Joint Embedding Variational Bayes

Anonymous authors
Paper under double-blind review

Abstract

We introduce Variational Joint Embedding (VJE), a framework that synthesizes joint embedding and variational inference to enable self-supervised learning of probabilistic representations in a reconstruction-free, non-contrastive setting. Compared to energy-based predictive objectives that optimize pointwise discrepancies, VJE maximizes a symmetric conditional evidence lower bound (ELBO) for a latent-variable model defined directly on encoder embeddings. We instantiate the conditional likelihood with a heavy-tailed Student–t model using a polar decomposition that explicitly decouples directional and radial factors to prevent norm-induced instabilities during training. VJE employs an amortized inference network to parameterize a diagonal Gaussian variational posterior whose feature-wise variances are shared with the likelihood scale to capture anisotropic uncertainty without auxiliary projection heads. Across ImageNet-1K, CIFAR-10/100, and STL-10, VJE achieves performance comparable to standard non-contrastive baselines under linear and k-NN evaluation. We further validate these probabilistic semantics through one-class CIFAR-10 anomaly detection, where likelihood-based scoring under the proposed model outperforms comparable self-supervised baselines.

1 Introduction

Joint embedding architectures have emerged as a powerful paradigm for self-supervised representation learning in computer vision. These architectures can be broadly categorized into contrastive and non-contrastive methods. Contrastive approaches, such as SimCLR (Chen et al., 2020a) and MoCo (He et al., 2020; Chen et al., 2020b), learn representations by maximizing similarity between pairs of semantically related (positive) samples, and minimizing similarity between unrelated (negative) samples. Since these methods are unsupervised, they must approximate 'negative' associations through heuristic choices of negatives (typically other instances in the minibatch, and in some cases external memory banks or queues) to ensure the presence of sufficiently diverse samples. This estimation procedure increases computational and memory demands, and can complicate training in domains where modelling patterns can be sensitive to erroneous negative associations.

Non-contrastive methods, including BYOL (Grill et al., 2020), SimSiam (Chen and He, 2021), VICReg (Bardes et al., 2022), and Barlow Twins (Zbontar et al., 2021), avoid the need for negative samples by learning from paired views of the same input. To prevent the model from collapsing to a trivial solution (e.g., mapping all inputs to the same constant representation), these approaches incorporate architectural asymmetries and auxiliary heads (e.g., prediction/projection heads), as well as redundancy-reduction objectives, largely as heuristic choices. By eliminating the dependence on negative samples, non-contrastive methods simplify training and can extend the applicability of self-supervised learning to settings where the notion of similarity and what constitutes a 'negative' sample are inherently ambiguous. A prominent viewpoint within this landscape is the Joint Embedding Predictive Architecture (JEPA) (LeCun, 2022), which formulates training in terms of minimizing a pointwise compatibility or energy between predicted and target embeddings in representation space, with recent instantiations including I-JEPA (Assran et al., 2023) and LeJEPA (Balestriero and LeCun, 2025).

Despite their empirical success, both contrastive and non-contrastive objectives are commonly trained to produce deterministic point embeddings, as each input is mapped to a single point in latent space and learning proceeds by optimizing pointwise similarity or discrepancy objectives between paired embeddings. As a

result, such representations do not naturally support distributional semantics under a normalized probabilistic formulation in representation space, and therefore do not readily yield density-based scoring or calibrated uncertainty signals. In applications such as medical diagnosis (Begoli et al., 2019; Ghesu et al., 2019; Gal et al., 2017), anomaly detection (Schlegl et al., 2019; Zimmerer et al., 2018; Wang and Huang, 2018; Liang et al., 2018), and reinforcement learning (Depeweg et al., 2018; Chua et al., 2018; Ha and Schmidhuber, 2018), the inability to represent uncertainties over latent factors can severely limit the reliability and interpretability of the resulting embeddings, or can limit the effectiveness of downstream tasks.

Variational methods, and in particular Variational Autoencoders (VAEs) (Kingma and Welling, 2014) and their extensions (Higgins et al., 2017; van den Oord et al., 2017; Maaløe et al., 2019), provide a principled probabilistic framework by modelling latent variables as distributions. This is accomplished through a reconstruction-based objective, which ensures that latent variables capture detailed generative factors associated with pixel-level input data. However, when the end goal is to obtain high-level semantic representations for downstream tasks, enforcing pixel-level fidelity can impose significant and often unnecessary computational overhead. This motivates latent-variable objectives defined directly on representations rather than on pixels, while retaining the robust probabilistic representations that make VAEs attractive in uncertainty-sensitive applications.

In this work, we introduce **Variational Joint Embedding (VJE)**, a framework that synthesizes **variational inference** and **joint embedding** to provide a latent-variable formulation of self-supervised learning without relying on input reconstruction (i.e., pixel-level) or contrastive objectives. In contrast to pointwise compatibility objectives that directly optimize discrepancies between embeddings, VJE adopts a normalized probabilistic formulation in representation space as its training primitive by defining an explicit latent-variable likelihood on target embeddings. An amortized inference network parameterizes a variational posterior over the corresponding latent variables. The model is trained by maximizing a *symmetric conditional* evidence lower bound (ELBO) on paired embeddings, where the target branch is detached to implement fixed-observation conditioning in the likelihood term, yielding probabilistic representations with feature-wise (i.e., anisotropic) uncertainty. We formalize this objective-level distinction between pointwise energy-based predictive losses and normalized likelihood-based training in Appendix B.

Empirically, VJE attains performance comparable to leading non-contrastive methods on representation learning benchmarks across ImageNet (Russakovsky et al., 2015), STL-10 (Coates et al., 2011), and CIFAR (Krizhevsky, 2009), while additionally exhibiting coherent probabilistic semantics as demonstrated in a downstream anomaly detection task on CIFAR-10. Taken together, these results position VJE as a principled alternative to pointwise energy-based objectives, with normalized probabilistic modelling as the underlying training primitive for non-contrastive representation learning.

2 Background

Non-contrastive self-supervised learning (SSL). The central principle of non-contrastive SSL is to relate representations derived from different views of the same input, without relying on negative samples. Under the JEPA viewpoint (LeCun, 2022; Assran et al., 2023), this is instantiated as prediction in representation space under a pointwise compatibility objective. More specifically, a context representation is used to predict a corresponding target representation from a paired view. Formally, given two related views x_1 and x_2, the encoder f_θ and predictor g_ϕ are learned by minimizing:

$$\mathcal{L}_{\text{JEPA}}(x_1, x_2) = d\big(g_\phi(f_\theta(x_1)),\, f_\xi(x_2)\big), \tag{1}$$

where f_ξ denotes a target encoder providing target embeddings for the predictive loss, and the metric $d(\cdot,\cdot)$ is generally some distance or similarity measure (e.g., cosine or Euclidean distance). Many non-contrastive objectives can be written in closely related paired-view forms, including BYOL (Grill et al., 2020) and SimSiam (Chen and He, 2021), which rely on asymmetric branches, as well as VICReg (Bardes et al., 2022) and Barlow Twins (Zbontar et al., 2021), which impose explicit variance/covariance or redundancy reduction penalties. Despite their empirical success, these approaches yield deterministic embeddings by mapping each input to a single point in latent space, and their pointwise energy/compatibility primitives cannot be defined through a *tractable* normalized probabilistic model in representation space (LeCun, 2022).

Variational inference and uncertainty quantification. Variational inference provides a general framework for probabilistic representation learning by representing each input as a distribution over latent factors. A canonical instance is the Variational Autoencoder (VAE) (Kingma and Welling, 2014), in which an encoder produces an approximate posterior $q_\phi(\boldsymbol{s} \mid x)$ over latent variables $\boldsymbol{s}$, and a decoder reconstructs the input via a generative model $p_\theta(x \mid \boldsymbol{s})$. Training proceeds by maximizing the evidence lower bound (ELBO):

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q_\phi(\boldsymbol{s}|x)}\big[\log p_\theta(x \mid \boldsymbol{s})\big] - D_{\text{KL}}\big(q_\phi(\boldsymbol{s} \mid x) \,\|\, p(\boldsymbol{s})\big), \tag{2}$$

where the prior $p(\boldsymbol{s})$ is typically taken to be a standard Gaussian (Kingma and Welling, 2014), $p(\boldsymbol{s}) = \mathcal{N}(0, I)$. To enable gradient-based optimization, the reparameterization trick (Kingma and Welling, 2014; Rezende et al., 2014):

$$\boldsymbol{s} = \boldsymbol{\mu}(x) + \boldsymbol{\sigma}(x) \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I) \tag{3}$$

is used to express stochastic samples as a differentiable transformation of $\boldsymbol{\epsilon}$. The encoder learns both a central estimate $\boldsymbol{\mu}$ and a feature-wise dispersion $\boldsymbol{\sigma}$, yielding an explicit posterior distribution over latent variables. The KL divergence term regularizes the posterior toward the prior. While this establishes a rigorous and tractable probabilistic foundation, its dependence on pixel-level reconstruction motivates alternative approaches suited to representation learning settings where such reconstruction is not required.

Variational inference in joint embedding architectures. Recent works have attempted to combine the efficiency of non-contrastive self-supervised learning with the probabilistic foundations of variational inference to produce uncertainty-aware representations, though several challenges remain. Notably, VI-SimSiam (Nakamura et al., 2023) modifies SimSiam by wrapping each unit-length embedding in a Power-Spherical (PS) (Cao et al., 2020) density parameterized by a mean direction $\boldsymbol{\mu}_i$ and concentration κ_i, with one branch's embedding frozen via stop-gradient and $\kappa_i = u_\theta(x_i)$ predicted by an additional scalar head. The resulting loss is defined as a PS log-likelihood applied to unit-norm embeddings:

$$\mathcal{L}_{\text{align}} = \tfrac{1}{2}\Big[- \log \text{PS}\big(\boldsymbol{z}_2; \boldsymbol{\mu}_1, \kappa_1\big) - \log \text{PS}\big(\boldsymbol{z}_1; \boldsymbol{\mu}_2, \kappa_2\big)\Big]. \tag{4}$$

In their variational interpretation, this likelihood term is augmented with a KL divergence between the PS density and a hyperspherical prior on $\mathbb{S}^{D-1}$, yielding an ELBO-like objective on unit embeddings. Mathematically, this defines a directional density on the unit sphere, with the scalar concentration κ_i controlling dispersion around $\boldsymbol{\mu}_i$. This is a coherent probabilistic formulation of non-contrastive alignment, though its uncertainty mechanism is inherently limited since κ_i is a single scalar, and decreasing it pushes the density toward the uniform distribution on $\mathbb{S}^{D-1}$, which directly weakens the alignment penalty on hard or noisy pairs. As a result, the learned κ_i largely functions as an example-dependent temperature controlling how strongly directional agreement is enforced, rather than as a structured uncertainty representation. Moreover, the scalar form restricts uncertainty to isotropic dispersion on the sphere and cannot represent feature-wise uncertainty.

VSSL (Yavuz and Yanikoglu, 2025) takes a different approach, whereby it couples a student encoder with a momentum-updated teacher, both outputting diagonal Gaussians over latent features. The teacher processes view x_1 to define a data-dependent "prior" $p_{\theta_t}(\boldsymbol{s} \mid x_1) = \mathcal{N}(\boldsymbol{s}; \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2)$, while the student processes view x_2 to produce a "posterior" $q_{\theta_s}(\boldsymbol{s} \mid x_2) = \mathcal{N}(\boldsymbol{s}; \boldsymbol{\mu}_2, \boldsymbol{\sigma}_2^2)$. The objective combines a likelihood term evaluating student samples under the teacher Gaussian:

$$\mathbb{E}_{q(\boldsymbol{s}|x_2)}\big[\log \mathcal{N}\big(\boldsymbol{s};\, \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2\big)\big], \tag{5}$$

with a KL penalty $\text{KL}\big(q_{\theta_s}(\boldsymbol{s} \mid x_2) \,\|\, p_{\theta_t}(\boldsymbol{s} \mid x_1)\big)$. This is framed as a self-supervised ELBO with the teacher as prior. However, since the teacher is defined by a moving-average update of the student, the KL term primarily enforces teacher–student consistency rather than regularization toward a fixed Bayesian prior.

More fundamentally, replacing the analytic Gaussian KL and log-likelihood with cosine-based alternatives breaks the probabilistic interpretation: these cosine quantities depend only on the directions of the mean and variance vectors and are not the KL divergence or log-likelihood of any Gaussian in $\mathbb{R}^D$. The resulting objective consequently becomes a heuristic angular-alignment loss, insensitive to parameter norms and at odds with the Euclidean geometry that underlies the Gaussian distribution. As a result, the variance parameters

are optimized as directional features rather than calibrated uncertainty measures, and their probabilistic semantics become unclear.

Collectively, these efforts highlight the potential of integrating variational inference into non-contrastive joint embedding architectures, while also demonstrating the difficulty of obtaining normalized probabilistic semantics when pointwise compatibility objectives remain the underlying training primitive. This motivates our work, in which a normalized latent-variable model is constructed from the ground up in representation space.

3 Model Architecture

Our implementation of the Variational Joint Embedding (VJE) framework follows the standard training structure of non-contrastive self-supervised learning (Chen and He, 2021; Grill et al., 2020), while adopting a normalized probabilistic latent-variable objective. A stochastic augmentation τ is applied twice to an input x to produce two views, $x_1 = \tau^{(1)}(x)$ and $x_2 = \tau^{(2)}(x)$, which are processed by a shared encoder $f_\theta : \mathcal{X} \to \mathbb{R}^D$ into deterministic embeddings $\mathbf{z}_1 = f_\theta(x_1)$ and $\mathbf{z}_2 = f_\theta(x_2)$. The framework consists of two asymmetric branches: an *inference branch* that maps $\mathbf{z}_i$ to the parameters of a stochastic latent code for view i, and a *target branch* that treats the opposite embedding $\mathbf{z}_j$ as a fixed observation during training via stop-gradient. Note that this asymmetry is used to implement the fixed-observation conditioning required by the conditional ELBO derived in Section 4, rather than serving as a stabilization heuristic.

Consistent with the theoretical derivations in Section 4, we represent each embedding by its direction $\hat{\mathbf{z}}$ and magnitude $\|\mathbf{z}\|$. This separation decouples directional matching from norm matching, allowing each to be modelled with a dedicated likelihood term that stabilizes optimization while respecting the geometry of the embedding space. We factorize the conditional likelihood into an extrinsic directional term (24) and a one-dimensional radial term (28), both instantiated as heavy-tailed Student–t likelihoods as detailed in Section 4. The overall architecture is depicted in Figure 1.

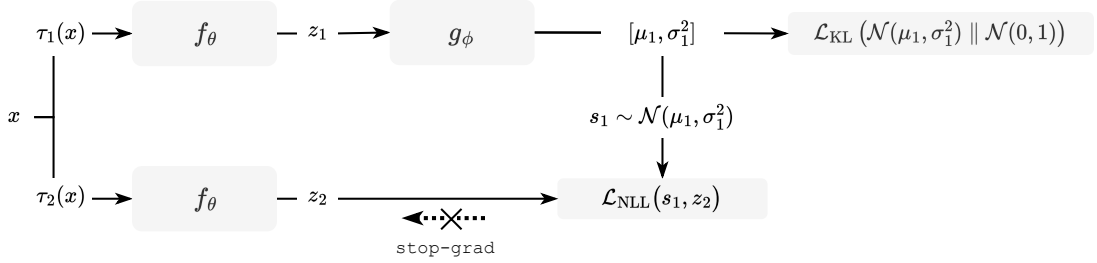


Figure 1: The asymmetric forward pass for one conditional direction in VJE, from view 1 to view 2. An encoder f_θ produces $\mathbf{z}_1$, and an amortized inference network g_ϕ maps it to a latent distribution $q_1(\mathbf{s}) = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2)$. A sample $\mathbf{s}_1$ is drawn and the conditional likelihood of the reparameterized target observation $y_2 = (\hat{\mathbf{z}}_2, \Delta r_{12})$ is evaluated under this latent code, where $\Delta r_{12} = \|\mathbf{z}_2\| - \|\mathbf{s}_1\|$. The target branch is detached (stop-gradient), enforcing fixed-observation semantics for the conditional likelihood term. The loss consists of directional (ℓ_{dir}) and radial (ℓ_{rad}) negative log-likelihoods (NLLs), jointly denoted $\mathcal{L}_{\mathrm{NLL}}$, together with a Kullback–Leibler (KL) divergence term $\mathcal{L}_{\mathrm{KL}}$.

Encoder and inference network. The encoder $f_\theta : \mathcal{X} \to \mathbb{R}^D$ is a shared backbone across views, and an inference network g_ϕ maps each $\mathbf{z}_i$ to the parameters of a diagonal Gaussian variational posterior, $q_i(\mathbf{s}) = \mathcal{N}\big(\boldsymbol{\mu}_i,\ \mathrm{diag}(\boldsymbol{\sigma}_i^2)\big)$, with $q_i(\mathbf{s}) \equiv q_\phi(\mathbf{s} \mid \mathbf{z}_i)$. While g_ϕ is architecturally akin to the predictor networks used in non-contrastive methods, we refer to it as an *amortized inference network* to reflect its variational role, since it parameterizes an instance-conditional posterior from which reparameterized samples are drawn and evaluated under the conditional likelihood. The network g_ϕ is implemented as a bottleneck MLP in

which each layer applies a linear transformation followed by layer normalization and a nonlinear activation function. Its final hidden representation is mapped through two linear output heads to produce $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i^2$.

The same variance vector $\boldsymbol{\sigma}_i^2$ is used both as the diagonal covariance of the variational posterior q_i and, via $\Sigma = \mathrm{diag}(\boldsymbol{\sigma}_i^2)$, as the whitening matrix in the directional Student–t likelihood (24). This tying ensures that feature-wise dispersion governs both the posterior sampling distribution and the conditional likelihood in a consistent manner. We adopt a diagonal covariance to keep the number of parameters linear in D and to ensure that sampling and KL evaluation scale linearly with the embedding dimension.

Unlike other joint-embedding architectures (Chen and He, 2021; Bardes et al., 2022; Grill et al., 2020; Assran et al., 2023), we do not introduce a separate projection head, as doing so would define an auxiliary representation space whose geometry is not constrained to relate to that of the encoder output. In VJE, both the conditional likelihood and the variational posterior are defined directly in the encoder's embedding space, so the inference network parameterizes latent structure within this same geometry rather than operating through an additional projection with unrelated semantics.

Latent sampling and likelihood evaluation. A latent sample $\mathbf{s}_i = \boldsymbol{\mu}_i + \boldsymbol{\sigma}_i \odot \boldsymbol{\varepsilon}_i$, with $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, is drawn using the reparameterization trick (Kingma and Welling, 2014; Rezende et al., 2014). Defining $\hat{\mathbf{s}}_i = \mathbf{s}_i / \|\mathbf{s}_i\|$, the negative log-likelihood averages directional and radial terms across both conditional directions $(i, j) \in \{(1, 2), (2, 1)\}$:

$$\mathcal{L}_{\mathrm{NLL}} = \tfrac{1}{2} \sum_{(i,j)\in\{(1,2),(2,1)\}} \mathbb{E}_{\mathbf{s}_i \sim q_i} \left[\ell_{\mathrm{dir}}\big(\mathrm{sg}(\hat{\mathbf{z}}_j), \hat{\mathbf{s}}_i; \boldsymbol{\sigma}_i^2\big) + \ell_{\mathrm{rad}}\big(\Delta r_{ij}\big) \right], \tag{6}$$

where $\mathrm{sg}(\cdot)$ denotes a stop-gradient and terms constant with respect to θ, ϕ are omitted for clarity. The *directional* term ℓ_{dir} is an extrinsic Student–t negative log-likelihood on unit vectors:

$$Q_{\mathrm{dir}} = \sum_{d=1}^{D} \frac{(\hat{s}_{i,d} - \hat{z}_{j,d})^2}{\sigma_{i,d}^2}, \qquad \ell_{\mathrm{dir}} = \tfrac{\nu+D}{2} \log\Big(1 + \tfrac{Q_{\mathrm{dir}}}{\nu}\Big) + \tfrac{1}{2} \sum_{d=1}^{D} \log \sigma_{i,d}^2. \tag{7}$$

The *radial* term ℓ_{rad} is a one-dimensional Student–t negative log-likelihood acting on the norm residual:

$$\Delta r_{ij} = \|\mathbf{z}_j\| - \|\mathbf{s}_i\|, \qquad \ell_{\mathrm{rad}} = \tfrac{\nu+1}{2} \log\Big(1 + \tfrac{\Delta r_{ij}^2}{\nu}\Big). \tag{8}$$

KL regularization and total objective. To anchor the posteriors, each $q_i(\mathbf{s})$ is regularized toward a standard Gaussian prior $p(\mathbf{s}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ using the analytic KL divergence:

$$\mathcal{L}_{\mathrm{KL}} = \frac{1}{2} \sum_{i=1}^{2} \sum_{d=1}^{D} \Big(\sigma_{i,d}^2 + \mu_{i,d}^2 - 1 - \log \sigma_{i,d}^2 \Big). \tag{9}$$

The final training loss combines the likelihood and regularization terms with a weighting factor β:

$$\mathcal{L} = \mathcal{L}_{\mathrm{NLL}} + \beta\, \mathcal{L}_{\mathrm{KL}}. \tag{10}$$

Appendix A provides pseudocode for the forward pass and loss computation of VJE.

4 Latent variable model

We begin our theoretical formulation by defining a likelihood model $p_\psi(\mathbf{z} \mid \mathbf{s})$ that formalizes the probabilistic relationship between the latent variable $\mathbf{s}$ and the observed embedding $\mathbf{z}$, both D-dimensional vectors in $\mathbb{R}^D$. Here, $\mathbf{z}$ is produced by the encoder and $\mathbf{s}$ is sampled from a variational posterior $q(\mathbf{s} \mid \mathbf{z})$, with the two corresponding to different views of the same input. The likelihood model evaluates how well a latent representation inferred from one view explains the embedding observed from another view. We ultimately evaluate the likelihood on the reparameterized observation $y = (\hat{\mathbf{z}}, \|\mathbf{z}\|)$, where $\hat{\mathbf{z}} := \mathbf{z}/\|\mathbf{z}\|$, and write the

final model as $p_\psi(y \mid \mathbf{s})$. This construction provides the foundation for our conditional evidence lower bound (ELBO) objective.

Our approach makes several explicit modelling choices that are motivated by geometric and statistical considerations. These choices are developed throughout this section and empirically evaluated in Section 5. The distinction between the normalized likelihood-based formulation presented here and pointwise energy-based objectives is formalized at the objective level in Appendix B.

Likelihood distribution. The choice of likelihood distribution $p_\psi(\mathbf{z} \mid \mathbf{s})$ is a central design element, as it determines how residuals between $\mathbf{z}$ and $\mathbf{s}$ are scored under a normalized density and how sensitive the resulting objective is to large deviations. To motivate the final form we adopt, we first examine the behaviour of a Gaussian likelihood and its limitations. Choosing a Gaussian likelihood yields:

$$p_\psi^{\mathcal{N}}(\mathbf{z} \mid \mathbf{s}) = (2\pi\lambda)^{-D/2} \exp\!\big(-\tfrac{1}{2\lambda}\|\mathbf{z}-\mathbf{s}\|^2\big)\,, \tag{11}$$

with corresponding negative log-likelihood (NLL) and gradient:

$$\ell_{\mathcal{N}}(\mathbf{z};\mathbf{s}) = \tfrac{1}{2\lambda}\|\mathbf{z}-\mathbf{s}\|^2, \qquad \nabla_{\mathbf{z}}\ell_{\mathcal{N}} = \tfrac{1}{\lambda}(\mathbf{z}-\mathbf{s}). \tag{12}$$

The gradient norm $\|\nabla_{\mathbf{z}}\ell_{\mathcal{N}}\|$ grows linearly with $\|\mathbf{z}-\mathbf{s}\|$, leading to unbounded influence from large residuals.

Since the Gaussian loss grows quadratically with $\|\mathbf{z}-\mathbf{s}\|$, large deviations in high-dimensional spaces can dominate training dynamics. To mitigate this behaviour in a model-based manner, we instead adopt a heavy-tailed Student–t likelihood. The Student–t distribution has a bounded influence function (Huber and Ronchetti, 2009), ensuring that large residuals continue to contribute to the objective without exerting unbounded influence on gradients or parameter estimates. This robustness arises from the probabilistic form of the likelihood itself, rather than from ad-hoc clipping or heuristic penalties. The resulting likelihood is given by (Kotz and Nadarajah, 2004; Bishop, 2006):

$$p_\psi^{t,\Sigma}(\mathbf{z} \mid \mathbf{s}) = C_{\nu,D}\,|\Sigma|^{-1/2}\left[1 + \tfrac{1}{\nu}(\mathbf{z}-\mathbf{s})^\top\Sigma^{-1}(\mathbf{z}-\mathbf{s})\right]^{-\frac{\nu+D}{2}}, \tag{13}$$

yielding NLL and gradient functions:

$$\begin{aligned} \ell_{t,\Sigma}(\mathbf{z};\mathbf{s}) &= \tfrac{\nu+D}{2}\log\!\big(1 + \tfrac{1}{\nu}(\mathbf{z}-\mathbf{s})^\top\Sigma^{-1}(\mathbf{z}-\mathbf{s})\big) + \tfrac{1}{2}\log|\Sigma|, \\ \nabla_{\mathbf{z}}\ell_{t,\Sigma} &= \frac{\nu+D}{\nu + (\mathbf{z}-\mathbf{s})^\top\Sigma^{-1}(\mathbf{z}-\mathbf{s})}\,\Sigma^{-1}(\mathbf{z}-\mathbf{s}), \end{aligned} \tag{14}$$

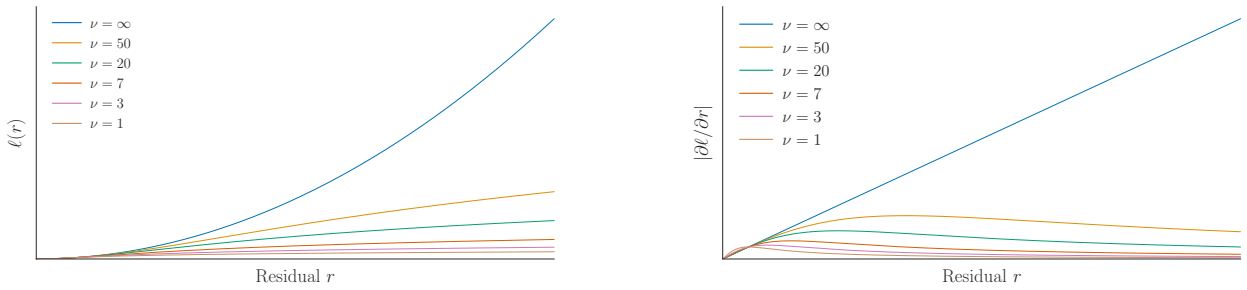
where $\nu > 0$ controls tail heaviness, Σ is a symmetric positive-definite scale matrix, and $C_{\nu,D}$ is the normalizing constant. As $\nu \to \infty$, the Student–t likelihood recovers the Gaussian case.

This formulation ensures that the NLL grows only logarithmically for large residuals, while the gradient magnitude decays with the Mahalanobis distance $(\mathbf{z}-\mathbf{s})^\top\Sigma^{-1}(\mathbf{z}-\mathbf{s})$. Figure 2 illustrates this behaviour, highlighting the bounded influence of the Student–t likelihood compared to the unbounded Gaussian case.

Although the Student–t likelihood provides robustness to large deviations, the Mahalanobis term $(\mathbf{z}-\mathbf{s})^\top\Sigma^{-1}(\mathbf{z}-\mathbf{s})$ still couples angular misalignment and differences in norms into a single error channel. Expanding this term,

$$(\mathbf{z}-\mathbf{s})^\top\Sigma^{-1}(\mathbf{z}-\mathbf{s}) = \mathbf{z}^\top\Sigma^{-1}\mathbf{z} + \mathbf{s}^\top\Sigma^{-1}\mathbf{s} - 2\,\mathbf{z}^\top\Sigma^{-1}\mathbf{s}, \tag{15}$$

reveals both the individual quadratic norms $\mathbf{z}^\top\Sigma^{-1}\mathbf{z}$, $\mathbf{s}^\top\Sigma^{-1}\mathbf{s}$ and their inner product $\mathbf{z}^\top\Sigma^{-1}\mathbf{s}$. As a consequence, the contribution of an angular discrepancy between $\mathbf{z}$ and $\mathbf{s}$ is scaled by their norms (in the Σ^{-1} metric), so that large-norm embeddings can produce large Mahalanobis residuals even for moderate angular error. In contrast, a large Mahalanobis residual does not reveal whether it is dominated by a mismatch in direction or magnitude. This coupling between scale and orientation motivates a reparameterization in which directional and radial contributions are modelled by separate likelihood factors. To tackle this, we reformulate the likelihood in a space where angular and radial variations are explicitly decoupled, and so



(a) Negative log-likelihood $\ell(r)$

(b) Gradient magnitude $|\partial \ell / \partial r|$

Figure 2: One-dimensional views of the Student–t loss used in VJE, plotted as functions of the residual r for different degrees of freedom ν (with the Gaussian limit at $\nu = \infty$). Panel (a) illustrates how heavy tails moderate the growth of the negative log-likelihood for large residuals, while panel (b) shows the corresponding influence functions, where gradients saturate and then decay so that outliers contribute only a bounded amount of signal. This underlines the choice of Student–t likelihoods in VJE to stabilize training without ad-hoc heuristics.

each embedding is represented not by its Euclidean coordinates, but by its *direction* and *magnitude*. In the radial channel, discrepancies are evaluated relative to the predicted norm via the residual:

$$\Delta r := \|\mathbf{z}\| - \|\mathbf{s}\| \in \mathbb{R}. \tag{16}$$

This representation allows us to treat directional and radial channels separately in the likelihood, each with its own stability properties and normalization behaviour. We define a product-form likelihood on the observation $y = (\hat{\mathbf{z}}, \|\mathbf{z}\|)$, with the radial factor expressed in terms of the residual Δr.

In Section 4.1, we develop this decomposition by examining how a polar factorization of the isotropic Student–t distribution motivates the separation of directional and radial channels. In Section 4.2, we recover feature-wise (i.e., anisotropic) uncertainty by introducing directional whitening through a diagonal variance vector $\boldsymbol{\sigma}^2$, which allows the model to weight dimensions according to their uncertainty. Section 4.3 formalizes the radial residual Δr and its associated likelihood, and Section 4.4 establishes the conditional evidence lower bound (ELBO) that combines these components into a symmetric objective suitable for joint embedding architectures.

4.1 Polar decomposition of the likelihood

To motivate the separation of directional and radial channels, we begin from an isotropic Student–t likelihood, whose rotational symmetry admits a well-defined polar decomposition into independent radial and directional factors (Mardia and Jupp, 2000). The isotropic Student–t likelihood can be written as:

$$p_{\psi}^{t,\lambda}(\mathbf{z} \mid \mathbf{s}) = C_{\nu,D}\, \lambda^{-D/2} \left[1 + \frac{\|\mathbf{z} - \mathbf{s}\|^2}{\nu\lambda}\right]^{-\frac{\nu+D}{2}}, \tag{17}$$

where $\lambda > 0$ is a scalar scale parameter. This corresponds to the isotropic special case $\Sigma = \lambda \mathbf{I}$ of the elliptical Student–t likelihood in Eq. (13).

We introduce polar coordinates for the displacement vector $\mathbf{z} - \mathbf{s}$:

$$\rho := \|\mathbf{z} - \mathbf{s}\| \in (0, \infty), \qquad \boldsymbol{\omega} := \frac{\mathbf{z} - \mathbf{s}}{\|\mathbf{z} - \mathbf{s}\|} \in \mathbb{S}^{D-1}, \tag{18}$$

and the corresponding Jacobian for this change of variables:

$$d\mathbf{z} = \rho^{D-1}\, d\rho\, d\boldsymbol{\omega}, \tag{19}$$

where $d\boldsymbol{\omega}$ denotes the uniform surface measure on the unit sphere $\mathbb{S}^{D-1}$ (Mardia and Jupp, 2000). Since the isotropic density in Eq. (17) depends only on ρ, the direction $\boldsymbol{\omega}$ is uniformly distributed and independent of ρ. The likelihood therefore factorizes into independent radial and directional components:

$$p_{\psi}^{t,\lambda}(\mathbf{z} \mid \mathbf{s}) = p_{\mathrm{rad}}(\rho)\, p_{\mathrm{dir}}(\boldsymbol{\omega}), \qquad p_{\mathrm{dir}}(\boldsymbol{\omega}) = \frac{1}{\mathrm{vol}(\mathbb{S}^{D-1})}. \tag{20}$$

The explicit form of the radial factor is given by:

$$p_{\mathrm{rad}}(\rho) = \frac{2\,\Gamma\big(\frac{\nu+D}{2}\big)}{\Gamma\big(\frac{\nu}{2}\big)\,\Gamma\big(\frac{D}{2}\big)\,(\nu\lambda)^{D/2}}\,\rho^{D-1}\left[1+\frac{\rho^2}{\nu\lambda}\right]^{-\frac{\nu+D}{2}}, \qquad \rho > 0, \tag{21}$$

which provides a clear geometric interpretation, as the isotropic Student–t distributes mass uniformly over directions, and the entire radial structure is captured by the corresponding one-dimensional term.

This polar decomposition motivates our use of a product-form likelihood (i.e., joint likelihood) with independent directional and radial factors, enabling separate treatment of angular and radial errors. While the isotropic case yields a well-defined factorization, our full model extends this structure to incorporate anisotropic scaling and alternative radial parameterizations. This design choice provides flexibility to model uncertainty while retaining probabilistic and geometric coherence. The specific parameterization of these factors is developed in subsequent sections.

4.2 Directional whitening and feature-wise uncertainty

An immediate limitation of the isotropic form derived in Section 4.1 is the shared scale λ, which prevents the likelihood from expressing feature-wise variation. We enable anisotropic weighting of dimensions by introducing a diagonal variance vector $\boldsymbol{\sigma}^2 \in \mathbb{R}_+^D$ in the directional term through per-dimension whitening. In the polar parametrization of the displacement $\mathbf{z}-\mathbf{s}$, the radial variable $\rho = \|\mathbf{z}-\mathbf{s}\|$ is one-dimensional, so any anisotropy is naturally confined to the directional term. We define the directional scale matrix as $\Sigma = \mathrm{diag}(\boldsymbol{\sigma}^2)$ and the corresponding whitening operator $W = \mathrm{diag}(\boldsymbol{\sigma}^{-1})$.

The directional discrepancy is computed by first normalizing each embedding to unit length and then applying coordinate-wise whitening:

$$\hat{\mathbf{z}} = \frac{\mathbf{z}}{\|\mathbf{z}\|}, \qquad \hat{\mathbf{s}} = \frac{\mathbf{s}}{\|\mathbf{s}\|}, \qquad \mathbf{u} = W(\hat{\mathbf{z}}-\hat{\mathbf{s}}), \qquad Q_{\mathrm{dir}} = \|\mathbf{u}\|^2. \tag{22}$$

Here, Q_{dir} measures the normalized squared difference after rescaling each coordinate by its variance.

To implement this whitening of the isotropic scale $\lambda\mathbf{I}$ to the diagonal matrix $\Sigma = \mathrm{diag}(\boldsymbol{\sigma}^2)$, we work with an extrinsic directional representation in $\mathbb{R}^D$ rather than an intrinsic one on the hypersphere $\mathbb{S}^{D-1}$. An intrinsic construction would require explicit Jacobian terms from Eq. (19) and would complicate the whitening operator, as the diagonal matrix $W = \mathrm{diag}(\boldsymbol{\sigma}^{-1})$ would need to be projected onto tangent spaces at each point. In contrast, the extrinsic formulation treats unit-normalized embeddings as vectors in the ambient space and absorbs the sphere Jacobian into the normalizing constant, allowing the whitening transformation to be applied directly in $\mathbb{R}^D$. This also gives each variance component σ_d^2 a clear interpretation as uncertainty along a specific encoder feature, whereas on the sphere, coordinates are arbitrary due to rotational symmetry.

We model the directional component directly by an *extrinsic* Student-t distribution in $\mathbb{R}^D$:

$$p_{\mathrm{dir}}(\hat{\mathbf{z}} \mid \hat{\mathbf{s}}, \boldsymbol{\sigma}^2) = C_{\nu,D}\,|\Sigma|^{-1/2}\big(1+\tfrac{1}{\nu}Q_{\mathrm{dir}}\big)^{-\frac{\nu+D}{2}}, \tag{23}$$

which retains heavy-tailed robustness while simplifying computation. From this expression, the per-sample negative log-likelihood, dropping constants independent of $(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, becomes:

$$\ell_{\mathrm{dir}} = \frac{\nu+D}{2}\log\Big(1+\tfrac{1}{\nu}Q_{\mathrm{dir}}\Big) + \frac{1}{2}\log|\Sigma| \;=\; \frac{\nu+D}{2}\log\Big(1+\tfrac{1}{\nu}Q_{\mathrm{dir}}\Big) + \frac{1}{2}\sum_{d=1}^{D}\log\sigma_d^2. \tag{24}$$

The quadratic term Q_{dir} down-weights high-variance dimensions and emphasizes low-variance ones, while the $\frac{1}{2}\log|\Sigma|$ term penalizes variance inflation.

To ensure that feature-wise uncertainty jointly governs the directional likelihood and the variational posterior, we tie the variance vector $\boldsymbol{\sigma}^2$ between the two distributions. The inference network output $\boldsymbol{\sigma}^2$ is shared across the Gaussian posterior $q(\mathbf{s} \mid \mathbf{z})$ and the directional likelihood $p_{\text{dir}}(\hat{\mathbf{z}} \mid \hat{\mathbf{s}}, \boldsymbol{\sigma}^2)$, with the same matrix $\Sigma = \text{diag}(\boldsymbol{\sigma}^2)$ appearing in both. This choice makes each σ_d^2 a shared per-feature scale parameter: it controls the weighting of the directional residual in the likelihood while also serving as the variance parameter of the posterior, so that feature-wise uncertainty is expressed consistently across both distributions.

4.3 Radial reparameterization and final likelihood

While the polar factorization in Section 4.1 yields a one-dimensional radial variable $\rho = \|\mathbf{z} - \mathbf{s}\|$, using this Euclidean distance directly as the radial term still couples angular misalignment with differences in norms. The squared distance expansion

$$\|\mathbf{z} - \mathbf{s}\|^2 \;=\; \|\mathbf{z}\|^2 + \|\mathbf{s}\|^2 - 2\,\|\mathbf{z}\|\,\|\mathbf{s}\|\,\cos\theta \tag{25}$$

reveals that the Euclidean distance inherently couples angular alignment with magnitude through the cosine term, so angular discrepancies contribute to the radial error in proportion to the product of norms. For a Student-t likelihood, the angular gradient takes the form

$$\frac{\partial \ell_t}{\partial \theta} \;=\; \frac{\nu + D}{\nu\lambda + \|\mathbf{z} - \mathbf{s}\|^2}\,\|\mathbf{z}\|\,\|\mathbf{s}\|\,\sin\theta, \tag{26}$$

which remains susceptible to norm amplification despite the bounded prefactor.

To address this coupling, we reparameterize the radial channel as the *difference of norms*:

$$\Delta r \;:=\; \|\mathbf{z}\| - \|\mathbf{s}\|. \tag{27}$$

This parameterization measures magnitude discrepancy independently of angular alignment: the radial residual is zero when predicted and observed magnitudes agree, regardless of their relative directions. The reference point shifts from the Euclidean distance $\|\mathbf{z} - \mathbf{s}\|$ to the predicted norm $\|\mathbf{s}\|$, enabling a clean separation of scale from orientation in the radial term.

Consequently, translation invariance is no longer preserved, since $\|(\mathbf{z} + \mathbf{a}) - (\mathbf{s} + \mathbf{a})\| = \|\mathbf{z} - \mathbf{s}\|$ but $\|\mathbf{z} + \mathbf{a}\| - \|\mathbf{s} + \mathbf{a}\| \neq \|\mathbf{z}\| - \|\mathbf{s}\|$. However, this does not introduce inconsistencies, as we anchor both the posterior $q(\mathbf{s} \mid \mathbf{z})$ and the likelihood $p_\psi(\mathbf{z} \mid \mathbf{s})$ to the origin by introducing the standard Gaussian prior in Section 4.4. Since both the likelihood and the posterior are defined relative to this fixed coordinate system, the reparameterization remains geometrically consistent with the choice of a fixed origin.

The resulting radial likelihood is given by a one-dimensional Student-t kernel

$$p_{\text{rad}}^{(\Delta)}(\Delta r) \;=\; \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi\lambda}\,\Gamma(\frac{\nu}{2})}\left(1 + \frac{(\Delta r)^2}{\nu\lambda}\right)^{-\frac{\nu+1}{2}}, \qquad \Delta r \in \mathbb{R}. \tag{28}$$

Therefore the per-sample NLL, omitting constants, is

$$\ell_{\text{rad}}(\Delta r) = \frac{\nu+1}{2}\log\!\left(1 + \frac{\Delta r^2}{\nu\lambda}\right), \tag{29}$$

with the derivative of its loss

$$\left|\frac{\partial}{\partial\,\Delta r}\big[-\log p_{\text{rad}}^{(\Delta)}(\Delta r)\big]\right| \;=\; \frac{\nu+1}{\nu\lambda}\,\frac{|\Delta r|}{1 + (\Delta r)^2/(\nu\lambda)} \;\leq\; \frac{\nu+1}{2\sqrt{\nu\lambda}}, \tag{30}$$

which is bounded for any finite ν. In the radial kernel $p_{\text{rad}}^{(\Delta)}(\Delta r)$, the parameter $\lambda > 0$ acts purely as a global scale: the density depends on Δr only through the combination $\Delta r^2/(\nu\lambda)$. A change in λ is therefore

equivalent to a rescaling of the radial coordinate and does not increase the expressiveness of the model. Moreover, for any fixed $\Delta r \neq 0$ and $\nu > 0$, increasing λ decreases the radial NLL $\ell_{\mathrm{rad}}(\Delta r)$. An unconstrained maximum-likelihood solution would therefore push $\lambda \to \infty$, effectively eliminating the radial penalty. To avoid this degenerate path, we fix the radial scale to $\lambda = 1$, and leave ν as the sole shared radial hyperparameter.

Combining the stabilized radial term (28) with the directional factor (23) yields the product likelihood used in training on the observation $y = (\hat{\mathbf{z}}, \|\mathbf{z}\|)$:

$$p_\psi(y \mid \mathbf{s}, \boldsymbol{\sigma}^2) \;=\; p_{\mathrm{rad}}^{(\Delta)}(\Delta r) \cdot p_{\mathrm{dir}}(\hat{\mathbf{z}} \mid \hat{\mathbf{s}}, \boldsymbol{\sigma}^2), \qquad \Delta r := \|\mathbf{z}\| - \|\mathbf{s}\|. \tag{31}$$

Constants independent of $(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ are omitted from the loss. The Jacobian terms associated with the change of variables from $\mathbf{z}$ to $(\hat{\mathbf{z}}, \|\mathbf{z}\|)$ do not depend on the trainable parameters and are absorbed into these omitted constants.

Together, the radial factor derived in this section and the directional factor defined in Eq. (23) specify the complete likelihood model $p_\psi(y \mid \mathbf{s}, \boldsymbol{\sigma}^2)$ on representation-space observations.

4.4 Variational posterior and evidence lower bound

To establish the evidence lower bound (ELBO) underlying the VJE objective, we combine the product-form likelihood $p_\psi(y \mid \mathbf{s}, \boldsymbol{\sigma}^2)$ of Eq. (31), defined on the observation $y = (\hat{\mathbf{z}}, \|\mathbf{z}\|)$, with a diagonal-Gaussian variational posterior. This formulation defines a conditional variational inference framework in which each view predicts the latent representation of its paired view under the factorized likelihood.

For each view $i \in \{1, 2\}$, we denote by q_i the variational posterior associated with $\mathbf{z}_i$, parameterized as a diagonal Gaussian

$$q_i(\mathbf{s}) = \mathcal{N}\big(\boldsymbol{\mu}_i,\, \mathrm{diag}(\boldsymbol{\sigma}_i^2)\big)\,, \tag{32}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i^2$ are produced by the inference network for view i, and the variance $\boldsymbol{\sigma}_i^2$ is shared with the directional likelihood introduced in Section 4.2. Samples are drawn via the reparameterization trick (Kingma and Welling, 2014; Rezende et al., 2014):

$$\mathbf{s}_i = \boldsymbol{\mu}_i + \boldsymbol{\sigma}_i \odot \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{33}$$

and regularized toward the standard normal prior $p(\mathbf{s}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ using the analytic KL divergence

$$\mathcal{L}_{\mathrm{KL}}(q_i \| p) = \tfrac{1}{2} \sum_{d=1}^{D} \left(\sigma_{id}^2 + \mu_{id}^2 - 1 - \log \sigma_{id}^2\right). \tag{34}$$

We use a fixed standard Gaussian prior rather than a learned or data-dependent prior. This choice provides a consistent reference frame that anchors all embeddings to the origin, enabling the norm-difference radial parameterization developed in Section 4.3.

For the variational posterior, we adopt a diagonal Gaussian to keep the number of covariance parameters and the cost of sampling and KL evaluation linear in D, and to reuse the same diagonal scale matrix $\Sigma = \mathrm{diag}(\boldsymbol{\sigma}^2)$ that appears in the Student–t directional likelihood. This tying makes each σ_d^2 a shared per-feature scale parameter across likelihood and posterior, ensuring that feature-wise uncertainty is parameterized consistently in both components of the model.

For each conditional direction $i \to j$ with $(i, j) \in \{(1, 2), (2, 1)\}$, we treat the target embedding $\mathbf{z}_j$ $(j \neq i)$ as a fixed observation by detaching it from the computation graph. This implements the conditioning semantics of the variational posterior q_i predicting the opposite view: in the conditional ELBO, y_j appears only as an observed variable in the likelihood term and is not a parameter to be optimized. The one-way conditional ELBO is then

$$\mathcal{F}_{i \to j} = \mathbb{E}_{\mathbf{s}_i \sim q_i}\left[\log p_{\mathrm{rad}}^{(\Delta)}(\|\mathbf{z}_j\| - \|\mathbf{s}_i\|) + \log p_{\mathrm{dir}}\big(\hat{\mathbf{z}}_j \mid \hat{\mathbf{s}}_i,\, \boldsymbol{\sigma}_i^2\big)\right] - \mathcal{L}_{\mathrm{KL}}(q_i \| p), \tag{35}$$

which, by Jensen's inequality (Kingma and Welling, 2014; Rezende et al., 2014), provides a lower bound on the conditional log-likelihood $\log p_\psi(y_j \mid \mathbf{z}_i)$ for the observation

$$y_j = \big(\hat{\mathbf{z}}_j,\, \|\mathbf{z}_j\|\big). \tag{36}$$

Since the Jacobian of the mapping from $\mathbf{z}_j$ to $(\hat{\mathbf{z}}_j, \|\mathbf{z}_j\|)$ does not depend on the model parameters $(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2)$, the same bound holds for $\log p_\psi(\mathbf{z}_j \mid \mathbf{z}_i)$ up to an additive constant that is omitted from the loss.

Symmetrizing over both directions yields the β-weighted objective

$$\overline{\mathcal{F}}^{(\beta)} = \frac{1}{2} \sum_{(i,j)\in\{(1,2),(2,1)\}} \left\{ \mathbb{E}_{\mathbf{s}_i\sim q_i} \left[\log p_{\mathrm{rad}}^{(\Delta)}(\|\mathbf{z}_j\| - \|\mathbf{s}_i\|) + \log p_{\mathrm{dir}}\big(\hat{\mathbf{z}}_j \mid \hat{\mathbf{s}}_i,\, \boldsymbol{\sigma}_i^2\big) \right] - \beta\, \mathcal{L}_{\mathrm{KL}}(q_i \| p) \right\}. \tag{37}$$

Here, β weights the KL regularization analogously to β-VAE (Higgins et al., 2017): $\beta = 1$ corresponds to the unweighted conditional ELBO, $\beta < 1$ emphasizes likelihood fitting, and $\beta > 1$ increases regularization strength.

For training, we minimize the negative of Eq. (37). Writing the likelihood terms as their negative log-likelihoods (NLLs) gives

$$\mathcal{L}_{\mathrm{train}} = \underbrace{\frac{1}{2} \sum_{(i,j)\in\{(1,2),(2,1)\}} \mathbb{E}_{\mathbf{s}_i\sim q_i} \left[\ell_{\mathrm{dir}}\big(\hat{\mathbf{z}}_j, \hat{\mathbf{s}}_i;\, \boldsymbol{\sigma}_i^2\big) + \ell_{\mathrm{rad}}(\Delta r_{ij}) \right]}_{\mathcal{L}_{\mathrm{NLL}}} + \beta \underbrace{\frac{1}{2} \sum_{i=1}^{2} \mathcal{L}_{\mathrm{KL}}(q_i \| p)}_{\mathcal{L}_{\mathrm{KL}}}, \tag{38}$$

where $\Delta r_{ij} := \|\mathbf{z}_j\| - \|\mathbf{s}_i\|$. Here, $\mathbf{z}_j$ is always detached, ensuring that gradients flow only through the inference network of branch i, consistent with the conditional ELBO formulation in which $\mathbf{z}_j$ is treated as a fixed observation. Note that constants independent of $(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ are dropped from the loss.

5 Experiments

We evaluate the proposed Variational Joint Embedding (VJE) framework along two complementary axes. First, we assess whether VJE retains discriminative representation quality comparable to deterministic joint-embedding baselines, establishing that the normalized likelihood objective supports competitive feature learning. We evaluate this on both large-scale ImageNet-1K (Russakovsky et al., 2015) and smaller-scale CIFAR–10 (Krizhevsky, 2009) benchmarks (Section 5.2), and perform related ablations to quantify the contributions of key design choices (Section 5.4) using CIFAR–10, CIFAR–100 (Krizhevsky, 2009), and STL–10 (Coates et al., 2011) datasets. Second, and more centrally, we test whether VJE yields probabilistically coherent uncertainty estimates through one-class anomaly detection on CIFAR–10 (Section 5.3), evaluating joint-likelihood scores (directional + radial) and related uncertainty signals (entropy, variance) using AUROC averaged over classes. We sweep the KL weight β and the Student–t degrees of freedom ν on a grid to map the operating regime, and report class-averaged trends, per-class results, and diagnostics across the (β, ν) surface.

5.1 Experimental setup

All experiments use a ResNet backbone (He et al., 2016) for consistency, followed by an inference network that outputs posterior parameters $(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ as described in Section 3. For ImageNet-1K (Russakovsky et al., 2015) we use a ResNet–50 backbone as our encoder, while for CIFAR–10, CIFAR–100, and STL–10 we use a smaller ResNet–18 model. For CIFAR–10 and CIFAR–100, the architecture is adapted to 32×32 inputs by modifying the first convolutional layer to 3×3 with stride 1 and removing the initial max-pooling layer. The inference network is a three-layer bottleneck MLP interleaved with Layer Normalization and nonlinear activation, with the bottleneck dimension set to 512 for ResNet–50 and 128 for ResNet–18. Two separate linear heads then produce the posterior mean $\boldsymbol{\mu}$ and diagonal variance $\boldsymbol{\sigma}^2$. Recall that the standard implementation does not include a separate projection head, as the inference network directly models the latent distribution over encoder features.

5.2 Representation learning and discriminative capacity

ImageNet-1K. We begin with a large-scale representation learning experiment on ImageNet-1K (Russakovsky et al., 2015) to assess whether introducing a variational objective fundamentally compromises

discriminative capacity. Training follows a 100-epoch pretraining schedule with ResNet–50 and the standard MoCo v2 augmentation recipe (Chen et al., 2020b). We use stochastic gradient descent with momentum 0.9, batch size 256, cosine learning-rate decay (Loshchilov and Hutter, 2017) from an initial value of 0.05, and weight decay of 5×10^{-6}, with a 10-epoch linear warm-up preceding the cosine phase (Goyal et al., 2017). We set β=1.0 to match the exact symmetric conditional ELBO formulation and ν=3.0 for a moderately heavy-tailed Student–t likelihood. After pretraining, we train a linear classifier on frozen backbone features for 100 epochs using the same optimizer configuration, excluding the warm-up phase.

Table 1: Linear evaluation of comparable non-contrastive methods on ImageNet-1K, with ResNet–50 and 100 pretraining epochs. VJE (ours) was pretrained and evaluated over three runs to obtain mean and standard deviation. All baseline values are taken directly from cited sources where available. Entries for *SimCLR* and *BYOL* are 100-epoch results reproduced by Chen and He (2021) and marked with an asterisk (*).

Method	**Top-1 (%)**
SimCLR (Chen et al., 2020a)	66.5*
BYOL (Grill et al., 2020)	66.5*
SwAV (Caron et al., 2020)	66.5
SimSiam (Chen and He, 2021)	68.1
VICReg (Bardes et al., 2022)	68.6
VJE (ours)	65.6 ± 0.4

VJE achieves 65.6% top-1 accuracy (Table 1) and 86.8% top-5 accuracy after 100 epochs of pretraining, retaining discriminative capacity comparable to deterministic baselines under the same training configuration. The results demonstrate that the stochasticity introduced by the variational formulation does not fundamentally compromise representation quality, and so the model is able to learn useful features while simultaneously capturing uncertainty information, although a modest gap to the strongest deterministic baselines remains.

CIFAR–10. We next evaluate VJE on CIFAR–10 (Krizhevsky, 2009) using ResNet–18 adapted for 32×32 inputs. Training runs for 800 epochs with the MoCo v2 augmentation recipe (Chen et al., 2020b) excluding Gaussian blur, which is less effective at this resolution. We use stochastic gradient descent with momentum 0.9, cosine learning-rate decay from an initial value of 0.05, batch size 256, and weight decay 5×10^{-6}. As before, we set β=1.0 and ν=3.0.

As baselines, we reproduce SimSiam (Chen and He, 2021) with a 2-layer 2048-dimensional projector and VICReg (Bardes et al., 2022) with a 3-layer 4096-dimensional projector, matching their original architectural choices while using the common training schedule and augmentation pipeline described above.

Table 2: CIFAR–10 evaluation with ResNet–18 after 800 epochs. VJE is evaluated on both encoder output z and posterior mean μ using k–NN ($k = 30$) and a linear probe trained on frozen backbone features. All results are obtained under identical training conditions.

Method	**Projector**	**k–NN Top-1 (%)**	**Linear Top-1 (%)**
SimSiam (reproduced)	2048-dim, 2-layer	90.60	91.8
VICReg (reproduced)	4096-dim, 3-layer	86.37	89.4
VJE (ours), z	—	89.98	92.1
VJE (ours), μ	—	89.02	90.8

Table 2 shows that VJE trails SimSiam by approximately 0.6 percentage points in k–NN accuracy ($k = 30$) while outperforming VICReg by 3.6 points. For linear evaluation, VJE achieves 92.1% accuracy on the encoder output, outperforming both SimSiam (91.8%) and VICReg (89.4%). The close alignment between k–NN and linear evaluation performance on z (89.98%, 92.1%) and μ (89.02%, 90.8%) indicates that the variational posterior concentrates around the encoder embedding while maintaining non-trivial variance, as

we further explore in the anomaly detection experiments (Section 5.3). Figure 3 shows the k–NN training trajectories, with both z and μ curves exhibiting stable convergence throughout the 800-epoch schedule.

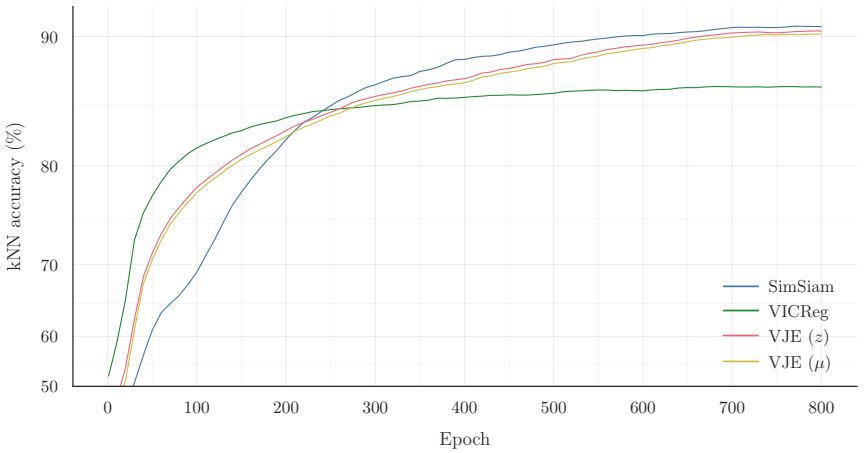


Figure 3: k–NN accuracy over training on CIFAR–10 for SimSiam, VICReg, and VJE ($k = 30$). VJE curves are shown for both encoder output z and posterior mean μ, demonstrating stable convergence and close alignment between the two representations throughout training.

Projection heads in VJE. Although VJE defines the latent variable model and conditional ELBO directly on the encoder feature space, we investigate whether a projection head offers empirical benefits, analogous to its role in deterministic self-supervised methods. We train a variant mirroring SimSiam by inserting a 2-layer, 2048-dimensional projection MLP after the encoder. To ensure posterior statistics remain independent of batch composition, we replace all Batch Normalization layers in the projector and the inference network (matching the SimSiam predictor) with Layer Normalization. This variant attains a final k–NN accuracy of approximately 89.5% at epoch 800, slightly below the projection-free baseline (89.98%), suggesting that the projection head provides no tangible benefit despite the added architectural complexity.

5.3 Anomaly detection and probabilistic semantics

To probe the probabilistic semantics of VJE that are central to its design, we adopt a one-class anomaly detection protocol on CIFAR–10 (Krizhevsky, 2009), where a model is trained on a single class treated as normal and evaluated on its ability to separate inliers (the training class) from outliers (the other 9 classes) at test time. A well-defined probabilistic model is expected to assign high likelihood to inliers and low likelihood to outliers. This protocol directly tests the density estimation capacity and likelihood structure developed in Section 4.

Our experiment follows a one-vs-rest CIFAR–10 anomaly detection protocol commonly used in prior work (Ruff et al., 2018; Tack et al., 2020), where we train 10 models independently, one per class, and evaluate on the full test set, measuring AUROC for separating the training class (inlier) from the other 9 classes (outliers). Each model uses ResNet–18 with the architecture described in Section 5.1. Training runs for 800 epochs with SGD using momentum 0.9, batch size 64, learning rate $5{\times}10^{-6}$ with cosine decay, and weight decay $5{\times}10^{-6}$. The reduced batch size and learning rate reflect the one-class setting, where training data has limited diversity. We use the same MoCo v2 augmentation recipe (Chen et al., 2020b) as in Section 5.2, again excluding Gaussian blur.

We sweep $\beta \in \{0, 0.5, 0.75, 1.0, 1.25, 1.5, 2.0\}$ and $\nu \in \{0.25, 0.5, 1.0, 2.0, 3.0, 5.0, 10.0, 20.0\}$ across all 10 classes, producing a 7×8 grid of 56 configurations averaged across each of the 10 one-vs-rest splits (560 models total). The primary anomaly score is the joint negative log-likelihood, matching the per-sample NLL terms in the training objective of Eq. (38):

$$S_{\text{joint}}(\mathbf{x}) = \ell_{\text{dir}}(\hat{\mathbf{z}}, \hat{\mathbf{s}}; \boldsymbol{\sigma}^2) + \ell_{\text{rad}}(\Delta r), \qquad \Delta r := \|\mathbf{z}\| - \|\mathbf{s}\|, \tag{39}$$

where higher values indicate greater anomaly. This score aggregates directional and radial contributions under the learned model.

Figure 4: CIFAR–10 one-class detection: class-averaged AUROC across the $\beta \times \nu$ grid using the joint-likelihood score (Eq. 39). Lighter regions indicate higher AUROC. The optimal regime concentrates at $\beta \approx 1.0$ and small ν.

Figure 4 shows that AUROC reaches a broad optimum near β=1.0 (unweighted ELBO) and small $\nu \in \{0.25, 0.5, 1.0\}$ (strong heavy tails). At $(\beta, \nu) = (1.0, 0.5)$, the class-averaged AUROC is **0.903** with an across-class standard deviation of 0.043. Neighbouring settings $(1.0, 1.0)$ and $(1.0, 0.25)$ yield 0.899 and 0.897, respectively, indicating a stable operating region forming the ridge in Figure 4. These results empirically confirm the role of heavy tails as anticipated in Section 4: down-weighting large errors through a Student–t likelihood improves generalization at the inlier-outlier boundary and prevents overfitting to inlier noise. The consistency across classes indicates that the optimal hyperparameters generalize well without per-class tuning.

Table 3: Per-class AUROC at $(\beta, \nu) = (1.0, 0.5)$ using the joint-likelihood score.

Class	airp.	auto.	bird	cat	deer	dog	frog	horse	ship	truck
AUROC	0.872	0.976	0.851	0.828	0.884	0.897	0.909	0.935	0.940	0.937

Attempts to use Gaussian likelihoods ($\nu \to \infty$) resulted in posterior collapse across all classes, where learned $\boldsymbol{\sigma}^2$ converged to near-zero and the model effectively behaved as a deterministic encoder with an uninformative posterior. To evaluate other uncertainty signals, we compare the joint-likelihood score (Eq. 39) against alternative uncertainty-based scores at $(\beta, \nu) = (1.0, 0.5)$: variance sum $S_{\text{var}} = \sum_d \sigma_d^2$, and entropy $S_{\text{ent}} = \frac{1}{2} \sum_d \log(2\pi e\, \sigma_d^2)$. The joint-likelihood score performs best (0.903 AUROC), followed by variance sum (0.867) and entropy (0.852). Across the full (β, ν) grid, cell-averaged AUROC correlates strongly with posterior entropy (r=0.896) and posterior variance sum (r=0.817), indicating that the posterior's dispersion parameters track the difficulty of the anomaly detection task, with the joint-likelihood score being the most effective way to combine them.

Comparative performance. To contextualize this result, we benchmark VJE's class-averaged AUROC of 90.3% on CIFAR–10 against representative one-class anomaly detection methods (Table 4). We group methods by supervision regime to distinguish approaches trained with generic self-supervised/one-class objectives from approaches that incorporate anomaly-specific signals (e.g., synthetic anomalies, shifted negatives, or feature adaptation).

Within the *generic self-supervised / one-class training* regime, VJE achieves the strongest performance, outperforming Rot+Trans (89.8%) and GOAD (88.2%). Notably, VJE is evaluated without anomaly-specific modifications, using only joint likelihood scoring under the learned model.

For completeness, we also report several state-of-the-art methods that achieve higher absolute AUROC by leveraging anomaly-specific configurations, including CSI (Tack et al., 2020), PANDA (Reiss et al., 2021), CutPaste (Li et al., 2021), and SimpleNet (Liu et al., 2023). These methods rely on synthetic anomalies, shifted negatives, or feature adaptation and therefore operate under different supervision assumptions than the generic regime, and are included to provide a broader reference point on the state-of-the-art, rather than as directly comparable baselines.

Table 4: Comparison of one-class CIFAR–10 anomaly detection methods. Entries report mean AUROC (%) over ten one-vs-rest splits. Methods are grouped by training regime: the top block contains generic self-supervised or one-class objectives that are directly comparable to VJE, while the bottom block contains state-of-the-art approaches that utilize anomaly-specific signals or pretrained embeddings (synthetic anomalies, shifted negatives, or feature adaptation) included for broader context. For VJE, scores are computed using the joint likelihood score. Asterisks (*) mark values reproduced from Tack et al. (2020) and Reiss and Hoshen (2023).

Method	Backbone	Learning approach	AUROC (%)
VJE (ours)	ResNet–18	one-class variational SSL (joint likelihood)	**90.3**
Rot+Trans (Tack et al., 2020)	ResNet–18	rotation+translation self-supervision	89.8*
GOAD (Bergman and Hoshen, 2020)	ResNet–18	multi-head transformation prediction	88.2*
GEOM (Golan and El-Yaniv, 2018)	WRN-16-8	geometric self-supervision	86.0
DROCC (Goyal et al., 2020)	LeNet	adversarial robust one-class classifier	71.6*
Deep SVDD (Ruff et al., 2018)	LeNet	hypersphere one-class objective	64.8
SimpleNet (Liu et al., 2023)	ResNet–18	feature mixing with synthetic noise	99.6
PANDA (Reiss et al., 2021)	ResNet–18	pre-trained feature adaptation	96.2
CutPaste (Li et al., 2021)	ResNet–18	cut-paste synthetic anomalies	95.2
CSI (Tack et al., 2020)	ResNet–18	contrastive learning with shifted views	94.3

5.4 Ablation studies

We perform further ablations on the VJE objective along two axes: tail heaviness in the Student–t likelihood, and the contribution of each loss component (KL regularizer, directional Student–t term, radial Student–t term).

Student–t degrees of freedom. To understand the role of tail heaviness in representation learning, we sweep the degrees of freedom $\nu \in \{1.0, 3.0, 7.0, 20.0, 50.0, \infty\}$ while holding β fixed at 1.0. We train for 200 epochs across CIFAR–10, CIFAR–100, and STL–10 (Coates et al., 2011), and evaluate both k–NN ($k = 30$) accuracy on the encoder output z and linear probe accuracy on frozen features.

Table 5 shows that, unlike the anomaly detection setting (Section 5.3) where very small values of ν are crucial and performance drops sharply as ν increases, representation learning is less sensitive to the exact choice of tail heaviness. Across finite values $\nu \in \{1, 3, 7, 20, 50\}$, both k–NN and linear accuracy stay within a narrow band on all three datasets, with smaller ν still slightly preferred (e.g., , ν=1 on CIFAR–10/100 and ν=3 on STL–10). In contrast, the Gaussian limit ($\nu \to \infty$) exhibits clear signs of failure on all datasets, with substantially lower k–NN and linear accuracy. This confirms that the bounded influence of the heavy-tailed

Table 5: Ablation of Student–t degrees of freedom ν across three datasets. All models use β=1.0. The Gaussian limit ($\nu \to \infty$) consistently underperforms across all datasets, with severe degradation on CIFAR–100. Results are averaged across three runs.

	CIFAR–10		**CIFAR–100**		**STL–10**	
ν	k–NN	Linear	k–NN	Linear	k–NN	Linear
1.0	86.18	89.23	54.17	63.05	78.71	84.23
3.0	86.2	88.9	53.7	62.7	79.7	85.1
7.0	86.02	88.71	53.26	62.08	79.4	84.58
20.0	85.82	88.5	53.42	62.29	79.35	84.93
50.0	85.4	88.08	51.93	61.53	78.91	84.68
∞ (Gaussian)	29.0	62.0	3.3	23.4	37.8	55.7

Student–t likelihood is essential for stable optimization even when the primary objective is discriminative rather than probabilistic.

Loss components. To assess the contribution of the KL regularizer $\mathcal{L}_{\mathrm{KL}}$, the directional Student–t term $\mathcal{L}_{\mathrm{dir}}$, and the radial Student–t term $\mathcal{L}_{\mathrm{rad}}$, we ablate these components on CIFAR–10, CIFAR–100, and STL–10. We utilize the same architecture and training setup as in the previous experiment, with ν=3.0 and β=1.0, and evaluate six objectives: the full loss $\mathcal{L}_{\mathrm{dir}} + \mathcal{L}_{\mathrm{rad}} + \mathcal{L}_{\mathrm{KL}}$, variants that drop exactly one term ($\mathcal{L}_{\mathrm{dir}} + \mathcal{L}_{\mathrm{rad}}$, $\mathcal{L}_{\mathrm{rad}} + \mathcal{L}_{\mathrm{KL}}$, $\mathcal{L}_{\mathrm{dir}} + \mathcal{L}_{\mathrm{KL}}$), and extreme cases retaining only $\mathcal{L}_{\mathrm{rad}}$ or $\mathcal{L}_{\mathrm{dir}}$. For each configuration, we report k–NN (k=30) and linear accuracy, alongside the mean and coefficient of variation (CV) of the posterior variance $\boldsymbol{\sigma}^2$ across latent dimensions. Tables 6 and 7 summarize the discriminative performance and posterior statistics.

Table 6: Ablation of loss components on CIFAR–10, CIFAR–100, and STL–10 with ν=3.0. We report k–NN accuracy (k=30) and linear probe accuracy (Top-1, %) on frozen encoder features.

	CIFAR–10		**CIFAR–100**		**STL–10**	
Objective	k–NN	Linear	k–NN	Linear	k–NN	Linear
$\mathcal{L}_{\mathrm{dir}} + \mathcal{L}_{\mathrm{rad}} + \mathcal{L}_{\mathrm{KL}}$	86.18	88.91	53.65	62.70	79.74	85.11
$\mathcal{L}_{\mathrm{dir}} + \mathcal{L}_{\mathrm{KL}}$	86.08	88.97	53.34	62.00	79.41	84.40
$\mathcal{L}_{\mathrm{dir}} + \mathcal{L}_{\mathrm{rad}}$	19.86	22.68	3.69	6.30	10.10	10.01
$\mathcal{L}_{\mathrm{rad}} + \mathcal{L}_{\mathrm{KL}}$	25.41	31.63	4.53	9.38	23.58	27.83
$\mathcal{L}_{\mathrm{rad}}$	18.19	19.13	4.58	9.67	18.66	23.48
$\mathcal{L}_{\mathrm{dir}}$	13.44	15.66	3.90	5.52	9.78	10.00

The ablations reveal three qualitatively distinct regimes consistent with the latent-variable analysis in Section 4. The first regime corresponds to the full objective $\mathcal{L}_{\mathrm{dir}} + \mathcal{L}_{\mathrm{rad}} + \mathcal{L}_{\mathrm{KL}}$ and the variant $\mathcal{L}_{\mathrm{dir}} + \mathcal{L}_{\mathrm{KL}}$. These are the only configurations that maintain strong discriminative performance and yield a nondegenerate posterior with moderate variance and clear anisotropy. For instance, on CIFAR–100, they attain linear accuracies of 62.70% and 62.00% (Table 6), with Var mean ≈ 0.33 and Var CV ≈ 0.74 (Table 7). This finite, anisotropic variance confirms that the model learns structured uncertainty while retaining discriminative capacity.

A second regime arises when $\mathcal{L}_{\mathrm{KL}}$ is absent ($\mathcal{L}_{\mathrm{dir}} + \mathcal{L}_{\mathrm{rad}}$ and $\mathcal{L}_{\mathrm{dir}}$). Without regularization, the optimizer can minimize the directional loss by collapsing the posterior variance to zero while aligning the mean direction with the target. This results in Var mean ≈ 0 and poor accuracy. In this collapsed regime, Var CV is undefined (and therefore omitted in Table 7) and does not indicate meaningful anisotropy, as the posterior is effectively deterministic.

Table 7: Posterior variance statistics for loss ablations. *Var mean* denotes the average posterior variance $\text{Mean}(\sigma_1^2, \ldots, \sigma_D^2)$, and *Var CV* is its coefficient of variation $\frac{\text{Std}(\sigma_1^2,\ldots,\sigma_D^2)}{\text{Mean}(\sigma_1^2,\ldots,\sigma_D^2)}$, which measures anisotropy across latent dimensions $d \in \{1, \ldots, D\}$. For configurations with nonzero Var mean, values of Var CV close to 0 imply isotropy, while higher values imply anisotropy. Omitted entries indicate cases where Var mean ≈ 0, in which CV is numerically unstable and not informative.

	CIFAR–10		**CIFAR–100**		**STL–10**	
Objective	Mean	CV	Mean	CV	Mean	CV
$\mathcal{L}_{\text{dir}} + \mathcal{L}_{\text{rad}} + \mathcal{L}_{\text{KL}}$	0.32	0.78	0.33	0.73	0.28	0.70
$\mathcal{L}_{\text{dir}} + \mathcal{L}_{\text{KL}}$	0.32	0.77	0.33	0.74	0.27	0.70
$\mathcal{L}_{\text{dir}} + \mathcal{L}_{\text{rad}}$	0.00	–	0.00	–	0.00	–
$\mathcal{L}_{\text{rad}} + \mathcal{L}_{\text{KL}}$	0.99	0.00	0.99	0.00	0.99	0.00
$\mathcal{L}_{\text{rad}}$	0.07	0.06	0.32	0.11	0.21	0.07
$\mathcal{L}_{\text{dir}}$	0.00	–	0.00	–	0.00	–

The third regime occurs when the directional likelihood is omitted. With $\mathcal{L}_{\text{rad}} + \mathcal{L}_{\text{KL}}$, retaining $\mathcal{L}_{\text{KL}}$ without $\mathcal{L}_{\text{dir}}$ allows the posterior to match the isotropic prior (Var mean ≈ 1, Var CV ≈ 0), leading to a sharp drop in accuracy as the model converges to it. With $\mathcal{L}_{\text{rad}}$ alone, the posterior remains nearly isotropic (small Var CV) as it shrinks to a roughly spherical shape with much smaller variance (Var mean ≈ 0.1–0.3), again discarding informative anisotropy and yielding poor discriminative capacity.

Taken together, these results demonstrate that a well-posed VJE objective minimally requires the directional Student–t term and the KL regularizer, as removing either leads to collapsed or prior-matching posteriors and a significant loss of discriminative power. Lastly, while the radial term $\mathcal{L}_{\text{rad}}$ only modestly affects linear accuracy in the full objective, it plays a geometric anchoring role by tying the radial scale of the posterior to the encoder features, which prevents arbitrary latent rescaling when combined with the directional likelihood.

6 Conclusion

In this work, we introduced **Variational Joint Embedding (VJE)**, a framework that presents a normalized probabilistic formulation of non-contrastive self-supervised learning. By using an amortized inference network and maximizing a symmetric conditional ELBO, VJE preserves the reconstruction-free training paradigm of joint embedding architectures while providing feature-wise uncertainty signals through an explicit variational posterior.

The theoretical robustness of the framework rests on defining a normalized likelihood on representation-space observations. We achieved this by factorizing the likelihood into decoupled directional and radial components, reparameterizing the radial term as a norm difference, and anchoring the geometry via analytic KL regularization. This construction explicitly decouples angular alignment from embedding norm, mitigating norm-induced inflation of the alignment signal and large, unstable updates that can arise when angle and scale are coupled under a single discrepancy, while ensuring that the learned geometry remains probabilistically coherent. Empirically, VJE maintains strong discriminative performance on standard benchmarks while producing non-degenerate posteriors that enable effective likelihood-based scoring, as demonstrated by its performance on anomaly detection against comparable baselines.

Furthermore, our theoretical analysis in Appendix B clarifies the relationship between this likelihood-based formulation and standard pointwise objectives used in non-contrastive learning. Our analysis formally distinguishes the normalized conditional modelling primitive of VJE from the pointwise compatibility primitive of JEPA-style objectives. We further demonstrate that some common discrepancy measures, such as squared-error and cosine-based losses, arise as objective-level limiting configurations of VJE under explicit restrictions.

Limitations and future directions. Although our experiments focus on augmented views in vision, the underlying formulation is agnostic to both modality and the mechanism by which paired views are generated, requiring only paired observations of a shared latent signal. While VJE is competitive across benchmarks, we observe a gap to the strongest deterministic baselines on ImageNet-1K under linear probing, which highlights a trade-off between the density-based semantics of VJE and the sharpness of point-estimate objectives optimized primarily for separability. Closing this gap while maintaining the normalized conditional objective remains an important open challenge.

A promising direction for future work is to extend this probabilistic formulation to hierarchical or patch-based architectures to enable fine-grained spatial uncertainty. Additionally, while we instantiated the likelihood using heavy-tailed Student–t factors, the framework can in principle admit other properly normalized kernels. Exploring alternative likelihood families could further refine the model's ability to capture complex data manifolds in non-visual or mixed modalities.

References

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

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

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

Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, 2019.

Liron Bergman and Yedid Hoshen. Classification-based anomaly detection for general data. In *International Conference on Learning Representations (ICLR)*, 2020.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006. ISBN 978-0387310732.

Nicola De Cao, Luca Falorsi, Thomas Kipf, and Jakub M. Tomczak. The power spherical distribution. In *International Conference on Learning Representations (ICLR)*, 2020.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9912–9924, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607. PMLR, 2020a.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, 2021.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. In *arXiv preprint arXiv:2003.04297*, 2020b. URL `https://arxiv.org/abs/2003.04297`.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4754–4765, 2018.

Adam Coates, Honglak Lee, and Andrew Y. Ng. STL-10: A dataset for developing unsupervised feature learning, deep learning, self-taught learning and semi-supervised learning algorithms. `https://cs.stanford.edu/~acoates/stl10/`, 2011.

Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning (ICML)*, pages 1184–1193. PMLR, 2018.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1183–1192. PMLR, 2017.

Florin C Ghesu, Erik Hedman, Bogdan Georgescu, Sasa Grbic, Andreas Maier, and Dorin Comaniciu. Quantifying and leveraging classification uncertainty for chest radiograph assessment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 676–684. Springer, 2019.

Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: Training ImageNet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. URL `https://arxiv.org/abs/1706.02677`.

Sachin Goyal, Aditi Raghunathan, Moksh Jain, Harsha Vardhan Simhadri, and Prateek Jain. DROCC: Deep robust one-class classification. In *International Conference on Machine Learning (ICML)*, 2020.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 21271–21284, 2020.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

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)*, pages 9729–9738, 2020.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017. URL `https://openreview.net/forum?id=Sy2fzU9gl`.

Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2nd edition, 2009.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Samuel Kotz and Saralees Nadarajah. *Multivariate t Distributions and Their Applications*. Cambridge University Press, Cambridge, 2004. ISBN 978-0521838606.

Alex Krizhevsky. Learning multiple layers of features from tiny images (CIFAR-10). `https://www.cs.toronto.edu/~kriz/cifar.html`, 2009.

Yann LeCun. A path towards autonomous machine intelligence (v0.9.2). *OpenReview Preprint*, 2022. Technical Report.

Chun-Liang Li, Kihyuk Sohn, Jaehoon Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Yibin Liu et al. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017. URL `https://openreview.net/forum?id=Skq89Scxx`.

Lars Maaløe, Casper Kaae Sønderby, Søren Kaae Sønderby, and Ole Winther. Biva: A very deep hierarchy of latent variables for generative modelling. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 6548–6558, 2019.

Kanti V. Mardia and Peter E. Jupp. *Directional Statistics*. John Wiley & Sons, Chichester, 2nd edition, 2000.

Hiroki Nakamura, Masashi Okada, and Tadahiro Taniguchi. Representation uncertainty in self-supervised learning as variational inference, 2023. URL `https://arxiv.org/abs/2203.11437`.

Tal Reiss and Yedid Hoshen. Mean-shifted contrastive loss for anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. PANDA: Adapting pretrained features for anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning (ICML)*, pages 1278–1286, 2014.

Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Lucas Deecke, Shoaib A. Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International Conference on Machine Learning (ICML)*, pages 4393–4402. PMLR, 2018.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54:30–44, 2019.

Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. CSI: Novelty detection via contrastive learning on distributionally shifted instances. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Duo Wang and Junzhou Huang. Anomaly detection with variational autoencoders for high-dimensional data. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 216–225. IEEE, 2018.

Mehmet Can Yavuz and Berrin Yanikoglu. Variational self-supervised learning, 2025. arXiv preprint arXiv:2504.04318.

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

David Zimmerer, Simon A. A. Kohl, Jens Petersen, Fabian Isensee, and Klaus H. Maier-Hein. Context-encoding variational autoencoder for unsupervised anomaly detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 282–290. Springer, 2018.

A Pseudocode

This appendix provides pseudocode for the VJE objective described in Sections 3 and 4. The routines implement the symmetric conditional ELBO of Eq. (37) using the directional and radial Student–t likelihoods together with a diagonal Gaussian variational posterior. All likelihood terms are evaluated with the target view detached (stop-gradient), and constants independent of $(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ are omitted.

Directional negative log-likelihood. This routine evaluates the extrinsic directional Student–t negative log-likelihood on unit-normalized embeddings, corresponding to Eq. (24).

Algorithm 1 Directional negative log-likelihood (extrinsic Student–t)

Require: Target embedding $\mathbf{z}$, sample $\mathbf{s}$, variance $\boldsymbol{\sigma}^2$, degrees of freedom ν, dimension D
1: $\hat{\mathbf{z}} \leftarrow \text{normalize}(\mathbf{z})$
2: $\hat{\mathbf{s}} \leftarrow \text{normalize}(\mathbf{s})$
3: $\text{q\_dir} \leftarrow \sum_k (\hat{z}_k - \hat{s}_k)^2 / \sigma_k^2$
4: **return** $\frac{1}{2}(\nu + D)\log\big(1 + \text{q\_dir}/\nu\big) + \frac{1}{2}\sum_k \log \sigma_k^2$

Radial negative log-likelihood. This routine evaluates the radial Student–t negative log-likelihood on the norm difference $\Delta r = \|\mathbf{z}\| - \|\mathbf{s}\|$ as introduced in Section 4.3.

Algorithm 2 Radial negative log-likelihood (Student–t on Δr)

Require: Target embedding $\mathbf{z}$, sample $\mathbf{s}$, degrees of freedom ν
1: $r_z \leftarrow \|\mathbf{z}\|$
2: $r_s \leftarrow \|\mathbf{s}\|$
3: $\Delta r \leftarrow r_z - r_s$
4: **return** $\frac{1}{2}(\nu + 1)\log\big(1 + (\Delta r)^2/\nu\big)$

KL divergence. This routine computes the analytic KL divergence between a diagonal Gaussian posterior and the standard Normal prior, as in Eq. (9).

Algorithm 3 KL divergence (diagonal Gaussian vs. standard Normal)

Require: Mean vector $\boldsymbol{\mu}$, variance vector $\boldsymbol{\sigma}^2$
1: **return** $\frac{1}{2}\sum_k \big(\sigma_k^2 + \mu_k^2 - 1 - \log \sigma_k^2\big)$

VJE training step. This routine performs one symmetric VJE update, combining the directional and radial likelihood terms with the KL regularizer as in Eq. (38).

Algorithm 4 One training step of VJE

Require: Encoder f_θ, inference network g_ϕ, views (x_1, x_2); weights β, ν
1: $\mathbf{z}_1 \leftarrow f_\theta(x_1)$; $\quad \mathbf{z}_2 \leftarrow f_\theta(x_2)$
2: $(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2) \leftarrow g_\phi(\mathbf{z}_1)$; $\quad (\boldsymbol{\mu}_2, \boldsymbol{\sigma}_2^2) \leftarrow g_\phi(\mathbf{z}_2)$
3: $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$; $\quad \mathbf{s}_i \leftarrow \boldsymbol{\mu}_i + \sqrt{\boldsymbol{\sigma}_i^2} \odot \boldsymbol{\epsilon}_i \quad$ for $i = 1, 2$
4: $\ell_{\text{dir}} \leftarrow \frac{1}{2}\big[\text{nll\_dir}(\mathbf{z}_2.\text{detach}(), \mathbf{s}_1, \boldsymbol{\sigma}_1^2, \nu, D)$
$\qquad + \text{nll\_dir}(\mathbf{z}_1.\text{detach}(), \mathbf{s}_2, \boldsymbol{\sigma}_2^2, \nu, D)\big]$
5: $\ell_{\text{rad}} \leftarrow \frac{1}{2}\big[\text{nll\_rad}(\mathbf{z}_2.\text{detach}(), \mathbf{s}_1, \nu)$
$\qquad + \text{nll\_rad}(\mathbf{z}_1.\text{detach}(), \mathbf{s}_2, \nu)\big]$
6: $\ell_{\text{KL}} \leftarrow \frac{1}{2}\big[\text{kld}(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2) \; + \; \text{kld}(\boldsymbol{\mu}_2, \boldsymbol{\sigma}_2^2)\big]$
7: **return** $\ell_{\text{dir}} + \ell_{\text{rad}} + \beta\, \ell_{\text{KL}}$

Additional notes.

- **ELBO structure.** Likelihood terms use $\mathbf{z}_j$.detach(), implementing the conditional ELBO of Section 4.4 and Eq. (37), where each view predicts the latent representation of the opposite view under a fixed observation.
- **Variance parameterization.** The same diagonal variance $\boldsymbol{\sigma}^2$ parameterizes both the Gaussian posterior and the directional Student–t likelihood, as described in Section 4.2. In practice, g_ϕ may output an unconstrained vector that is passed through a `softplus` nonlinearity (and a small floor) to obtain $\boldsymbol{\sigma}^2$; for clarity, the pseudocode assumes g_ϕ directly returns $\boldsymbol{\sigma}^2$.
- **Stability.** In practice, a small constant (e.g., 10^{-6}) is used inside `normalize` (via $\hat{\mathbf{z}} = \mathbf{z}/\max(\|\mathbf{z}\|, \varepsilon)$) and as a floor on $\boldsymbol{\sigma}^2$ to avoid division by zero and degenerate variances; we omit this from the interface for clarity.
- **Radial scale.** The radial likelihood uses $\Delta r = \|\mathbf{z}\| - \|\mathbf{s}\|$ with the scale fixed to $\lambda = 1$, as described in Section 4.3.
- **Monte Carlo sampling.** Each expectation under $q(\mathbf{s} \mid \mathbf{z})$ in Eq. (38) is approximated with a single reparameterized sample per view.

B Energy-based Predictive Learning and Normalized Probabilistic Modelling

B.1 Overview and scope

Predictive joint embedding in non-contrastive self-supervised learning admits two distinct formulations. The first is an *energy-based* (or compatibility) formulation LeCun (2022), in which training minimizes a pointwise discrepancy between a prediction and a target embedding, without specifying an explicit normalized density over embeddings. The second is the *normalized probabilistic* formulation that we adopt in Variational Joint Embedding (VJE), in which a normalized likelihood is specified in representation space and optimized through a symmetric conditional ELBO (Section 4). While these paradigms may share architectural motifs (e.g., asymmetric branches, auxiliary heads, and momentum/target encoders (Grill et al., 2020; He et al., 2020; Chen et al., 2020b)), they optimize different primitives and therefore assign different semantic roles to their components.

In this appendix, we make an objective-level comparison between the primitive used by energy-based (deterministic) non-contrastive objectives and our VJE formulation. Architectural patterns vary across methods, and we do not conflate them with the energy-based or likelihood-based formulation itself, except where the normalized likelihood formulation assigns a specific semantic role to the inference network and the target branch.

Given two related views x_1, x_2 of an input x, let $\mathbf{z}_i = f_\theta(x_i) \in \mathbb{R}^D$ denote the encoder embedding for view $i \in \{1, 2\}$. In VJE, we perform a directed scoring from a source embedding $\mathbf{z}_i$ to a target embedding $\mathbf{z}_j$, averaged symmetrically over $(i, j) \in \{(1, 2), (2, 1)\}$. The inference network g_ϕ outputs parameters $(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2)$ of a diagonal Gaussian variational posterior (Eq. (32)),

$$q_i(\mathbf{s} \mid \mathbf{z}_i) = \mathcal{N}(\boldsymbol{\mu}_i, \mathrm{diag}(\boldsymbol{\sigma}_i^2)),$$

from which a latent sample $\mathbf{s}_i$ is drawn. The representation-space observation scored by the likelihood is:

$$y_j = (\hat{\mathbf{z}}_j, \|\mathbf{z}_j\|), \qquad \hat{\mathbf{z}}_j = \mathbf{z}_j / \|\mathbf{z}_j\|,$$

and the radial residual entering the likelihood is:

$$\Delta r_{ij} = \|\mathbf{z}_j\| - \|\mathbf{s}_i\|,$$

as used in Eq. (31).

B.2 Energy scores and normalized likelihoods

A typical joint-embedding objective takes the form of a pointwise discrepancy between a prediction and a target embedding:

$$\mathcal{L}_{\mathrm{JEPA}}(x_i, x_j) = d\big(g_\phi(\mathbf{z}_i),\, \mathbf{z}_j\big). \tag{40}$$

Here $d(\cdot, \cdot)$ is commonly a squared-error or cosine-style discrepancy. Optimization proceeds directly on this pointwise score, shaping embedding geometry by enforcing low discrepancy for compatible pairs. While such objectives can be related formally to energies and unnormalized conditional models, they are typically optimized as pointwise compatibility losses on embedding pairs rather than as likelihoods.

To develop VJE, we instead take a normalized likelihood as the primitive. We specify $p_\psi(y_j \mid \mathbf{s}_i, \boldsymbol{\sigma}_i^2)$ on the representation-space observation y_j, and we optimize this likelihood through a symmetric conditional ELBO by marginalizing latent codes to obtain the objective:

$$\overline{\mathcal{F}}^{(\beta)} = \frac{1}{2} \sum_{(i,j) \in \{(1,2),(2,1)\}} \left\{ \mathbb{E}_{\mathbf{s}_i \sim q_i} \left[\log p_\psi(y_j \mid \mathbf{s}_i, \boldsymbol{\sigma}_i^2) \right] - \beta\, \mathcal{L}_{\mathrm{KL}}(q_i \| p) \right\}, \tag{41}$$

where training minimizes $-\overline{\mathcal{F}}^{(\beta)}$ (Eq. (38)). In our formulation, the likelihood is normalized by construction on the observation space $y = (\hat{\mathbf{z}}, \|\mathbf{z}\|)$ (Section 4), so the negative log-likelihood terms have log-density semantics without introducing a partition function over embeddings.

B.3 Representation semantics: inference and fixed observations

A likelihood-based objective assigns specific semantic roles to components that otherwise appear as training heuristics in energy-based objectives.

Stop-gradient and fixed-observation semantics. In VJE, the likelihood term evaluates the log-density of a target-side observation $y_j = (\hat{\mathbf{z}}_j, \|\mathbf{z}_j\|)$, while $\mathbf{z}_j$ is produced by a trainable encoder. In probabilistic modelling, observations are treated as fixed while model parameters are optimized to explain them. In representation learning, y_j changes across training as the encoder evolves, and we enforce the intended semantics by treating y_j as fixed within each update step for the likelihood term, implemented by stop-gradient on the target branch (or alternatively by using a slowly updated target encoder).

To make this explicit, we differentiate the one-directional negative ELBO likelihood term and omit the KL term for clarity:

$$\mathcal{L}^{\mathrm{NLL}}_{i\to j}(\theta,\phi) = \mathbb{E}_{\mathbf{s}_i\sim q_i(\cdot\mid\mathbf{z}_i(\theta))}\left[-\log p_\psi\big(y_j(\theta)\mid \mathbf{s}_i, \boldsymbol{\sigma}_i^2\big)\right], \qquad y_j(\theta) = (\hat{\mathbf{z}}_j(\theta), \|\mathbf{z}_j(\theta)\|).$$

Taking gradients with respect to θ yields two pathways:

$$\nabla_\theta \mathcal{L}^{\mathrm{NLL}}_{i\to j} = \mathbb{E}_{\mathbf{s}_i\sim q_i}\left[\nabla_{\mathbf{s}_i}\Big(-\log p_\psi\big(y_j\mid \mathbf{s}_i,\boldsymbol{\sigma}_i^2\big)\Big)\nabla_\theta \mathbf{s}_i\right] + \mathbb{E}_{\mathbf{s}_i\sim q_i}\left[\nabla_{y_j}\Big(-\log p_\psi\big(y_j\mid \mathbf{s}_i,\boldsymbol{\sigma}_i^2\big)\Big)\nabla_\theta y_j\right]. \tag{42}$$

Our intended semantics correspond to retaining only the first pathway, in which model parameters are updated to explain a fixed observation under the conditional likelihood. Detaching the target branch in Eq. (38) sets $\nabla_\theta y_j = \mathbf{0}$ for the likelihood term. Allowing gradients to flow through both pathways changes the meaning of the objective by introducing a route in which the scored observation is modified by the same likelihood term.

Amortized inference network. Optimizing an ELBO requires an explicit variational posterior $q(\mathbf{s} \mid \mathbf{z})$ (Kingma and Welling, 2014; Rezende et al., 2014). In VJE, the inference network g_ϕ parameterizes this posterior by producing $(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2)$, enabling reparameterized sampling and efficient optimization of the ELBO. Deterministic objectives in energy-based joint embedding can be interpreted as learning point estimates, while VJE maintains an explicit distributional posterior that supports marginalization and density-based scoring.

Objective and architecture. The comparison above is objective-level, contrasting a pointwise energy score with a normalized likelihood. Many non-contrastive baselines apply their loss in a projected space. We omit a separate projection head since the likelihood semantics in VJE are attached directly to the geometry of the encoder embedding space, and both the posterior $q_i(\mathbf{s} \mid \mathbf{z}_i)$ and the likelihood $p_\psi(y_j \mid \mathbf{s}_i, \boldsymbol{\sigma}_i^2)$ are defined in relation to $\mathbf{z}$ (Section 3).

B.4 Unifying specific objectives as boundary configurations

With the primitives separated, we show two objective-level correspondences that recover common pointwise losses as boundary configurations under explicit likelihood choices and limits. These correspondences concern objective functions rather than full training pipelines, and they are not intended to claim complete equivalence of optimization dynamics.

Energy-to-likelihood mapping and partition functions. Given an energy $E(\mathbf{z}_j; \mathbf{z}_i)$, one may define an unnormalized conditional model $\tilde{p}(\mathbf{z}_j \mid \mathbf{z}_i) \propto \exp(-E(\mathbf{z}_j; \mathbf{z}_i))$. The corresponding normalized conditional density is:

$$p(\mathbf{z}_j \mid \mathbf{z}_i) = \frac{\exp(-E(\mathbf{z}_j;\mathbf{z}_i))}{Z(\mathbf{z}_i)}, \qquad Z(\mathbf{z}_i) = \int \exp(-E(\mathbf{z};\mathbf{z}_i))\, d\mathbf{z}. \tag{43}$$

Its negative log-likelihood decomposes as:

$$-\log p(\mathbf{z}_j \mid \mathbf{z}_i) = E(\mathbf{z}_j;\mathbf{z}_i) + \log Z(\mathbf{z}_i). \tag{44}$$

This identity explains why pointwise energy minimization does not generally coincide with likelihood maximization unless $\log Z(\mathbf{z}_i)$ is handled, for example by exact computation or sampling-based approximation, consistent with the energy-based predictive viewpoint of LeCun (2022).

Recovering squared-error objectives. A squared-error JEPA objective arises from our formulation when we replace the representation-space likelihood by an isotropic Gaussian density on embeddings, $p^{\mathcal{N}}(\mathbf{z}_j \mid \mathbf{s}_i) = \mathcal{N}(\mathbf{z}_j; \mathbf{s}_i, \lambda I)$, keep $\lambda > 0$ fixed, score the observation $\mathbf{z}_j$ rather than $y_j = (\hat{\mathbf{z}}_j, \|\mathbf{z}_j\|)$, take the point-estimate posterior limit $\boldsymbol{\sigma}_i^2 \to \mathbf{0}$ so that $\mathbf{s}_i \to \boldsymbol{\mu}_i$, and remove KL regularization by setting $\beta \to 0$. Under these limits and up to additive constants, the one-directional objective reduces to:

$$\mathcal{L}_{i \to j} \;\to\; \frac{1}{2\lambda} \|\mathbf{z}_j - \boldsymbol{\mu}_i\|^2, \qquad \boldsymbol{\mu}_i = g_\phi(\mathbf{z}_i),$$

which is a squared-error compatibility loss. Averaging symmetrically over both directions recovers the standard squared-error JEPA objective, and the same reduction applies when $\mathbf{z}_j$ represents a collection of patch embeddings, yielding the corresponding I-JEPA loss (Assran et al., 2023). The fixed-scale restriction is essential, since removing the KL term while retaining a learned per-sample scale no longer coincides with a deterministic squared-error energy, a distinction consistent with the variance degeneracy observed in the "without KL" ablation in Section 5.4.

Recovering directional objectives. A cosine-style objective arises from the directional channel when we restrict to the directional NLL in Eq. (24), fix the scale to be isotropic ($\boldsymbol{\sigma}_i^2 \equiv \mathbf{1}$), take the Gaussian limit $\nu \to \infty$ (Eq. (13)), take the point-estimate posterior limit, and remove the KL and radial terms. Under these limits and up to constants, we obtain:

$$\ell_{\text{dir}} \;\to\; \frac{1}{2} \|\hat{\boldsymbol{\mu}}_i - \hat{\mathbf{z}}_j\|^2 = 1 - \hat{\boldsymbol{\mu}}_i^\top \hat{\mathbf{z}}_j,$$

which is equivalent to cosine alignment. Summed symmetrically over both directions, this reproduces a SimSiam-style objective at the level of the loss (Chen and He, 2021), while SimSiam's empirical behavior additionally depends on its architectural asymmetry and optimization dynamics.

B.5 Geometric regularization

The formulations also differ in how embedding geometry is regulated. Energy-based objectives are commonly paired with architectural stabilization, such as stop-gradient or momentum targets, and may additionally include explicit penalties on batch statistics. VICReg and Barlow Twins, for example, enforce representation diversity through variance or covariance constraints computed over batches (Bardes et al., 2022; Zbontar et al., 2021). These mechanisms constrain aggregate embedding geometry but do not specify a conditional likelihood on representation-space observations.

In VJE, geometric regularization follows from the probabilistic structure. The analytic KL term in Eq. (38) anchors each instance-conditioned posterior to the explicit isotropic Gaussian prior $p(\mathbf{s}) = \mathcal{N}(\mathbf{0}, I)$ (Eq. (34)). Moreover, we share the posterior scale $\boldsymbol{\sigma}_i^2$ with the directional likelihood scale (Section 4.2), so the same feature-wise parameters govern both sampling in $q_i(\mathbf{s})$ and normalization or weighting in the directional likelihood. This yields a prior-relative notion of geometry within a normalized likelihood model and underlies likelihood-based scoring such as the anomaly detection experiments in Section 5.3.

LeJEPA (Balestriero and LeCun, 2025) provides a complementary perspective by advocating an isotropic Gaussian target geometry for embeddings and introducing SIGReg as an explicit global regularizer that encourages this structure. The relationship to VJE is geometric rather than a strict reduction, since LeJEPA enforces isotropy through a deterministic regularizer on encoder outputs, whereas VJE encourages prior-relative structure through an instance-conditioned posterior and analytic KL within a normalized likelihood model.