All Configurations.* Same protocol as Table 4, with every neighborhood count K, weight λ, and the local Gaussian control (LocaLeJEPA-SIGReg: cluster-wise SIGReg on uncentered embeddings).

		Figurine 0		Figurine 1	
Encoder	Regularizer	$R^2 \uparrow$	MAE (°)↓	$R^2 \uparrow$	MAE (°)↓
Global distribution matching					
Spherical ViT	SUSReg	0.9896 ± 0.0003	3.25 ± 0.05	0.9738 ± 0.0010	4.04 ± 0.07
nViT	SIGReg	0.9897 ± 0.0003	3.21 ± 0.09	0.9678 ± 0.0015	4.37 ± 0.08
nViT	SubReg	0.9891 ± 0.0002	3.26 ± 0.03	0.9643 ± 0.0012	4.63 ± 0.11
ViT	SIGReg	0.9930 ± 0.0002	2.75 ± 0.06	0.9775 ± 0.0008	3.85 ± 0.08
ViT	SubReg	0.9926 ± 0.0002	2.84 ± 0.06	0.9758 ± 0.0010	3.97 ± 0.07
ViT	RDMReg	0.9899 ± 0.0001	3.33 ± 0.03	0.9714 ± 0.0012	4.25 ± 0.09
Local distribution matching					
ViT	LocaLeJEPA (λ=0.1, K=2)	$\mathbf{0.9959 \pm 0.0001}$	$\mathbf{2.13 \pm 0.04}$	0.9840 ± 0.0007	3.21 ± 0.09
ViT	LocaLeJEPA (λ=0.1, K=1→8)	0.9938 ± 0.0002	2.58 ± 0.05	$\mathbf{0.9841 \pm 0.0003}$	$\mathbf{3.20 \pm 0.08}$
ViT	LocaLeJEPA (λ=0.1, K=4)	0.9946 ± 0.0001	2.44 ± 0.04	0.9821 ± 0.0003	3.42 ± 0.07
ViT	LocaLeJEPA (λ=0.3, K=2)	0.9919 ± 0.0002	2.89 ± 0.04	0.9808 ± 0.0007	3.48 ± 0.07
ViT	LocaLeJEPA-SIGReg (λ=0.1, K=4)	0.9919 ± 0.0002	2.94 ± 0.07	0.9770 ± 0.0009	3.84 ± 0.08

C.3.1 Complete Torus Results at Full Probe Budget

In this subsection, we provide the complete full-budget comparison of Table 2 in Section 5, together with ablations over the regularization weight $\lambda \in \{0.1, 0.3\}$ and the neighborhood count $K \in \{2, 4, 1\to8\}$.

Results. As can be observed in Table 15, across the three LocaLeJEPA variants with λ=0.1, local matching improves both rotation coordinates over every global baseline. Fixed K=2 gives the strongest overall recovery, reaching 2.13° and 3.21° MAE on θ_0 and θ_1, respectively. Increasing the partition granularity to K=4 degrades both readouts to 2.44° and 3.42°, indicating that finer

neighborhoods are not uniformly beneficial. The dynamic schedule $K{=}1{\to}8$ lies between the two fixed settings on θ_0 ($2.58°$), while attaining the best θ_1 result ($R^2{=}0.9841$, MAE $3.20°$). Thus, progressively increasing locality can match the best fixed setting on one factor, but does not improve the aggregate recovery over $K{=}2$. Performance is also sensitive to the strength of the local regularizer. At fixed $K{=}2$, increasing λ from 0.1 to 0.3 raises MAE from $2.13°$ to $2.89°$ on θ_0 and from $3.21°$ to $3.48°$ on θ_1. The lower-weight configuration is therefore clearly preferred for this readout: with $\lambda{=}0.3$, performance on θ_0 falls slightly below global ViT-SIGReg ($2.89°$ versus $2.75°$), although the advantage on θ_1 remains ($3.48°$ versus $3.85°$).

In this ablation, we also consider *LocaLeJEPA-SIGReg* as introduced in Section C.2, which tests whether local neighborhoods alone are sufficient. Keeping $\lambda{=}0.1$ and $K{=}4$ but replacing the centered ball construction with cluster-wise SIGReg on unentered embeddings yields $2.94°$ and $3.84°$ MAE, effectively eliminating the improvement over global ViT-SIGReg. This shows that the gains cannot be explained by locality alone and identifies the centering operation and the reference distribution as both important components of LocaLeJEPA. However, the experiment does not disentangle their individual contributions.

C.3.2 Probe Label-Budget Curves

In this subsection we extend Table 2 and ablate over multiple probe-training budgets, together with a subset of parameters $(\lambda, K) \in \{0.1, 0.3\} \times \{2, 4, 1{\to}8\}$. We analyze up to ten budgets, from 5 labeled frames to the full 5,182-frame training pool, and report MAE and R^2 separately for both rotation coordinates. All values are means $\pm$ sample standard deviations over the same ten probe splits used in the main paper, with the pretrained encoder fixed throughout.

Results. Across the two coordinates, LocaLeJEPA with $\lambda{=}0.1$ and fixed $K{=}2$ achieves the lowest mean angular error at every budget from 10 frames onward. The improvement becomes increasingly clear as the budget increases. Relative to global ViT-SIGReg, the two-coordinate mean MAE decreases from $39.2°$ to $36.0°$ with 10 labels, from $17.8°$ to $15.5°$ with 50, and from $11.0°$ to $9.0°$ with 100. At 500 frames and at full supervision, the corresponding errors decrease from $4.70°$ to $3.80°$ and from $3.30°$ to $2.67°$, respectively. The coordinate-wise curves reveal a more nuanced low-data regime. For θ_0, fixed-K LocaLeJEPA attains the lowest MAE from 10 frames onward and remains best throughout the remainder of the curve (Table 17a). For θ_1, the leading method varies at small budgets: RDMReg is best with 5 frames, ViT-SubReg with 10 and 20, dynamic-K LocaLeJEPA with 50, and the $\lambda{=}0.3$, $K{=}2$ variant with 100 (Table 17b). Once the probe has at least 500 training frames, dynamic-K gives the lowest θ_1 error, reaching $3.20°$ at the full budget compared with $3.21°$ for fixed $K{=}2$. Thus, the consistent advantage of fixed $K{=}2$ in the two-coordinate mean does not imply uniform dominance on each rotation factor individually. The R^2 curves mirror the MAE results. Fixed-K LocaLeJEPA gives the strongest θ_0 recovery from 10 frames onward, reaching $R^2 = 0.996$ at full supervision, while the preferred configuration for θ_1 depends more strongly on the available label budget. At five training frames, performance is close to the floor of the task: no configuration exceeds $R^2 = 0.08$ on θ_0, and all methods are near zero or negative on θ_1. All comparisons in Table 17 concern variability of the downstream probe only. A summary of the ablation over λ and K can be found also inside Table 16.

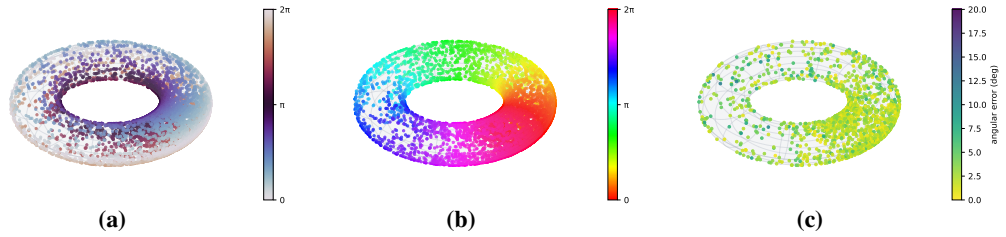
C.3.3 Spatial Structure of Rotation Recovery

We also investigate the visualization of the spatial structure of the rotation recovery experiment to expose where the errors occur over the underlying angular domain, a factor that the aggregate probe metrics do not reveal. For that purpose, we use (θ_0, θ_1) as coordinates on $\mathbb{T}^2$ and embed the resulting torus in $\mathbb{R}^3$ for visualization. These plots should be interpreted as visualizations of *coordinate recovery*. Their geometry is determined by the ground-truth or probe-decoded angles, rather than by a direct embedding of the encoder features themselves. They therefore show how faithfully the periodic factors are recovered across the angular domain, but do not by themselves establish the topology of the learned representation space.

Results. Figure 7 provides the basic construction. The first two panels show the ground-truth torus colored by the two periodic coordinates, while the final panel places held-out frames according to the angles decoded from fixed-K LocaLeJEPA ($\lambda{=}0.1$, $K{=}2$) and colors them by their angular error. The decoded coordinates cover the full angular domain, with most residual error remaining small

Table 16: *Mean Rotation-Recovery Error across Probe-label Budgets.* Wrapped angular MAE (°) averaged over the two rotation coordinates as the number of labeled frames used to fit the ridge probe increases. Results are mean ± sample standard deviation over ten shared probe splits, with the two coordinates averaged within each split. Per-coordinate results are reported in Table 17. Bold marks the lowest mean at each budget.

Method	10	50	100	500	Full
Spherical ViT SUSReg	44.6 ± 4.3	20.2 ± 2.0	11.0 ± 0.7	4.91 ± 0.09	3.64 ± 0.04
nViT SIGReg	41.8 ± 4.9	23.1 ± 2.0	11.5 ± 0.5	5.23 ± 0.10	3.79 ± 0.06
nViT SubReg	46.0 ± 4.5	26.7 ± 1.7	14.6 ± 0.7	5.65 ± 0.12	3.94 ± 0.05
ViT SIGReg	39.2 ± 3.6	17.8 ± 1.1	11.0 ± 0.4	4.70 ± 0.09	3.30 ± 0.05
ViT SubReg	37.9 ± 4.4	18.0 ± 1.7	11.2 ± 0.5	4.70 ± 0.10	3.40 ± 0.04
ViT RDMReg	39.8 ± 7.0	19.8 ± 2.0	10.5 ± 0.6	5.47 ± 0.08	3.79 ± 0.05
ViT LocaLeJEPA (λ=0.1, K=2)	$\mathbf{36.0 \pm 3.8}$	$\mathbf{15.5 \pm 1.2}$	$\mathbf{9.0 \pm 0.3}$	$\mathbf{3.80 \pm 0.05}$	$\mathbf{2.67 \pm 0.05}$
ViT LocaLeJEPA (λ=0.1, K=1→8)	37.5 ± 4.3	17.1 ± 1.1	10.9 ± 0.5	4.10 ± 0.10	2.89 ± 0.04
ViT LocaLeJEPA (λ=0.1, K=4)	37.6 ± 4.7	19.2 ± 1.8	9.6 ± 0.4	4.07 ± 0.08	2.93 ± 0.03
ViT LocaLeJEPA (λ=0.3, K=2)	40.3 ± 3.3	22.1 ± 1.6	9.7 ± 0.3	4.66 ± 0.07	3.19 ± 0.04
ViT LocaLeJEPA-SIGReg (λ=0.1, K=4)	41.5 ± 5.0	20.5 ± 1.3	9.8 ± 0.8	4.75 ± 0.06	3.39 ± 0.05



(a) **(b)** **(c)**

Figure 7: *Ground-truth and Recovered Angular Coordinates.* (a,b) The 8,097 frames positioned on $\mathbb{T}^2$ using the ground-truth rotation coordinates and colored by θ_0 and θ_1, respectively. (c) Held-out frames positioned using angles decoded from frozen LocaLeJEPA features (λ=0.1, K=2), with color indicating per-frame angular error (mean 2.7°).

relative to the range of the two periodic variables. Figure 8 extends this view across all evaluated encoders and probe-label budgets. Rows correspond to representations and columns to 10, 50, 100, 500, and all 5,182 probe-training frames. Each panel reports the MAE for the same held-out split and colors individual predictions using a shared angular error scale. The progression across columns makes the effect of probe supervision visible directly: sparse-label probes produce more dispersed and higher-error reconstructions, whereas increasing the training budget progressively concentrates the predictions around the underlying toroidal coordinate structure. To compare methods spatially, Figure 9 uses fixed-K LocaLeJEPA as the reference and visualizes the per-frame difference:

$$\Delta_{\text{err}} = \text{err}(\text{competitor}) - \text{err}(\text{LocaLeJEPA}) \tag{49}$$

Positive values therefore indicate frames on which fixed-K LocaLeJEPA is more accurate. Each panel additionally reports the mean error difference and the fraction of held-out frames won by LocaLeJEPA. At the 10-frame budget the comparison is mixed, whereas from 50 frames onward the mean difference is positive against every displayed competitor on this split. At the full budget, the mean advantage remains positive throughout the grid and LocaLeJEPA is more accurate on 53–69% of held-out frames, depending on the competitor. All three visualizations use one fixed probe split, whereas Section C.3.2 reports averages over ten shared splits.

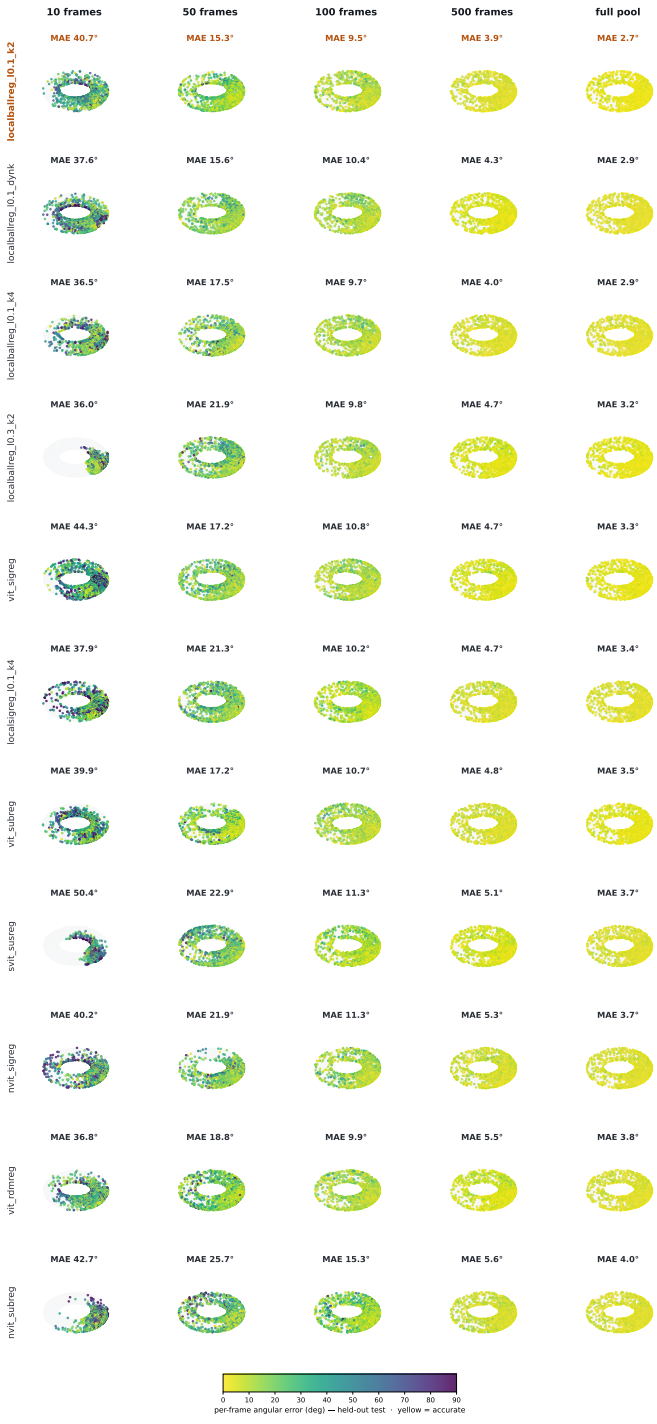


Figure 8: *Spatial Rotation Recovery across Encoders and Probe-label Budgets.* Held-out frames are positioned according to their probe-decoded rotation coordinates and colored by angular error using a common scale across all panels. Rows compare encoders and columns vary the number of labeled frames used to fit the probe. Visualizations correspond to one shared probe split. The aggregate results over ten splits are reported in Section C.3.2.

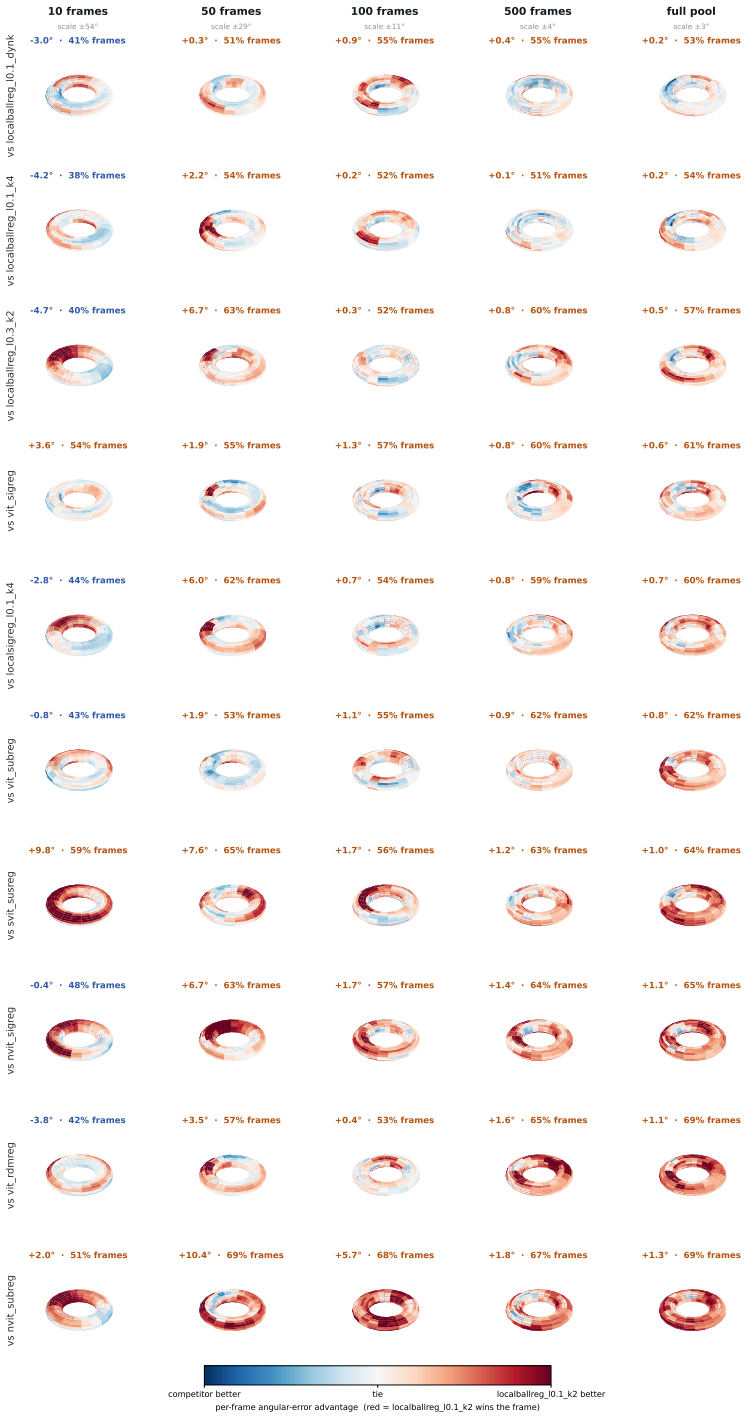


Figure 9: *Spatial Distribution of the LocaLeJEPA Recovery Advantage.* Regional mean of $\mathrm{err}(\mathrm{competitor}) - \mathrm{err}(\mathrm{LocaLeJEPA})$ relative to fixed-K LocaLeJEPA (λ=0.1, K=2). Positive values (red) favor LocaLeJEPA, while negative values (blue) favor the competitor. Panel titles report the mean error difference and the fraction of held-out frames on which LocaLeJEPA obtains lower error. *Full* denotes the complete 5,182-frame probe-training pool. All panels use the same fixed probe split.

Table 17: *Rotation Recovery across Probe-label Budgets.* Circular ridge probes are fitted on frozen representations using different labeled frame budgets $\in \{5, 10, 20, 50, 100, 500, 800, 2000, 4000, 5182\}$, with *Full=5182* denoting the complete training pool. Panels report angular MAE and R^2 separately for the two rotation coordinates. Ridge strength is selected on the same 1,296-frame validation set at every budget. Results are mean $\pm$ sample standard deviation over ten shared probe splits; bold marks the best mean in each column. Complete encoder configurations are given in Table 15.

(a) Figurine 0. Test MAE ($^\circ$, $\downarrow$)

Method	5	10	20	50	100	500	800	2000	4000	Full
SUSReg	73.69 ± 10.20	52.70 ± 7.03	42.24 ± 4.85	18.90 ± 2.34	10.68 ± 0.80	4.53 ± 0.15	4.07 ± 0.10	3.57 ± 0.06	3.32 ± 0.05	3.25 ± 0.05
SIGReg	62.17 ± 13.10	42.51 ± 7.38	31.74 ± 2.74	18.84 ± 3.00	9.43 ± 0.60	4.37 ± 0.15	3.94 ± 0.11	3.49 ± 0.07	3.28 ± 0.09	3.21 ± 0.09
SubReg	66.39 ± 8.12	50.91 ± 7.13	39.64 ± 4.80	25.49 ± 2.25	12.94 ± 0.87	4.77 ± 0.07	4.18 ± 0.06	3.58 ± 0.06	3.33 ± 0.03	3.26 ± 0.03
SIGReg	60.46 ± 9.80	40.02 ± 7.95	25.79 ± 6.50	15.24 ± 0.95	9.50 ± 0.61	4.01 ± 0.13	3.59 ± 0.11	3.06 ± 0.08	2.83 ± 0.07	2.75 ± 0.06
SubReg	56.99 ± 8.76	41.60 ± 7.97	29.86 ± 3.99	18.30 ± 2.44	10.17 ± 0.85	3.97 ± 0.12	3.57 ± 0.06	3.14 ± 0.07	2.92 ± 0.06	2.84 ± 0.06
RDMReg	69.09 ± 13.24	44.97 ± 12.60	29.27 ± 4.07	16.73 ± 2.34	9.14 ± 0.54	4.94 ± 0.11	4.42 ± 0.10	3.77 ± 0.06	3.43 ± 0.04	3.33 ± 0.03
LocaLeReg (λ=0.1, K=2)	60.98 ± 10.18	$\mathbf{34.90 \pm 6.01}$	$\mathbf{20.18 \pm 2.96}$	$\mathbf{12.37 \pm 1.30}$	$\mathbf{7.31 \pm 0.53}$	$\mathbf{3.08 \pm 0.08}$	$\mathbf{2.71 \pm 0.06}$	$\mathbf{2.33 \pm 0.04}$	$\mathbf{2.17 \pm 0.04}$	$\mathbf{2.13 \pm 0.04}$
LocaLeReg (λ=0.1, K=1→8)	$\mathbf{56.74 \pm 8.90}$	39.87 ± 8.14	27.53 ± 6.69	16.91 ± 1.07	10.08 ± 0.64	3.74 ± 0.07	3.29 ± 0.08	2.83 ± 0.06	2.63 ± 0.06	2.58 ± 0.05
LocaLeReg (λ=0.1, K=4)	59.09 ± 8.22	40.83 ± 7.16	27.27 ± 4.44	17.94 ± 3.03	7.98 ± 0.37	3.48 ± 0.07	3.09 ± 0.04	2.66 ± 0.04	2.50 ± 0.04	2.44 ± 0.04
LocaLeReg (λ=0.3, K=2)	60.37 ± 8.25	45.13 ± 5.35	29.27 ± 4.11	21.55 ± 1.78	9.12 ± 0.54	4.50 ± 0.12	3.99 ± 0.07	3.30 ± 0.05	2.98 ± 0.04	2.89 ± 0.04
LocaLeReg-SIGReg (λ=0.1, K=4)	63.91 ± 11.89	43.39 ± 6.62	31.05 ± 4.63	18.92 ± 1.83	8.97 ± 1.13	4.24 ± 0.10	3.83 ± 0.08	3.30 ± 0.07	3.02 ± 0.07	2.94 ± 0.07

(b) Figurine 1. Test MAE ($^\circ$, $\downarrow$)

Method	5	10	20	50	100	500	800	2000	4000	Full
SUSReg	44.72 ± 16.31	36.42 ± 5.34	30.06 ± 3.67	21.50 ± 2.25	11.30 ± 0.90	5.29 ± 0.13	4.90 ± 0.11	4.35 ± 0.08	4.12 ± 0.06	4.04 ± 0.07
SIGReg	49.03 ± 12.13	41.14 ± 5.63	37.05 ± 4.63	27.41 ± 2.95	13.48 ± 0.84	6.08 ± 0.11	5.40 ± 0.10	4.71 ± 0.09	4.44 ± 0.09	4.37 ± 0.08
SubReg	48.08 ± 12.75	41.04 ± 5.51	38.27 ± 6.95	27.95 ± 2.71	16.34 ± 1.14	6.53 ± 0.22	5.78 ± 0.17	5.03 ± 0.12	4.72 ± 0.10	4.63 ± 0.11
SIGReg	42.66 ± 10.10	38.45 ± 3.45	30.76 ± 3.04	20.40 ± 2.30	12.47 ± 0.64	5.39 ± 0.17	4.86 ± 0.16	4.25 ± 0.09	3.95 ± 0.08	3.85 ± 0.08
SubReg	43.27 ± 8.44	$\mathbf{34.18 \pm 3.03}$	$\mathbf{25.52 \pm 4.82}$	17.79 ± 2.18	12.18 ± 0.71	5.42 ± 0.15	4.91 ± 0.10	4.33 ± 0.09	4.05 ± 0.07	3.97 ± 0.07
RDMReg	$\mathbf{37.11 \pm 5.14}$	34.71 ± 5.75	30.30 ± 2.91	22.87 ± 2.30	11.84 ± 1.26	6.01 ± 0.12	5.48 ± 0.08	4.76 ± 0.09	4.38 ± 0.09	4.25 ± 0.09
LocaLeReg (λ=0.1, K=2)	43.38 ± 10.17	37.11 ± 4.53	27.84 ± 5.95	18.71 ± 2.38	10.67 ± 0.65	4.52 ± 0.08	4.05 ± 0.09	3.49 ± 0.09	3.26 ± 0.09	3.21 ± 0.09
LocaLeReg (λ=0.1, K=1→8)	43.61 ± 10.21	35.14 ± 4.66	26.95 ± 2.95	$\mathbf{17.21 \pm 1.78}$	11.81 ± 1.02	$\mathbf{4.47 \pm 0.19}$	$\mathbf{4.01 \pm 0.10}$	$\mathbf{3.48 \pm 0.08}$	$\mathbf{3.26 \pm 0.08}$	$\mathbf{3.20 \pm 0.08}$
LocaLeReg (λ=0.1, K=4)	41.99 ± 9.89	34.37 ± 3.81	27.69 ± 4.20	20.50 ± 2.38	11.15 ± 0.81	4.65 ± 0.15	4.20 ± 0.10	3.70 ± 0.09	3.48 ± 0.08	3.42 ± 0.07
LocaLeReg (λ=0.3, K=2)	39.12 ± 4.90	35.41 ± 3.49	31.10 ± 2.79	22.64 ± 1.94	$\mathbf{10.36 \pm 0.51}$	4.83 ± 0.12	4.36 ± 0.11	3.82 ± 0.09	3.56 ± 0.07	3.48 ± 0.07
LocaLeReg-SIGReg (λ=0.1, K=4)	42.44 ± 7.86	39.61 ± 5.12	32.31 ± 3.89	21.98 ± 1.77	10.55 ± 0.81	5.25 ± 0.07	4.78 ± 0.10	4.22 ± 0.10	3.92 ± 0.09	3.84 ± 0.08

(c) Figurine 0. Test R^2 ($^\circ$, $\uparrow$)

Method	5	10	20	50	100	500	800	2000	4000	Full
SUSReg	-0.270 ± 0.213	0.182 ± 0.207	0.324 ± 0.258	0.743 ± 0.045	0.897 ± 0.010	0.980 ± 0.001	0.984 ± 0.001	0.987 ± 0.000	0.989 ± 0.000	0.990 ± 0.000
SIGReg	-0.058 ± 0.294	0.352 ± 0.188	0.545 ± 0.043	0.759 ± 0.042	0.920 ± 0.008	0.981 ± 0.001	0.985 ± 0.001	0.988 ± 0.000	0.989 ± 0.000	0.990 ± 0.000
SubReg	-0.227 ± 0.315	0.169 ± 0.162	0.378 ± 0.141	0.612 ± 0.043	0.857 ± 0.014	0.977 ± 0.001	0.982 ± 0.000	0.987 ± 0.000	0.989 ± 0.000	0.989 ± 0.000
SIGReg	-0.034 ± 0.234	0.362 ± 0.198	0.642 ± 0.122	0.825 ± 0.021	0.915 ± 0.009	0.985 ± 0.001	0.988 ± 0.001	0.991 ± 0.000	0.993 ± 0.000	0.993 ± 0.000
SubReg	$\mathbf{0.077 \pm 0.258}$	0.315 ± 0.285	0.583 ± 0.066	0.772 ± 0.043	0.907 ± 0.011	0.985 ± 0.000	0.988 ± 0.000	0.991 ± 0.000	0.992 ± 0.000	0.993 ± 0.000
RDMReg	-0.301 ± 0.419	0.219 ± 0.285	0.588 ± 0.085	0.792 ± 0.040	0.920 ± 0.009	0.978 ± 0.001	0.982 ± 0.001	0.987 ± 0.000	0.989 ± 0.000	0.990 ± 0.000
LocaLeReg (λ=0.1, K=2)	-0.040 ± 0.256	$\mathbf{0.486 \pm 0.140}$	$\mathbf{0.741 \pm 0.045}$	$\mathbf{0.869 \pm 0.022}$	$\mathbf{0.947 \pm 0.007}$	$\mathbf{0.991 \pm 0.000}$	$\mathbf{0.993 \pm 0.000}$	$\mathbf{0.995 \pm 0.000}$	$\mathbf{0.996 \pm 0.000}$	$\mathbf{0.996 \pm 0.000}$
LocaLeReg (λ=0.1, K=1→8)	0.029 ± 0.231	0.366 ± 0.189	0.595 ± 0.109	0.785 ± 0.027	0.900 ± 0.012	0.986 ± 0.000	0.989 ± 0.000	0.992 ± 0.000	0.994 ± 0.000	0.994 ± 0.000
LocaLeReg (λ=0.1, K=4)	0.025 ± 0.223	0.390 ± 0.146	0.639 ± 0.062	0.753 ± 0.087	0.938 ± 0.006	0.989 ± 0.001	0.991 ± 0.000	0.994 ± 0.000	0.994 ± 0.000	0.995 ± 0.000
LocaLeReg (λ=0.3, K=2)	-0.032 ± 0.211	0.318 ± 0.100	0.586 ± 0.069	0.687 ± 0.031	0.916 ± 0.009	0.980 ± 0.001	0.985 ± 0.000	0.989 ± 0.000	0.991 ± 0.000	0.992 ± 0.000
LocaLeReg-SIGReg (λ=0.1, K=4)	-0.049 ± 0.233	0.303 ± 0.165	0.559 ± 0.126	0.751 ± 0.028	0.922 ± 0.017	0.983 ± 0.001	0.986 ± 0.000	0.990 ± 0.000	0.991 ± 0.000	0.992 ± 0.000

(d) Figurine 1. Test R^2 ($^\circ$, $\uparrow$)

Method	5	10	20	50	100	500	800	2000	4000	Full
SUSReg	-0.173 ± 0.397	0.045 ± 0.234	0.241 ± 0.110	0.505 ± 0.057	0.810 ± 0.025	0.952 ± 0.002	0.959 ± 0.001	0.969 ± 0.001	0.973 ± 0.001	0.974 ± 0.001
SIGReg	-0.238 ± 0.261	-0.093 ± 0.238	0.098 ± 0.168	0.357 ± 0.037	0.742 ± 0.020	0.938 ± 0.003	0.951 ± 0.002	0.962 ± 0.002	0.967 ± 0.002	0.968 ± 0.002
SubReg	-0.217 ± 0.300	0.011 ± 0.094	0.097 ± 0.153	0.323 ± 0.112	0.655 ± 0.032	0.929 ± 0.004	0.944 ± 0.002	0.958 ± 0.002	0.963 ± 0.001	0.964 ± 0.001
SIGReg	-0.083 ± 0.233	-0.044 ± 0.336	0.273 ± 0.120	0.560 ± 0.061	0.771 ± 0.016	0.955 ± 0.002	0.964 ± 0.001	0.973 ± 0.001	0.976 ± 0.001	0.978 ± 0.001
SubReg	-0.106 ± 0.210	$\mathbf{0.166 \pm 0.146}$	$\mathbf{0.424 \pm 0.132}$	0.638 ± 0.040	0.778 ± 0.025	0.954 ± 0.002	0.962 ± 0.001	0.971 ± 0.001	0.975 ± 0.001	0.976 ± 0.001
RDMReg	$\mathbf{0.005 \pm 0.150}$	0.148 ± 0.144	0.291 ± 0.089	0.456 ± 0.080	0.786 ± 0.040	0.940 ± 0.003	0.951 ± 0.002	0.964 ± 0.001	0.970 ± 0.001	0.971 ± 0.001
LocaLeReg (λ=0.1, K=2)	-0.093 ± 0.215	0.093 ± 0.130	0.380 ± 0.139	0.594 ± 0.056	0.821 ± 0.019	0.966 ± 0.001	0.973 ± 0.001	0.981 ± 0.001	0.983 ± 0.001	0.984 ± 0.001
LocaLeReg (λ=0.1, K=1→8)	-0.094 ± 0.247	0.122 ± 0.129	0.388 ± 0.096	$\mathbf{0.646 \pm 0.046}$	0.791 ± 0.029	$\mathbf{0.968 \pm 0.002}$	$\mathbf{0.974 \pm 0.001}$	$\mathbf{0.981 \pm 0.000}$	$\mathbf{0.983 \pm 0.000}$	$\mathbf{0.984 \pm 0.000}$
LocaLeReg (λ=0.1, K=4)	-0.040 ± 0.195	0.149 ± 0.143	0.369 ± 0.091	0.564 ± 0.052	0.804 ± 0.023	0.966 ± 0.002	0.972 ± 0.001	0.979 ± 0.000	0.981 ± 0.000	0.982 ± 0.000
LocaLeReg (λ=0.3, K=2)	-0.014 ± 0.124	0.162 ± 0.087	0.285 ± 0.065	0.471 ± 0.064	$\mathbf{0.830 \pm 0.016}$	0.961 ± 0.002	0.969 ± 0.001	0.977 ± 0.001	0.980 ± 0.001	0.981 ± 0.001
LocaLeReg-SIGReg (λ=0.1, K=4)	-0.070 ± 0.191	0.021 ± 0.184	0.248 ± 0.136	0.483 ± 0.062	0.821 ± 0.020	0.952 ± 0.003	0.962 ± 0.001	0.971 ± 0.001	0.976 ± 0.001	0.977 ± 0.001

C.4 IMAGENET-1K: TRAINING AND EVALUATION PROTOCOL

This section specifies the experimental choices particular to the ImageNet-1K study underlying Sections 5.1 and 5.2. We first describe the three encoder geometries and their optimization settings, and then detail the frozen-feature classification, downstream-transfer, and rotation-recovery protocols. Shared training, checkpointing, and representation conventions follow Section C.1, while the compared distribution-matching objectives are defined in Section C.2.

C.4.1 IMAGENET-1K: PRETRAINING DETAILS

Following the experimental setup of (Balestriero & LeCun, 2025), we pretrain all models for 100 epochs on the 1.28M-image ImageNet-1K (Deng et al., 2009) training split.

Encoder Families. We consider three encoder families spanning distinct representation geometries. The *Euclidean* encoder is a standard ViT-S/8 (21.7M parameters). The *Hyperbolic* encoder uses the same ViT-S/8 backbone and maps its output to a Poincaré ball of curvature -0.1, with feature clipping at radius 2.3 (Ermolov et al., 2022). Finally, the D_4*-equivariant* encoder is a 24.4M-parameter ResNet-50 implemented with escnn (Klee et al., 2023), retaining the regular representation at its output rather than pooling to an invariant feature. All three encoders feed the common Euclidean projector defined in Section C.1.

Details. All runs use AdamW with cosine learning-rate decay and weight decay 5×10^{-2}. Global SIGReg and RDMReg use one warmup epoch and a learning rate of 2×10^{-3} for the two ViT encoders and 10^{-3} for the D_4-ResNet-50. LocaLeJEPA and its local control use a learning rate of 10^{-3}, ten warmup epochs, and gradient clipping at 1.0. SIGReg uses λ=0.02, while the LocaLeJEPA values of λ and K are reported with the corresponding configurations. RDMReg uses the objective weighting defined in Section C.2. For the dynamic-neighborhood variant, we instantiate the schedule of Section C.1 with $K_{\max} = 10$, and $E_{\text{step}} = 10$, so that the partition evolves from K=1 to K=10 over the 100-epoch training schedule, reaching its maximum granularity at epoch 90. Following the shared checkpointing protocol described in Section C.1, every downstream evaluation uses the final epoch-100 checkpoint.

C.4.2 IMAGENET-1K FROZEN-FEATURE READOUTS

Following the in-domain evaluation protocol of SPHERE-JEPA (Nicollier et al., 2026a), the ImageNet-1K training split forms the feature bank and the validation split is used for evaluation. Images are resized and center-cropped to 128×128 and normalized using ImageNet statistics.

k-NN and Linear Probes. For what concerns *k-NN probe*, the frozen representations are ℓ_2-normalized and compared using cosine similarity. Class votes are weighted according to $\exp(\text{sim}/0.07)$. We evaluate $k \in \{10, 20, 100, 200\}$ and use k=20 in the main paper. Across the tested configurations, this value is either optimal or within 0.5 percentage points of the best setting, and the ordering of methods is unchanged across the evaluated neighborhood sizes. Instead, in the *linear probe*, frozen representations are standardized using statistics computed from the ImageNet-1K training split. We fit a single affine classifier for 50 epochs using SGD with momentum 0.9, zero weight decay, batch size 4096, and cosine learning-rate decay. The learning rate is selected from $\{0.01, 0.03, 0.1, 0.3, 1, 3, 10\}$, using a 2% holdout from the training split. After selection, the classifier is refitted on the complete training split and evaluated once on the ImageNet-1K validation set. Both readouts operate on the same frozen epoch-100 representations. ImageNet-1K classification results are point estimates from a single pretraining run per configuration rather than averages over independent pretraining seeds.

Downstream Transfer. We evaluate the same frozen ImageNet-pretrained encoders on seven downstream datasets spanning textures, fine-grained recognition, and generic object classification: DTD (Cimpoi et al., 2014), FGVC-Aircraft (Maji et al., 2013), CIFAR-10 and CIFAR-100 (Krizhevsky, 2009), Flowers-102 (Nilsback & Zisserman, 2008), Food-101 (Bossard et al., 2014), and Oxford-IIIT Pets (Parkhi et al., 2012). We use DTD partition 1 train+val (3,760 images, 47 classes), Aircraft trainval (6,667/100), CIFAR-10 (50,000/10), CIFAR-100 (50,000/100), Flowers-102 train+val (2,040/102), Food-101 (75,750/101), and Pets trainval (3,680/37). The evaluation uses the corresponding official test split. Images are resized and center-cropped to 128×128, with no augmentation applied during probe training. Frozen representations are standardized and a single linear classifier is optimized for

Table 18: *ImageNet-1K Frozen-feature Classification across All Configurations.* Top-1 k-NN and linear-probe accuracy (%) on the ImageNet-1K validation set using the final epoch-100 checkpoints. Linear results average five probe seeds. **Bold** denotes the best mean within each encoder family.

	k-NN top-1 (%)				Linear (%)	
Method	k=10	k=20	k=100	k=200	top-1	top-5
Plain ViT-S/8 (Euclidean, 21.7M)						
LeJEPA (SIGReg)	24.0 ± 0.19	24.3 ± 0.19	23.2 ± 0.19	22.1 ± 0.19	40.7 ± 0.23	66.9 ± 0.21
LocaLeJEPA (λ=0.1, K=4)	$\mathbf{33.9 \pm 0.21}$	$\mathbf{34.1 \pm 0.21}$	$\mathbf{32.6 \pm 0.21}$	$\mathbf{31.2 \pm 0.21}$	$\mathbf{49.6 \pm 0.23}$	$\mathbf{75.0 \pm 0.20}$
Hyperbolic ViT-S/8 (Poincaré ball, 21.7M)						
LeJEPA (SIGReg)	24.4 ± 0.19	25.1 ± 0.19	23.9 ± 0.19	22.8 ± 0.19	41.1 ± 0.22	67.1 ± 0.21
Rectified LpJEPA (RDMReg)	14.2 ± 0.16	15.1 ± 0.16	15.6 ± 0.16	15.1 ± 0.16	29.2 ± 0.21	53.4 ± 0.23
LocaLeJEPA (λ=0.1, K=2)	33.2 ± 0.21	33.7 ± 0.21	32.2 ± 0.21	30.9 ± 0.21	49.6 ± 0.22	75.4 ± 0.19
LocaLeJEPA (λ=0.1, K=4)	$\mathbf{34.4 \pm 0.21}$	$\mathbf{34.9 \pm 0.21}$	$\mathbf{33.2 \pm 0.21}$	$\mathbf{31.8 \pm 0.21}$	$\mathbf{50.5 \pm 0.23}$	$\mathbf{75.8 \pm 0.19}$
LocaLeJEPA (λ=0.1, K=1→10)	29.4 ± 0.20	29.9 ± 0.20	28.8 ± 0.20	27.4 ± 0.20	47.4 ± 0.22	73.1 ± 0.20
LocaLeJEPA (λ=0.3, K=2)	28.5 ± 0.20	29.4 ± 0.20	28.8 ± 0.20	27.8 ± 0.20	43.8 ± 0.22	69.9 ± 0.21
LocaLeJEPA-SIGReg (λ=0.1, K=4)	25.2 ± 0.19	26.0 ± 0.20	25.2 ± 0.19	24.2 ± 0.19	40.7 ± 0.22	66.5 ± 0.21
D_4-equivariant ResNet-50 (24.4M)						
LeJEPA (SIGReg)	22.4 ± 0.19	23.3 ± 0.19	22.8 ± 0.19	22.1 ± 0.19	35.7 ± 0.22	63.0 ± 0.22
Rectified LpJEPA (RDMReg)	18.8 ± 0.17	19.7 ± 0.18	19.5 ± 0.18	18.6 ± 0.17	27.9 ± 0.20	54.1 ± 0.22
LocaLeJEPA (λ=0.1, K=2)	31.6 ± 0.21	32.4 ± 0.21	31.1 ± 0.21	30.0 ± 0.20	41.8 ± 0.22	69.0 ± 0.21
LocaLeJEPA (λ=0.1, K=4)	$\mathbf{32.3 \pm 0.21}$	$\mathbf{33.0 \pm 0.21}$	$\mathbf{31.7 \pm 0.21}$	$\mathbf{30.4 \pm 0.21}$	$\mathbf{42.3 \pm 0.22}$	$\mathbf{69.8 \pm 0.21}$
LocaLeJEPA (λ=0.1, K=1→10)	32.0 ± 0.21	32.5 ± 0.21	31.4 ± 0.21	30.1 ± 0.21	42.2 ± 0.22	69.5 ± 0.21
LocaLeJEPA (λ=0.3, K=2)	29.2 ± 0.20	30.0 ± 0.20	29.2 ± 0.20	28.1 ± 0.20	38.2 ± 0.22	65.3 ± 0.21
LocaLeJEPA-SIGReg (λ=0.1, K=4)	26.3 ± 0.20	27.0 ± 0.20	26.2 ± 0.20	25.2 ± 0.19	35.9 ± 0.21	62.7 ± 0.22

100 epochs using SGD with momentum 0.9, zero weight decay, batch size 256, and cosine learning-rate decay. The learning rate is selected from $\{10^{-3}, 3 \times 10^{-3}, \ldots, 30\}$ using a 10% holdout from the training partition, after which the classifier is refitted on the complete training split. We report top-1 test accuracy; *Avg* denotes the unweighted mean across the seven datasets.

C.4.3 IMAGENET-1K: FULL RESULTS

Tables 18 and 19 expand the comparison to the complete set of ImageNet-1K configurations detailed in the Section 5.1, including a subset of parameters over regularization strength and neighborhood granularity.

Regularization Strength. Across all three encoder families, λ=0.1 consistently provides stronger ImageNet classification than λ=0.3, with differences of approximately four to six points in linear top-1 accuracy. The same setting is also preferred in aggregate transfer accuracy, although individual downstream datasets can exhibit different orderings. These results indicate that the strength of the local distribution-matching term is an important component of the optimization regime rather than a neutral scaling choice.

Neighborhood granularity. At λ=0.1, fixed K=2, fixed K=4, and the dynamic schedule remain within approximately 0.6 percentage points of one another on ImageNet-1K classification. Performance is therefore comparatively stable over the tested neighborhood counts. Differences become more visible under transfer: in particular, the dynamic schedule attains the highest mean transfer accuracy within the D_4-equivariant family.

Locality versus reference distribution. LocaLeJEPA-SIGReg preserves the same local partitioning as LocaLeJEPA while replacing the centered uniform-ball target with Gaussian matching on uncentered clusters. The control improves k-NN accuracy over global SIGReg but does not recover the same improvement in linear classification. On the D_4-equivariant encoder, its mean transfer accuracy remains 3.7 points below LocaLeJEPA at the same λ and K. Local partitioning alone is therefore insufficient to account for the complete empirical gain, although, as discussed in the Section C.2, this control does not separately identify the effects of centering and reference distribution.

RDMReg. RDMReg trails global SIGReg on the ImageNet classification readouts across the evaluated encoder families. This provides an additional distribution-matching baseline with a substantially different reference geometry and shows that replacing the Gaussian target alone is not sufficient to reproduce the LocaLeJEPA improvements.

Table 19: *Transfer Linear Probing across All Configurations.* Top-1 accuracy (%) on the official test splits of seven downstream datasets using frozen epoch-100 ImageNet-1K representations, averaging results over five linear-probe seeds. **Bold** marks the best mean within each encoder family and column.

Method	DTD	Aircraft	CIFAR-10	CIFAR-100	Flowers	Food	Pets	Avg
Plain ViT-S/8 (Euclidean, 21.7M)								
LeJEPA (SIGReg)	56.4 ± 1.46	26.1 ± 0.77	84.4 ± 0.38	61.2 ± 0.60	68.4 ± 0.60	48.5 ± 0.33	52.0 ± 0.85	56.7 ± 0.29
LocaLeJEPA (λ=0.1, K=4)	$\mathbf{63.0 \pm 1.22}$	$\mathbf{31.1 \pm 0.92}$	$\mathbf{86.0 \pm 0.35}$	$\mathbf{63.8 \pm 0.49}$	$\mathbf{77.4 \pm 0.92}$	$\mathbf{55.7 \pm 0.32}$	$\mathbf{64.2 \pm 1.62}$	$\mathbf{63.0 \pm 0.32}$
Hyperbolic ViT-S/8 (Poincaré ball, 21.7M)								
LeJEPA (SIGReg)	57.4 ± 2.02	27.3 ± 0.91	85.3 ± 0.38	60.3 ± 0.50	70.9 ± 2.60	48.6 ± 0.33	53.4 ± 0.85	57.6 ± 0.60
Rectified LpJEPA (RDMReg)	50.0 ± 1.22	17.1 ± 0.67	77.2 ± 0.42	51.6 ± 0.51	56.4 ± 0.81	40.3 ± 0.31	39.6 ± 0.84	47.5 ± 0.28
LocaLeJEPA (λ=0.1, K=2)	60.9 ± 2.88	$\mathbf{32.4 \pm 0.91}$	86.8 ± 0.34	$\mathbf{65.2 \pm 0.52}$	$\mathbf{77.9 \pm 0.59}$	55.9 ± 0.32	$\mathbf{66.1 \pm 0.96}$	63.6 ± 0.46
LocaLeJEPA (λ=0.1, K=4)	$\mathbf{62.1 \pm 1.18}$	32.1 ± 0.81	$\mathbf{87.5 \pm 0.35}$	65.0 ± 0.48	77.9 ± 3.26	$\mathbf{56.4 \pm 0.31}$	65.9 ± 0.94	$\mathbf{63.9 \pm 0.48}$
LocaLeJEPA (λ=0.1, K=1→10)	59.5 ± 1.42	28.6 ± 1.04	86.5 ± 0.34	64.0 ± 0.56	74.7 ± 0.57	54.0 ± 0.31	62.0 ± 1.01	61.3 ± 0.29
LocaLeJEPA (λ=0.3, K=2)	58.6 ± 1.35	27.9 ± 0.84	83.9 ± 0.37	61.4 ± 0.52	73.7 ± 0.60	50.4 ± 0.32	59.3 ± 0.88	59.3 ± 0.28
LocaLeJEPA-SIGReg (λ=0.1, K=4)	55.7 ± 1.26	27.0 ± 0.88	83.1 ± 0.39	59.0 ± 0.60	70.5 ± 0.61	49.0 ± 0.32	56.5 ± 0.84	57.2 ± 0.27
D_4-equivariant ResNet-50 (24.4M)								
LeJEPA (SIGReg)	51.3 ± 1.38	16.4 ± 0.72	83.8 ± 0.37	55.8 ± 0.50	52.9 ± 0.73	34.3 ± 0.30	42.7 ± 1.09	48.2 ± 0.32
Rectified LpJEPA (RDMReg)	45.2 ± 1.18	14.1 ± 0.79	79.2 ± 0.42	47.8 ± 0.50	45.8 ± 1.24	27.2 ± 0.31	38.9 ± 1.30	42.6 ± 0.44
LocaLeJEPA (λ=0.1, K=2)	54.2 ± 1.19	21.3 ± 0.72	$\mathbf{85.0 \pm 0.36}$	57.5 ± 0.50	62.4 ± 0.81	39.0 ± 0.31	54.4 ± 1.19	53.4 ± 0.27
LocaLeJEPA (λ=0.1, K=4)	55.1 ± 1.19	20.7 ± 1.44	82.3 ± 0.45	55.3 ± 0.50	$\mathbf{66.4 \pm 0.63}$	$\mathbf{41.8 \pm 0.31}$	$\mathbf{56.6 \pm 0.85}$	54.0 ± 0.31
LocaLeJEPA (λ=0.1, K=1→10)	$\mathbf{56.7 \pm 1.19}$	$\mathbf{21.7 \pm 1.49}$	84.8 ± 0.36	$\mathbf{58.5 \pm 0.50}$	63.5 ± 1.69	40.5 ± 0.31	54.6 ± 0.97	$\mathbf{54.3 \pm 0.40}$
LocaLeJEPA (λ=0.3, K=2)	52.6 ± 1.40	21.2 ± 1.57	82.4 ± 0.38	55.3 ± 0.50	59.3 ± 0.72	37.6 ± 0.31	53.1 ± 0.84	51.6 ± 0.37
LocaLeJEPA-SIGReg (λ=0.1, K=4)	50.0 ± 1.24	20.4 ± 0.88	82.8 ± 0.42	54.9 ± 0.50	59.0 ± 0.69	32.9 ± 0.30	51.8 ± 0.95	50.3 ± 0.29

D Qualitative Embedding Analysis

In this section, we present a series of qualitative analyses and comparisons between the embeddings attained by the different JEPA recipes. In Section D.1, we complement the results presented in Figure 3, showing the recovery of topology in latent space. In Section D.2, we provide extended details about Figure 1, which shows that embeddings are able to capture meaningful semantics across geometric non-Euclidean embeddings. Lastly, in Section D.3 we perform a class consistency analysis across neigborhoods coming from different regularizations.

D.1 Topology Preservation on COIL Dataset

Data and training. COIL-100 (Nene et al., 1996) contains 7200 images of 100 objects, each took on a turntable at 72 poses 5° apart. For a fixed object the frames sample a closed curve $\theta \mapsto x_\theta$, $\theta \in S^1$, in image space, and we ask whether an encoder f maps this curve to a closed curve in embedding space or collapses it. We train ViT-S/8 encoders with LocaLeReg and SIGReg for 100 epochs with appearance-only augmentations (colour jitter, grayscale, Gaussian blur) and analyse the 512-d encoder output.

Persistent homology. To measure whether the 72 embeddings $\{f(x_\theta)\}$ of an object lie on a loop we use persistent homology (Edelsbrunner et al., 2002; Zomorodian & Carlsson, 2005). For a scale $r \geq 0$, the Vietoris–Rips complex VR_r contains a simplex for every set of points whose pairwise distances are at most r. These complexes are nested as r grows, and each topological feature appears at some scale b (its birth) and disappears at a scale d (its death): in dimension 0 the features are connected components, in dimension 1 they are loops. The persistence diagram plots every feature as the point (b, d), and its lifetime $d - b$ is the range of scales over which it exists. For a sampled closed curve, the loop is born once consecutive samples connect, and dies once the hole fills in; a long lifetime therefore indicates a robust hole, while features close to the diagonal are likely due to noise; this is made precise by the stability of persistence diagrams (Chazal et al., 2009). We compute diagrams with Ripser (Bauer, 2021) on the full 512-d embeddings after dividing all distances by the median pairwise distance among the 72 points of the object, so that lifetimes are comparable across representations of different scale and dimension.

Results. In pixel space the median longest H_1 lifetime over the 100 objects is 0.387, so the loop is present in the input. But encoders shorten it: under either regularizer the median is at most 0.196 for every seed and checkpoint. Over 3 training seeds, the median lifetime at the final weights is higher under LocaLeReg for every seed (0.190 ± 0.005 against 0.175 ± 0.002, mean and standard deviation over seeds); at the checkpoint with the best accuracy of the online linear probe, it is higher for two of the three (0.188 ± 0.007 against 0.181 ± 0.012). Figure 10 shows examples for several objects:

for each method, the persistence diagram of the object's 72 embeddings and their projection on the first 2 principal components, colored by rotation angle. The same runs also measure how well each representation separates the 100 objects (Table 20): the probe accuracy on held-out poses (every 6-th angle of every object) and the nearest-neighbor purity, the fraction of frames whose nearest neighbor in embedding space belongs to the same object, computed on 24 frames per object. Both are higher under LocaLeReg than under SIGReg for each of the 3 seeds, at the final weights and at the best checkpoint.

Table 20: Object identity on COIL-100, mean $\pm$ standard deviation over 3 seeds. The best checkpoint is the one with the highest probe accuracy.

	Probe accuracy		NN purity	
	final	best	final	best
LocaLeReg	0.384 ± 0.047	0.436 ± 0.026	0.492 ± 0.004	0.500 ± 0.006
SIGReg	0.264 ± 0.016	0.360 ± 0.022	0.431 ± 0.023	0.469 ± 0.018

D.2 Patch-Level Representation Structure

The quantitative frozen-feature evaluations characterize global representation quality but do not reveal how spatial information is organized before pooling. We therefore complement them with a qualitative PCA visualization of patch-level features, following the general visualization strategy used in LeJEPA [Balestriero & LeCun (2025)] and DINOv2 [Oquab et al. (2023)].

Results. For each backbone–objective pair, we select five ImageNet-1K validation images from the golden-retriever class (class 207) and extract spatial features from the frozen encoder immediately before global aggregation. For the Euclidean and hyperbolic ViT-S/8 encoders, we use the 384-dimensional tokens after the final transformer normalization, excluding prefix tokens, pooling, and output heads. At the 128×128 input resolution, these features form a 16×16 spatial grid. For the hyperbolic encoder, features are taken immediately before the Poincaré-ball mapping. For the D_4-equivariant ResNet-50, we use the 512-channel equivariant feature map immediately before average pooling. For each model, spatial features from the five images are concatenated into an $N \times d$ matrix, where the N spatial locations are observations and the d feature channels are variables. PCA is fitted once to this matrix, and the three leading components are retained. We fix the sign of each component by requiring its largest-magnitude loading to be positive and independently rescale the resulting component values to $[0, 1]$. The three channels are then mapped to RGB and reshaped to the native spatial grid. For visualization only, the result is upsampled to 512×512 using nearest-neighbor interpolation and overlaid on the input image at 55% opacity. No foreground masking or thresholding is applied. Figure 11 compares LocaLeJEPA (λ=0.1, K=4) with the corresponding SIGReg baseline on four representative images. Both objectives exhibit visible object–background structure in their leading principal components. These qualitative plots allow the PCA to identify directions of feature variance rather than semantic labels, and each model uses an independently fitted PCA basis. Consequently, colors are comparable across images within a row, but not across independently fitted models, and the visualization should not be interpreted as a quantitative segmentation or localization result.

D.3 Class Consistency of Frozen k-NN Neighborhoods

The largest ImageNet-1K improvement of LocaLeJEPA appears under the k-NN readout (Table 18), which depends directly on the local geometry of the frozen representation. We therefore examine whether this improvement is reflected in the composition of the neighborhoods on which the classifier operates. Specifically, we ask whether the nearest ImageNet-1K training examples of a validation image are more likely to share its semantic class under LocaLeJEPA than under SIGReg. We use exactly the representation space employed by the main k-NN evaluation: the 128×128 evaluation view, frozen 512-dimensional encoder features, cosine similarity, and the complete 1.28M-image ImageNet-1K training split as the retrieval bank.

Qualitative Retrieval. Figure 12 visualizes the eight nearest training examples for three query images under the Euclidean ViT-S/8. As shown, for the polar-bear query, the number of same-class

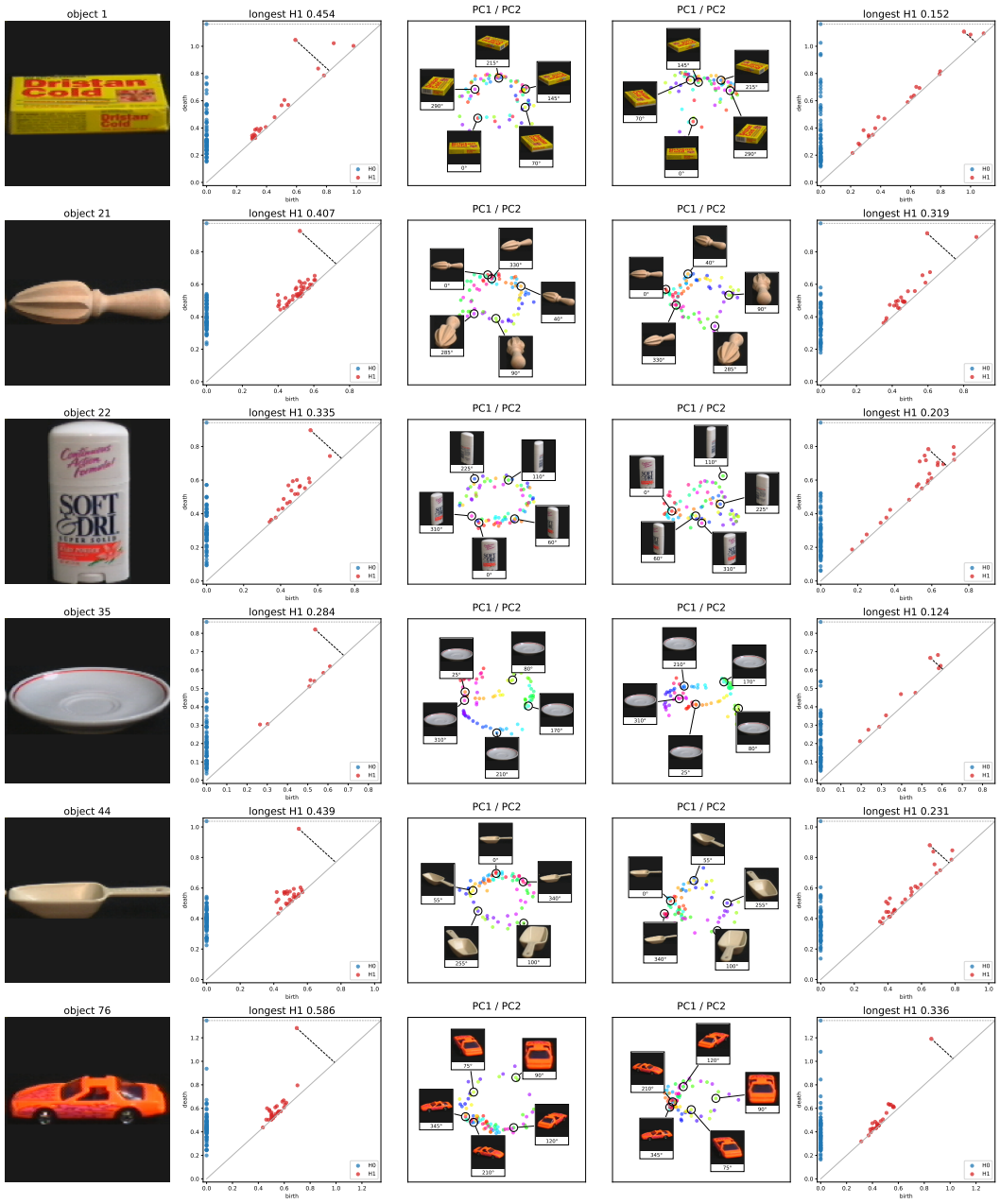


Figure 10: Per-object geometry of COIL-100 embeddings (ViT-S/8, 100 epochs, $\lambda = 0.1$; LocaLeReg with $k = 8$ clusters). Each row is one object: (a) its frame at 0°; (b) LocaLeReg and (c) SIGReg, each with the persistence diagram and the 2-d PCA projection of the embeddings of the object's 72 frames. *Persistence*: Vietoris–Rips diagram of the 72 embeddings, rescaled to unit median pairwise distance; the dashed segment marks the most persistent H_1 class.*PC1/PC2*: points are colored by rotation angle.

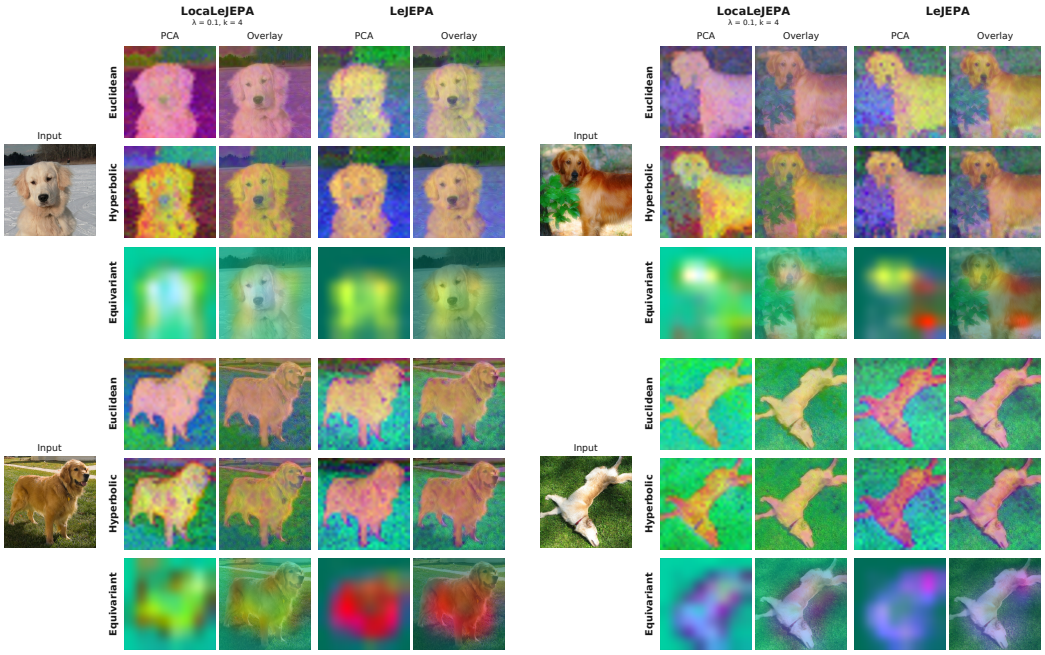


Figure 11: *Patch-level Principal Components for LocaLeJEPA and LeJEPA.* Frozen spatial features from the same ImageNet-1K validation images are projected onto their first three principal components and mapped to RGB. Each model uses an independently fitted PCA basis, so colors are directly comparable across images within a model but not across models.

Table 21: *Class Consistency of Frozen k-NN Neighborhoods by Semantic Group.* Mean top-20 same-class neighborhood consistency on the ImageNet-1K validation set, using the full training split as the retrieval bank. LocaLeJEPA uses λ=0.1 and K=4. Group sizes are shown in parentheses, and the final row reports the number of individual classes for which LocaLeJEPA yields higher consistency. **Bold** denotes the higher value within each encoder family.

	ViT-S/8		H-ViT-S/8		D_4-RN50	
	SIGReg	LocaLeJEPA	SIGReg	LocaLeJEPA	SIGReg	LocaLeJEPA
Fish (16)	0.25	**0.34**	0.26	**0.35**	0.28	**0.33**
Birds (38)	0.13	**0.22**	0.13	**0.22**	0.14	**0.19**
Amphibians (8)	0.11	**0.21**	0.11	**0.21**	0.11	**0.22**
Reptiles (36)	0.08	**0.15**	0.09	**0.16**	0.10	**0.17**
Invertebrates (86)	0.16	**0.26**	0.16	**0.26**	0.18	**0.25**
Marine mammals (4)	0.29	**0.40**	0.28	**0.41**	0.26	**0.31**
Dogs (118)	0.07	**0.12**	0.07	**0.12**	0.06	**0.12**
Wild canids (12)	0.09	**0.16**	0.09	**0.16**	0.09	**0.15**
Cats (13)	0.22	**0.31**	0.22	**0.31**	0.20	**0.30**
Bears (4)	0.16	**0.26**	0.17	**0.27**	0.15	**0.29**
Other mammals (63)	0.12	**0.20**	0.12	**0.21**	0.13	**0.21**
Objects, food, scenes (602)	0.14	**0.20**	0.14	**0.21**	0.14	**0.20**
All 1000 classes	0.13	**0.20**	0.13	**0.21**	0.13	**0.20**
LocaLeJEPA higher	989 / 1000		984 / 1000		971 / 1000	

neighbors increases from four under SIGReg to seven under LocaLeJEPA; for the Komodo dragon, from one to five; and for the triceratops, from three to four. Several SIGReg neighbors instead share strong visual attributes with the query (such as color or background) without sharing its class. The aggregate analysis below evaluates whether this pattern extends beyond these examples.



Figure 12: *Nearest-neighbor Retrieval in Frozen ViT-S/8 Features.* For each preselected query, we show the eight nearest ImageNet-1K training images under cosine similarity for SIGReg and LocaLeJEPA (λ=0.1, K=4). Green borders denote the query class and red borders a different class, with labels reporting class and cosine similarity.

Table 22: *Class Consistency of Frozen k-NN Neighborhoods by Class.* Mean top-20 same-class consistency over the 50 ImageNet-1K validation images of each class. LocaLeJEPA uses λ=0.1 and K=4. **Bold** denotes the higher value within each encoder family. The final row reports the average over all 1,000 classes.

	ViT-S/8		H-ViT-S/8		D_4-RN50	
	SIGReg	LocaLeJEPA	SIGReg	LocaLeJEPA	SIGReg	LocaLeJEPA
Bears, reptiles, marine mammals						
Polar bear	0.19	**0.31**	0.22	**0.32**	0.14	**0.26**
Brown bear	0.11	**0.20**	0.12	**0.23**	0.09	**0.22**
Komodo	0.14	**0.25**	0.21	**0.31**	0.29	**0.33**
Triceratops	0.20	**0.35**	0.20	**0.40**	0.33	**0.52**
Alligator	0.05	**0.08**	0.05	**0.07**	0.05	**0.10**
Cobra	0.05	**0.11**	0.06	**0.10**	0.10	**0.24**
Grey whale	0.31	**0.41**	0.31	**0.42**	**0.30**	0.27
Orca	0.40	**0.56**	0.39	**0.58**	0.37	**0.50**
Dugong	0.38	**0.49**	0.32	**0.49**	0.27	**0.38**
Sea lion	0.06	**0.12**	0.09	**0.13**	0.08	**0.11**
Other animals						
Goldfish	0.18	**0.35**	0.19	**0.34**	0.26	**0.29**
Great white shark	0.35	**0.40**	0.38	**0.43**	0.25	**0.32**
Bald eagle	0.16	**0.27**	0.14	**0.25**	0.14	**0.22**
Flamingo	0.28	**0.43**	0.31	**0.46**	0.29	**0.38**
King penguin	0.32	**0.57**	0.29	**0.53**	0.45	**0.68**
Golden retriever	0.04	**0.09**	0.04	**0.08**	0.04	**0.08**
Tabby cat	0.19	**0.30**	0.20	**0.31**	0.23	**0.33**
Tiger	0.32	**0.47**	0.32	**0.42**	0.24	**0.34**
Zebra	0.77	**0.85**	0.79	**0.85**	0.72	**0.78**
African elephant	0.25	**0.35**	0.23	**0.34**	0.30	**0.32**
Giant panda	0.46	**0.61**	0.54	**0.66**	0.49	**0.64**
Ladybug	0.05	**0.16**	0.06	**0.13**	0.07	**0.10**
Monarch butterfly	0.51	**0.72**	0.59	**0.77**	0.33	**0.52**
Objects, food, scenes						
Acoustic guitar	0.06	**0.12**	0.07	**0.13**	0.05	**0.12**
Airliner	0.53	**0.58**	0.52	**0.59**	0.50	**0.52**
Ambulance	0.34	**0.36**	0.31	**0.38**	0.22	**0.33**
Castle	0.25	**0.28**	0.22	**0.27**	0.16	**0.23**
Grand piano	0.09	**0.22**	0.08	**0.21**	0.07	**0.23**
Laptop	0.04	**0.10**	0.04	**0.10**	0.04	**0.06**
School bus	0.12	**0.20**	0.15	**0.22**	0.10	**0.15**
Sports car	0.17	**0.22**	0.14	**0.24**	0.19	**0.23**
Teapot	0.05	**0.07**	**0.07**	0.07	0.06	**0.13**
Banana	0.06	**0.12**	0.06	**0.15**	0.07	**0.11**
Pizza	0.22	**0.37**	0.23	**0.34**	0.18	**0.26**
Volcano	0.25	**0.32**	0.23	**0.28**	0.27	**0.33**
Daisy	0.72	**0.84**	0.78	**0.83**	**0.85**	0.83
All 1000 classes	0.13	**0.20**	0.13	**0.21**	0.13	**0.20**

Neighborhood Class Consistency. For a validation image x with class label $c(x)$, let $\mathcal{N}_K(x)$ denote its K nearest training examples under cosine similarity. We define the neighborhood class consistency as:

$$q_K(x) = \frac{1}{K} \sum_{y \in \mathcal{N}_K(x)} \mathbb{1}[c(y) = c(x)] \tag{50}$$

Thus, $q_K(x)$=1 when every retrieved neighbor belongs to the query class and $q_K(x)$=0 when none does. We use K=20, matching the headline k-NN readout, and average the statistic either over the 50 validation images of each class or over the complete 50k-image validation set.

Results. Table 21 reports the statistic over all 1,000 ImageNet-1K classes, grouped by WordNet category. Under SIGReg, approximately 13% of the top-20 neighborhood shares the query class for each of the three encoder families. LocaLeJEPA increases this fraction to approximately 20% for the Euclidean ViT-S/8 and D_4-equivariant ResNet-50, and to 21% for the hyperbolic ViT-S/8. The improvement is consistent in all classes. LocaLeJEPA increases class consistency in every reported WordNet group and in 989, 984, and 971 of the 1,000 individual classes for the Euclidean, hyperbolic, and equivariant encoders, respectively. Table 22 provides a representative 36-class breakdown, where LocaLeJEPA yields higher consistency in 105 of the 108 class–encoder comparisons. Absolute neighborhood consistency varies substantially across semantic groups. Fine-grained families containing many visually similar categories, such as dog breeds and reptiles, exhibit lower class purity

under both objectives, whereas visually distinctive categories such as zebra, daisy, and monarch butterfly show substantially higher values. This behavior is expected for ImageNet-1K: the weighted k-NN prediction requires only that the correct class receive the largest aggregate vote, not that a majority of retrieved samples share the query label. Consequently, the 23–35% top-1 accuracies in Table 18 are compatible with neighborhoods whose majority consists of other classes. Together, the retrieval examples and full-validation statistics show that the k-NN improvement of LocaLeJEPA is accompanied by a systematic increase in local semantic class consistency across architectures and categories. This analysis directly characterizes the neighborhoods used by the classifier, without requiring an additional learned readout.