# EMBRACE AMBIGUITY: RETHINKING CONTRASTIVE LEARNING THROUGH SIMILARITY UNCERTAINTY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

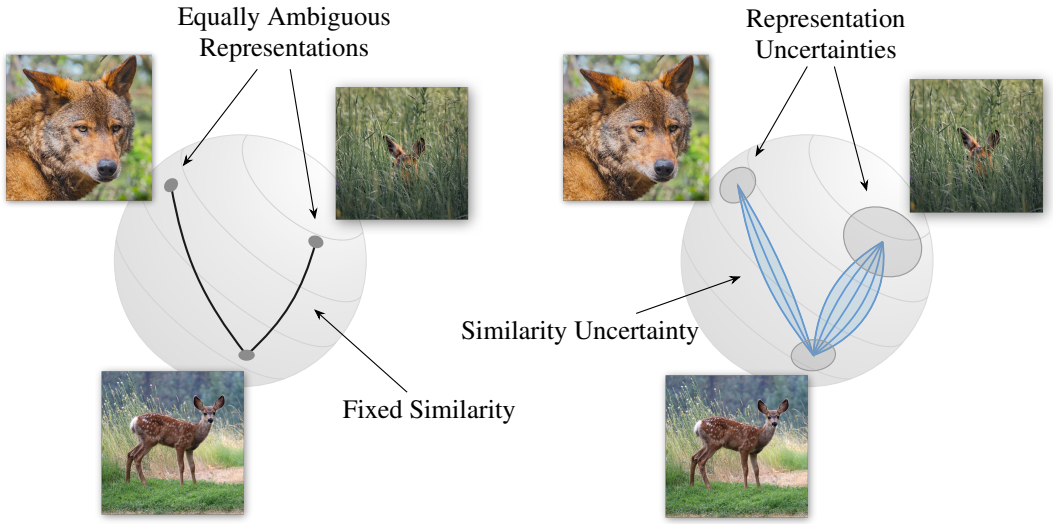
Contrastive learning has become a dominant paradigm for self-supervised representation learning, but two key limitations persist: (1) embedding similarities are treated as deterministic, causing ambiguous pairs to be optimized with unwarranted confidence and potentially distorting fine-grained semantic structure; and (2) prior probabilistic approaches model uncertainty in individual representations or learn temperatures heuristically, leaving the uncertainty of the pairwise similarity itself, which directly governs contrastive objectives, unmodeled. To address these limitations, we propose Distributional Contrastive Learning (DistCLR), a self-supervised framework that models both representations and pairwise similarities distributionally, inducing similarity uncertainty in closed form from the underlying representation posteriors. Motivated by an evidence lower bound (ELBO), DistCLR formulates an uncertainty-aware InfoNCE objective in which the global temperature is replaced by this per-pair similarity uncertainty, softening ambiguous pairs while preserving sharp alignment for confident ones. DistCLR changes only the pairwise contrastive score: it requires no Monte Carlo sampling, adds less than 1% wall-clock overhead, recovers InfoNCE under homoscedastic uncertainty, and plugs into existing InfoNCE-based frameworks without altering their training pipelines. Experiments on CIFAR-10, CIFAR-100, ImageNet-100, and ImageNet-1k show that DistCLR (1) matches or improves classification performance across three contrastive frameworks and two architectures, (2) induces clearer semantic organization in the representation space, as evidenced by neighborhood consistency, effective rank, clustering, and cross-domain transfer assessments, (3) improves strict open-set recognition, robustness under corrupted data, and misclassification detection, and (4) outperforms representative uncertainty-aware contrastive methods in both linear-probe accuracy and error detection. Code is provided in the supplementary material.

## 1 INTRODUCTION

Self-supervised learning (SSL) aims to imitate human learning by acquiring knowledge from observation alone, achieving remarkable results (Gui et al., 2024). Contrastive learning has emerged as a dominant SSL paradigm, structuring latent spaces across vision, language, and the natural sciences (Radford et al., 2021; Gao et al., 2021; Yu et al., 2023). At its core, it treats instance similarity as deterministic, rigidly pulling augmented views together and pushing all others apart. Yet human cognition treats "similar" as a *probable* association, a nuance standard contrastive objectives fail to capture. Prior work (Oh et al., 2019; Kirchhof et al., 2023; Park et al., 2022) models uncertainty in individual representations, but similarities themselves remain deterministic once embeddings are produced. This yields two limitations: (1) a sample's relations to others may remain uncertain even when its embedding is well estimated, causing ambiguous pairs to be optimized with unwarranted confidence; and (2) pointwise uncertainty cannot capture the relational ambiguity governing contrastive objectives, which depend on pairwise reliability instead of isolated embeddings. Existing methods can therefore over-commit to noisy alignments and degrade fine-grained structure. This gap motivates our work to establish a principled foundation for self-supervised contrastive learning through modeling *both* instance representations and their pairwise similarities as latent distributions (Figure 1).



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




(a) Deterministic Contrastive Learning (b) Distributional Contrastive Learning

Figure 1: Standard vs. proposed contrastive learning. **(a)** Deterministic methods embed instances as points with fixed scalar similarities and rigid separation. **(b)** DistCLR (ours) models *both* representations and pairwise similarities as latent distributions; the derived uncertainties modulate the objective, tightening alignment for confident pairs and relaxing it for ambiguous ones.

Central to contrastive learning, InfoNCE (Oord et al., 2018) aligns positive pairs through deterministic point comparisons. However, it relies on a fixed temperature $\tau$ that indiscriminately enforces hyperspherical alignment and uniformity (Wang & Isola, 2020), entirely ignoring per-pair ambiguity. Equally forcing the alignment of ambiguous and genuine pairs drives the model toward ambiguity invariance, erasing instance-specific variations (Xiao et al., 2021). Preventing this collapse calls for a mechanism that sets alignment strength per pair instead of applying a global heuristic. Probabilistic latent-variable models (Kingma & Welling, 2013; Suzuki & Matsuo, 2022) provide one: marginalizing the contrastive probability over representation uncertainty softens each comparison in proportion to how uncertain it is.

In this work, we propose Distributional Contrastive Learning (DistCLR), a self-supervised approach that retains the proven geometric inductive biases of contrastive learning while incorporating analytically propagated similarity uncertainty, yielding three key advantages. First, DistCLR maps *both* instances and their relation to parameterized latent distributions, enabling probabilistic affinity estimation in place of deterministic point comparisons. Second, it generalizes the InfoNCE objective such that the heuristic temperature is replaced by a pair-specific, heteroscedastic uncertainty metric, termed *similarity uncertainty*, that naturally applies sharp gradient updates to concordant views while attenuating the penalty for ambiguous pairs, preventing forced disruption of the underlying representation manifold. Third, DistCLR combines geometry-aware contrastive learning with learned uncertainty estimation, improving semantic organization while providing informative uncertainty scores.

Our main contributions are summarized as follows:

1. We connect variational inference with contrastive learning through a distributional formulation of pairwise similarity, yielding an uncertainty-aware generalization of InfoNCE.

2. We introduce *similarity uncertainty*, a pair-specific heteroscedastic temperature with a closed-form expression combining a shared augmentation baseline and an analytically propagated encoder component, replacing the global temperature $\tau$ and adapting contrastive scores to per-view uncertainty and pairwise geometry.

3. Across three representative contrastive frameworks (SimCLR, NNCLR, All4One), DistCLR matches or improves classification performance while providing uncertainty scores that rank misclassifications and unknown inputs, with additional benefits demonstrated in transfer and robustness evaluations.

4. Our learned representations exhibit richer semantic organization, yielding consistent improvements in neighborhood consistency, effective rank, and clustering, and greater robustness under distribution shifts.

## 2 PRELIMINARIES

**Self-Supervised Contrastive Learning.** Early formulations such as the pairwise contrastive loss (Chopra et al., 2005) and N-pair loss (Sohn, 2016) laid the foundation for current self-supervised contrastive representation learning. In this setting, an input sample $\mathbf{x}$ is encoded into a feature vector $\mathbf{z} \in \mathbb{R}^L$, typically $\ell_2$-normalized, and learning proceeds by bringing semantically related views closer while pushing unrelated views apart. This direction later crystallized into the noise-contrastive objective, InfoNCE (Oord et al., 2018):

$$\mathcal{L}_{\text{InfoNCE}_{i,j}} = -\log \frac{\exp(\mathbf{z}_i^\top \mathbf{z}_j^+ / \tau)}{\sum_{k=1}^{N} \mathbf{1}_{[k \neq i]} \exp(\mathbf{z}_i^\top \mathbf{z}_k / \tau)},$$

where $\mathbf{z}_i$ and $\mathbf{z}_j^+$ denote a positive pair and $\tau$ is a temperature parameter. InfoNCE underlies self-supervised systems such as SimCLR (Chen et al., 2020a). NNCLR (Dwibedi et al., 2021) and All4One (Estepa et al., 2023) use a support set to find positive pairs to capture meaningful semantic distributions. UniCon (Sui et al., 2026) has aided contrastive alignment via exact kernelized updates. However, these approaches remain purely deterministic. Our approach aims at simultaneously preserving contrastive geometry and improving generalization through capturing inherent ambiguity of the data.

**Uncertainty-aware Contrastive Learning.** Probabilistic formulations of contrastive learning emerged to capture ambiguity that deterministic embeddings cannot express. Hedged Instance Embedding (Oh et al., 2019) and its video extension (Park et al., 2022) model embeddings as distributions and derive uncertainty-aware contrastive losses, though with temperatures that are either fixed or learned through auxiliary networks. Following the alignment–uniformity perspective (Wang & Isola, 2020), TaU learns instance-specific temperatures (Zhang et al., 2021), while AMCL learns pair-adaptive temperatures across multiple projection heads (Wang et al., 2024). MCInfoNCE marginalizes contrastive probabilities over probabilistic embeddings (Kirchhof et al., 2023). Complementary regularization strategies followed: variance–covariance penalties such as VICReg (Bardes et al., 2021) prevent representation collapse through batch-level statistics, AnInfoNCE (Rusak et al., 2025) addresses identifiability failures under anisotropic augmentations, albeit at the cost of downstream accuracy, and VarCon (Wang et al., 2025) utilizes uncertainty for soft labeling under strict supervision. Our approach connects variational inference with self-supervised contrastive learning through analytically propagated similarity uncertainty.

**Variational Inference.** Variational Autoencoders (VAEs) (Kingma & Welling, 2013; Rezende et al., 2014) established the foundation for this paradigm, utilizing amortized encoder-decoder architectures to learn continuous latent representations. Bayesian weight uncertainty (Blundell et al., 2015) and learned priors have long served to enrich posterior flexibility and calibration, with subsequent work extending latent geometries to von Mises-Fisher distributions (Davidson et al., 2018) to better reflect the unit-norm structure inherent to contrastive embeddings. Building on these foundations, deep multimodal generative models (Ivanovic et al., 2020; Suzuki & Matsuo, 2022) established a principled framework for learning structured latent spaces through evidence lower bound maximization. In self-supervised settings, the integration of probabilistic objectives with contrastive losses has gained significant traction: the Noise-Contrastive Prior tightens the posterior-prior gap (Aneja et al., 2021), while information-bottleneck formulations augment instance-discrimination with explicit compression terms (Wang et al., 2022). Our method builds upon these foundations by modeling *both* representations and their pairwise similarities as distributions, preserving contrastive learning's geometric properties while incorporating explicit probabilistic semantics.

## 3 METHOD

Consider $M$ views $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^M$, with $\mathbf{x}_i = a_i(\mathbf{x})$ for augmentations $a_i \sim p_A$ applied to a shared input with $L$-dimensional latent $\mathbf{z} \in \mathbb{R}^L$, which are drawn from the induced distribution

$p_{\mathcal{D}}(\mathcal{X})$. For each view $\mathbf{x}_i$, an encoder parameterized by $\phi$ produces a Gaussian posterior expert $q_\phi(\mathbf{z}_i|\mathbf{x}_i) = \mathcal{N}(\boldsymbol{\mu}_i, \sigma_i^2\,\mathbf{I})$, and we adopt an isotropic Gaussian prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \sigma_p^2\,\mathbf{I})$ with $\sigma_p > 0$. Following multimodal generative modeling (Wu & Goodman, 2018), we combine the view-conditioned posterior experts through a view-only product-of-experts rule:

$$q_\phi(\mathbf{z}|\mathcal{X}) = \mathcal{N}(\boldsymbol{\mu}_{\mathcal{X}}, \sigma_{\mathcal{X}}^2\,\mathbf{I}), \qquad \sigma_{\mathcal{X}}^{-2} = \textstyle\sum_{i=1}^{M} \sigma_i^{-2}, \qquad \boldsymbol{\mu}_{\mathcal{X}} = \sigma_{\mathcal{X}}^2 \textstyle\sum_{i=1}^{M} \sigma_i^{-2}\boldsymbol{\mu}_i. \tag{1}$$

The Gaussian posterior parameterizes uncertainty in the Euclidean latent space, while the contrastive score is defined over the induced direction $\hat{\mathbf{z}}_i := \boldsymbol{\mu}_i/\|\boldsymbol{\mu}_i\|$ on the unit hypersphere $S^{L-1}$. Encoder uncertainty enters the contrastive objective analytically through the per-pair temperature (§3.2) and the scale regularizer (§3.3), so similarities are evaluated at the projected posterior means instead of reparameterized samples. This avoids Monte Carlo gradient variance while propagating encoder uncertainty analytically through the local cosine geometry. We motivate our contrastive surrogate using the evidence lower bound (ELBO) of a latent-variable model:

$$\log p_\theta(\mathcal{X}) \;\geq\; \mathcal{J}_{\text{ELBO}} \;=\; \mathbb{E}_{q_\phi(\mathbf{z}|\mathcal{X})}[\log p_\theta(\mathcal{X}|\mathbf{z})] - \underbrace{D_{\text{KL}}(q_\phi(\mathbf{z}|\mathcal{X})\,\|\,p(\mathbf{z}))}_{=:\,\mathcal{K}_{\text{ELBO}}}, \tag{2}$$

where $D_{\text{KL}}$ and $\mathcal{J}_{\text{ELBO}}$ denote the Kullback-Leibler divergence and ELBO objective, respectively.

We use this decomposition to motivate a per-pair uncertainty-aware InfoNCE (Section 3.1), derive a closed-form temperature splitting augmentation and encoder uncertainty (Section 3.2), and present the full prior-regularized objective (Section 3.3).

### 3.1 Uncertainty-aware InfoNCE

To decompose the reconstruction term in Eq. (2), we take the logarithm of Bayes' theorem, yielding $\log p_\theta(\mathcal{X}|\mathbf{z}) = \log p_\theta(\mathbf{z}|\mathcal{X}) + \log p_\theta(\mathcal{X}) - \log p_\theta(\mathbf{z})$ and substitute the aggregated-posterior approximation $p_\theta(\mathbf{z}) \approx \mathbb{E}_{p_{\mathcal{D}}}[p_\theta(\mathbf{z}|\mathcal{X})] =: p_{\mathcal{D}_\theta}(\mathbf{z})$ (Tomczak & Welling, 2018). Evaluating the expectation at the posterior mean, the negative ELBO takes an alignment–uniformity form (full derivation in Appendix C.1):

$$-\mathcal{J}_{\text{ELBO}} \;\approx\; \underbrace{-\log p_\theta(\mathbf{z}|\mathcal{X})}_{\text{alignment}} + \underbrace{\log p_{\mathcal{D}_\theta}(\mathbf{z})}_{\text{uniformity}} + \mathcal{K}_{\text{ELBO}} - \log p_\theta(\mathcal{X}). \tag{3}$$

We first construct the contrastive objective from the alignment and uniformity terms in Eq. (3). The prior regularizer and omission of the data-evidence term are addressed in Section 3.3.

**Definition 1** (Pairwise Directional Score)**.** *For each ordered positive pair $(r, t)$, we represent alignment by the negative pairwise score and uniformity by the log-average of exponentiated candidate scores. We construct these scores from the vMF log-density, with a concentration determined separately for each pair.*

For a fixed mean direction $\boldsymbol{\mu}$ and concentration $\kappa$, the von Mises–Fisher (vMF) density on $S^{L-1}$ is

$$p_{\text{vMF}}(\mathbf{x};\,\boldsymbol{\mu},\kappa) \;=\; C_L(\kappa)\,\exp\big(\kappa\,\boldsymbol{\mu}^\top\mathbf{x}\big), \qquad C_L(\kappa) = \frac{\kappa^{L/2-1}}{(2\pi)^{L/2}\,I_{L/2-1}(\kappa)}, \tag{4}$$

where $I_\nu(\cdot)$ is the modified Bessel function of the first kind. Defining the pairwise temperature $\upsilon_{r,t} := 1/\kappa_{r,t}$, setting $\boldsymbol{\mu} = \hat{\mathbf{z}}_r$, $\mathbf{x} = \hat{\mathbf{z}}_t$, and $\kappa = 1/\upsilon_{r,t}$ in Eq. (4) and taking logarithms yields:

$$\ell_{r,t} \;:=\; \frac{\hat{\mathbf{z}}_r^\top\hat{\mathbf{z}}_t}{\upsilon_{r,t}} \;+\; \underbrace{\log C_L\big(\tfrac{1}{\upsilon_{r,t}}\big)}_{=:\,b_{r,t}}, \tag{5}$$

where each pair carries its own temperature $\upsilon_{r,t}$, derived in Section 3.2. Let $\mathcal{B} = \{t\} \cup \mathcal{N}$ contain the positive and negative candidates. Using Eq. (5), the pairwise counterparts of the two terms in Eq. (3) combine as

$$\underbrace{-\ell_{r,t}}_{\text{alignment}} \underbrace{+\log\left(\frac{1}{|\mathcal{B}|}\sum_{j\in\mathcal{B}}\exp(\ell_{r,j})\right)}_{\text{uniformity}} = -\log\frac{\exp(\ell_{r,t})}{\sum_{j\in\mathcal{B}}\exp(\ell_{r,j})} - \log|\mathcal{B}|. \tag{6}$$

Dropping the parameter-independent constant $-\log|\mathcal{B}|$ and expanding the scores from Eq. (5) gives uncertainty-aware InfoNCE ($\mathcal{U}$-InfoNCE). Writing $\hat{\mathbf{z}}^+ = \hat{\mathbf{z}}_t$ and using superscripts $+$ and $-$ for positive- and negative-pair temperatures and corrections, respectively:

$$\mathcal{L}^{(r,t)}_{\mathcal{U}\text{-InfoNCE}} \;=\; -\log \frac{\exp\!\left(\dfrac{\hat{\mathbf{z}}_r^\top \hat{\mathbf{z}}^+}{\upsilon^+} + b^+\right)}{\exp\!\left(\dfrac{\hat{\mathbf{z}}_r^\top \hat{\mathbf{z}}^+}{\upsilon^+} + b^+\right) + \displaystyle\sum_{j=1}^{|\mathcal{N}|} \exp\!\left(\frac{\hat{\mathbf{z}}_r^\top \hat{\mathbf{z}}_j^-}{\upsilon_j^-} + b_j^-\right)}. \tag{7}$$

Equation (7) is the negative log-probability of the positive among the candidates (Appendix C.2). The final $\mathcal{L}_{\mathcal{U}\text{-InfoNCE}}$ averages it over all ordered pairs $(r,t)$ of distinct views in $\mathcal{X}$. Including the positive in the denominator retains the $O(1/|\mathcal{N}|)$ self-normalization bias under the assumptions in Appendix C.3 (Poole et al., 2019).

The log-normalizers $b^+, b_j^-$ provide an intrinsic pair-dependent regularization without additional hyperparameters: for any imperfectly aligned pair with similarity $s<1$, $\kappa s + \log C_L(\kappa) \to -\infty$ as $\kappa \to \infty$, penalizing overconfident temperatures whenever the observed alignment does not support them. In contrast to heuristic temperature clipping (Kirchhof et al., 2023), this score correction comes from the vMF normalizing constant.

**Remark 1** (Reduction to InfoNCE)**.** *Under homoscedastic uncertainty ($\upsilon^+ = \upsilon_j^- = \tau$), the log-normalizers $b^+ = b_j^-$ cancel, recovering standard InfoNCE:*

$$\mathcal{L}_{\mathcal{U}\text{-InfoNCE}} \xrightarrow{\upsilon=\tau} -\log \frac{\exp\!\big(\hat{\mathbf{z}}_r^\top \hat{\mathbf{z}}^+/\tau\big)}{\exp\!\big(\hat{\mathbf{z}}_r^\top \hat{\mathbf{z}}^+/\tau\big) + \sum_j \exp\!\big(\hat{\mathbf{z}}_r^\top \hat{\mathbf{z}}_j^-/\tau\big)}. \tag{8}$$

### 3.2 Similarity Uncertainty on the Hypersphere

Let $u_i$ denote the per-view directional variance and $s := \hat{\mathbf{z}}_r^\top \hat{\mathbf{z}}_t$ the mean-evaluated cosine. Under encoder noise the realized cosine $\tilde{s}$ is random, so the probability of identifying the positive among the candidates is an expectation of the candidate softmax over this noise. MCInfoNCE estimates this expectation by Monte Carlo sampling (Kirchhof et al., 2023); we approximate it in closed form. For logits $\tilde{s}_j/\tau$ with variance $V_j/\tau^2$, the probit approximation to the Gaussian–softmax integral (MacKay, 1992; Daxberger et al., 2021) moderates each logit:

$$\mathbb{E}\big[\mathrm{softmax}_j(\tilde{\mathbf{s}}/\tau)\big] \;\approx\; \mathrm{softmax}_j\!\left(s_j \Big/ \sqrt{\tau^2 + \lambda V_j}\right), \qquad \lambda = \pi/8. \tag{9}$$

To first order in $V_j/\tau^2$, $\sqrt{\tau^2+\lambda V_j} \approx \tau + \frac{\lambda}{2\tau} V_j$: similarity noise raises the temperature of each pair additively, in proportion to its variance. By the law of total variance, $\mathrm{Var}(\tilde{s}) = \mathbb{E}_{\mathcal{A}}[\mathrm{Var}_{\boldsymbol{\varepsilon}}(\tilde{s} \mid \mathcal{A})] + \mathrm{Var}_{\mathcal{A}}[\mathbb{E}_{\boldsymbol{\varepsilon}}(\tilde{s} \mid \mathcal{A})]$ separates encoder and augmentation noise. Using the realized encoder variance as a plug-in for its augmentation expectation gives the additive per-pair temperature

$$\upsilon_{r,t} \;=\; \upsilon^{\mathrm{A}}_{r,t} \;+\; \upsilon^{\mathrm{enc}}_{r,t}, \tag{10}$$

where $\upsilon^{\mathrm{enc}}_{r,t}$ is proportional to $\mathrm{Var}_{\boldsymbol{\varepsilon}}(\tilde{s} \mid \mathcal{A})$ and the constant $\lambda/(2\tau)$ is absorbed into the learned scales. The moderated logit $s/\upsilon_{r,t}$ is the vMF exponent of Eq. (5) at $\kappa = 1/\upsilon_{r,t}$. We derive each component below.

**Encoder component.** Projecting the Gaussian posterior onto $S^{L-1}$ yields an approximate vMF with inverse concentration $u_i \approx \sigma_i^2/\|\boldsymbol{\mu}_i\|^2$, $i \in \{r,t\}$ (Mardia & Jupp, 2000). Conditioned on $\mathcal{A}$, we calculate $\upsilon^{\mathrm{enc}}_{r,t}$:

**Proposition 1** (First-order similarity variance)**.** *Let $s = \hat{\mathbf{z}}_r^\top \hat{\mathbf{z}}_t$ and let $\tilde{s}$ denote the cosine similarity of independently sampled, normalized Gaussian representations. In the concentrated-posterior regime, the first-order delta method gives:*

$$\mathrm{Var}_{\boldsymbol{\varepsilon}}(\tilde{s} \mid \mathcal{A}) \;\approx\; (u_r + u_t)(1 - s^2). \tag{11}$$

*The general formulation including tangent-plane cross-covariance is provided with the proof in Appendix C.4.*

We use the propagated similarity variance to adapt the pairwise temperature:

$$\upsilon_{r,t}^{\text{enc}} \approx (u_r + u_t)(1 - s^2). \tag{12}$$

The concentrated-posterior regime also controls the gap between the expected cosine and its mean-evaluated counterpart (Appendix C.5). Operationally, $\upsilon_{r,t}^{\text{enc}}$ yields a similarity-modulated temperature that softens ambiguous pairs while keeping confident pairs close to the baseline.

**Augmentation component.** The remaining component, $\upsilon_{r,t}^{\text{A}} = \tau + \frac{\lambda}{2\tau}\text{Var}_{\mathcal{A}}[\mathbb{E}_{\boldsymbol{\epsilon}}(\tilde{s} \mid \mathcal{A})]$, requires marginalizing over the augmentation distribution for each pair, which is intractable. When encoder uncertainty vanishes ($\upsilon_{r,t}^{\text{enc}} = 0$), the temperature reduces to a global constant analogous to the temperature $\tau$ in deterministic contrastive learning (Wang & Isola, 2020). We therefore model it by a shared scalar, $\upsilon_{r,t}^{\text{A}} \approx \tau_A > 0$.

**Full temperature.** Substituting $\tau_A$ and the encoder component into (10):

$$\upsilon_{r,t} \;=\; \tau_A \;+\; (u_r + u_t)(1 - s^2), \tag{13}$$

which is computed in closed form from $(\hat{\mathbf{z}}_r, \hat{\mathbf{z}}_t, u_r, u_t)$ with $<1\%$ wall-clock overhead relative to deterministic InfoNCE (Appendix C.6). DistCLR predicts one scale per view, with no additional pairwise prediction head. Proposition 1 derives the partner-dependent factor $(1 - s^2)$ under concentrated, independent encoder noise; Table 5(b) ablates geometry and scale assignment. Gradients propagate through both the learned scales and the cosine-dependent factor, with no stop-gradient.

### 3.3 Distributional Contrastive Learning

We omit $-\log p_\theta(\mathcal{X})$ to define a decoder-free contrastive surrogate; with the mean evaluation in Eq. (3), this precludes an exact ELBO interpretation. Applying the product-of-experts rule to the directional scales gives $u_{\mathcal{X}}^{-1} = \sum_{i=1}^{M} u_i^{-1}$. We regularize the aggregated directional variance using the variance component of the Gaussian KL, normalized per dimension:

$$\mathcal{L}_{\text{prior}} = \frac{1}{2}\left(\frac{u_{\mathcal{X}}}{\sigma_p^2} - 1 - \log\frac{u_{\mathcal{X}}}{\sigma_p^2}\right). \tag{14}$$

Here $\sigma_p^2$ sets the reference variance. The regularizer targets the scale-free directional variance rather than $\sigma_i^2$, leaving directional alignment to the contrastive term. The full DistCLR objective reads:

$$\mathcal{L}_{\text{DistCLR}} = \mathcal{L}_{\mathcal{U}\text{-InfoNCE}} + \lambda_{\text{prior}}\mathcal{L}_{\text{prior}}, \tag{15}$$

where $\lambda_{\text{prior}}$ controls the strength of the directional prior regularizer. The objective is symmetric over views (Chen et al., 2020a; Dwibedi et al., 2021); its surrogate approximations are analyzed in Appendix C.7.

## 4 Experiments

We evaluate DistCLR across four standard datasets: CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009), ImageNet-100, and ImageNet (Russakovsky et al., 2015; Deng et al., 2009). Downstream classification is assessed by freezing the pretrained encoder and training a single linear classification layer. To isolate the effect of the learning objective, all data augmentations and encoder architectures are kept consistent with prior contrastive learning work (Estepa et al., 2023), using ResNet-18 and ResNet-50 (He et al., 2016). The learning rate follows the linear-scaling rule $\text{lr} = lr_{\text{base}}|\mathcal{B}|/256$ with cosine decay, where $|\mathcal{B}|$ is the mini-batch size. All experiments use the LARS optimizer (You et al., 2017); full details are in Appendix B.1.

**Hyper-parameters.** For each deterministic baseline, we denote by $\tau_d$ its global tuned temperature, e.g., $\tau_d = 0.1$ for ImageNet. The prior variance is set equal to $\sigma_p^2 = \tau_d/10$. To ensure a fair comparison, the augmentation baseline $\tau_A$ is tuned so that the empirical average temperature satisfies $\mathbb{E}_{(r,t)}[\upsilon_{r,t}] \approx \tau_d$. Since the encoder term in Eq. (13) is nonnegative, we search $\tau_A$ in the interval $(\tau_d - \sigma_p^2, \tau_d)$. Finally, $\lambda_{\text{prior}} = 1$ for SimCLR and All4One and $0.1$ for NNCLR.

### 4.1 Classification Performance

Table 1 reports linear-evaluation results on CIFAR-10, CIFAR-100, and ImageNet-100 using a ResNet-18 backbone. DistCLR improves on its deterministic counterpart in every dataset and metric across all three frameworks. Gains are most pronounced for SimCLR, where the distributional variant improves Top-1 accuracy by $0.31$, $0.99$, and $0.48$ percentage points on CIFAR-10, CIFAR-100, and ImageNet-100. For the stronger NNCLR and All4One baselines, gains are smaller but consistent, with DistCLR-All4One setting new overall best scores of $93.32\%$ on CIFAR-10, $72.08\%$ on CIFAR-100, and $82.02\%$ Top-1 / $95.84\%$ Top-5 on ImageNet-100.

Table 1: Linear-evaluation Top-1 and Top-5 accuracy (%; ResNet-18) of self-supervised baselines and of the deterministic and distributional (ours) SimCLR, NNCLR, and All4One. All experiments were conducted using solo-learn (Da Costa et al., 2022). For the contrastive frameworks, values are mean $\pm$ standard deviation over five independent pretraining runs. Underlined: the distributional variant outperforms its deterministic counterpart; **bold and underlined**: overall best.

| | CIFAR-10 | | CIFAR-100 | | ImageNet-100 | |
|---|---|---|---|---|---|---|
| Method | Top-1 ↑ | Top-5 ↑ | Top-1 ↑ | Top-5 ↑ | Top-1 ↑ | Top-5 ↑ |
| BYOL (Grill et al., 2020) | $92.60_{\pm 0.06}$ | $99.80_{\pm 0.03}$ | $70.51_{\pm 0.11}$ | $91.96_{\pm 0.04}$ | $80.18_{\pm 0.07}$ | $94.98_{\pm 0.05}$ |
| DC V2 (Caron et al., 2020) | $88.81_{\pm 0.09}$ | $99.60_{\pm 0.02}$ | $63.59_{\pm 0.12}$ | $88.07_{\pm 0.06}$ | $75.40_{\pm 0.08}$ | $93.19_{\pm 0.04}$ |
| DINO (Caron et al., 2021) | $89.53_{\pm 0.08}$ | $99.68_{\pm 0.02}$ | $66.83_{\pm 0.10}$ | $90.34_{\pm 0.05}$ | $74.89_{\pm 0.09}$ | $92.89_{\pm 0.04}$ |
| MoCoV2+ (Chen et al., 2020b) | $92.88_{\pm 0.09}$ | $99.76_{\pm 0.02}$ | $69.91_{\pm 0.09}$ | $91.62_{\pm 0.07}$ | $78.17_{\pm 0.11}$ | $95.51_{\pm 0.03}$ |
| MoCoV3 (Chen et al., 2021) | $93.08_{\pm 0.07}$ | $99.79_{\pm 0.01}$ | $68.80_{\pm 0.08}$ | $90.56_{\pm 0.06}$ | $80.39_{\pm 0.09}$ | $95.21_{\pm 0.04}$ |
| ReSSL (Zheng et al., 2021) | $90.63_{\pm 0.08}$ | $99.59_{\pm 0.03}$ | $65.93_{\pm 0.09}$ | $89.72_{\pm 0.05}$ | $76.93_{\pm 0.10}$ | $94.16_{\pm 0.04}$ |
| BT (Zbontar et al., 2021) | $92.04_{\pm 0.07}$ | $99.72_{\pm 0.03}$ | $70.88_{\pm 0.11}$ | $91.94_{\pm 0.04}$ | $80.42_{\pm 0.06}$ | $95.30_{\pm 0.05}$ |
| SimSiam (Chen & He, 2021) | $90.55_{\pm 0.10}$ | $99.75_{\pm 0.02}$ | $66.05_{\pm 0.08}$ | $89.59_{\pm 0.06}$ | $74.48_{\pm 0.12}$ | $93.17_{\pm 0.04}$ |
| SwAV (Caron et al., 2020) | $89.19_{\pm 0.08}$ | $99.68_{\pm 0.02}$ | $64.91_{\pm 0.10}$ | $88.79_{\pm 0.05}$ | $73.99_{\pm 0.12}$ | $92.67_{\pm 0.04}$ |
| VibCReg (Lee & Aune, 2021) | $91.15_{\pm 0.09}$ | $99.76_{\pm 0.02}$ | $67.32_{\pm 0.09}$ | $90.07_{\pm 0.06}$ | $79.82_{\pm 0.08}$ | $94.97_{\pm 0.03}$ |
| VICReg (Bardes et al., 2021) | $92.01_{\pm 0.10}$ | $99.73_{\pm 0.01}$ | $68.47_{\pm 0.08}$ | $90.86_{\pm 0.07}$ | $79.23_{\pm 0.09}$ | $95.05_{\pm 0.04}$ |
| W-MSE (Ermolov et al., 2021) | $88.71_{\pm 0.11}$ | $99.70_{\pm 0.02}$ | $61.32_{\pm 0.12}$ | $87.23_{\pm 0.05}$ | $67.58_{\pm 0.11}$ | $90.97_{\pm 0.06}$ |
| **SimCLR** | | | | | | |
| Deterministic | $90.71_{\pm 0.07}$ | $99.72_{\pm 0.01}$ | $65.80_{\pm 0.12}$ | $88.99_{\pm 0.06}$ | $77.61_{\pm 0.10}$ | $94.11_{\pm 0.05}$ |
| **Distributional (ours)** | $\underline{91.02}_{\pm 0.11}$ | $\underline{99.74}_{\pm 0.03}$ | $\underline{66.79}_{\pm 0.11}$ | $\underline{89.86}_{\pm 0.05}$ | $\underline{78.09}_{\pm 0.07}$ | $\underline{94.90}_{\pm 0.05}$ |
| **NNCLR** | | | | | | |
| Deterministic | $91.86_{\pm 0.09}$ | $99.79_{\pm 0.01}$ | $69.59_{\pm 0.08}$ | $91.51_{\pm 0.06}$ | $79.82_{\pm 0.07}$ | $95.31_{\pm 0.03}$ |
| **Distributional (ours)** | $\underline{91.97}_{\pm 0.09}$ | $\underline{\mathbf{99.82}}_{\pm 0.05}$ | $\underline{69.72}_{\pm 0.06}$ | $\underline{91.54}_{\pm 0.01}$ | $\underline{80.01}_{\pm 0.15}$ | $\underline{95.42}_{\pm 0.06}$ |
| **All4One** | | | | | | |
| Deterministic | $93.01_{\pm 0.08}$ | $99.79_{\pm 0.02}$ | $71.82_{\pm 0.09}$ | $92.98_{\pm 0.04}$ | $81.71_{\pm 0.06}$ | $95.79_{\pm 0.04}$ |
| **Distributional (ours)** | $\underline{\mathbf{93.32}}_{\pm 0.07}$ | $\underline{\mathbf{99.82}}_{\pm 0.01}$ | $\underline{\mathbf{72.08}}_{\pm 0.08}$ | $\underline{\mathbf{93.00}}_{\pm 0.04}$ | $\underline{\mathbf{82.02}}_{\pm 0.06}$ | $\underline{\mathbf{95.84}}_{\pm 0.03}$ |

**ImageNet-1k.** Table 2 reports linear evaluation on ImageNet-1k with a ResNet-50 backbone. Layered on top of NNCLR, the strongest contrastive baseline in this comparison, DistCLR matches its Top-1 ($75.5\%$ vs. $75.4\%$) and Top-5 accuracy ($92.5\%$ vs. $92.3\%$), showing that the per-pair temperature and prior regularizer scale to ImageNet-1k without optimization or stability cost.

### 4.2 Transfer Learning Performance

Table 3 evaluates transfer to seven downstream datasets using frozen ResNet-18 encoders pretrained on ImageNet-100. Adapting All4One, the strongest deterministic baseline, with DistCLR raises

Table 2: Linear-evaluation Top-1 and Top-5 accuracy (%) on ImageNet-1k with a ResNet-50 backbone; DistCLR is applied to NNCLR. **Best** scores are bold.

| Accuracy | SimCLR | MoCo v2 | SimSiam | InfoMin | BT | VICReg | OBoW | BYOL | NNCLR | **DistCLR** |
|---|---|---|---|---|---|---|---|---|---|---|
| Top-1 ↑ | 69.3 | 71.1 | 71.3 | 73.0 | 73.2 | 73.2 | 73.8 | 74.3 | 75.4 | **75.5** |
| Top-5 ↑ | 89.0 | - | - | 91.1 | 91.0 | 91.1 | - | 91.6 | 92.3 | **92.5** |

the average score across the seven datasets from 65.63 to 66.16 (+0.53 points), with the gain concentrated on fine-grained domains: PETS (+2.25), Flowers-102 (+0.72), Aircraft (+0.27), and Food-101 (+0.21). It also improves CIFAR-100 by 0.48 points, stays within 0.2 points on CIFAR-10 and VOC2007, and achieves the best score on five of the seven datasets. This is consistent with the higher effective rank in Table 4: DistCLR preserves intra-class variation that fine-grained probes can exploit while leaving coarse-grained boundaries unchanged.

Table 3: Transfer to seven datasets with frozen ResNet-18 encoders pretrained on ImageNet-100 (means over three pretraining seeds). We report mean-per-class accuracy for Aircraft, Flowers-102, and PETS, 11-point mAP for VOC2007, and top-1 accuracy otherwise. **Best** scores are bold.

| Method | CIFAR-10 | CIFAR-100 | Aircraft | VOC2007 | Flowers-102 | PETS | Food-101 | **Average** |
|---|---|---|---|---|---|---|---|---|
| SimCLR | 79.39 | 54.49 | 30.65 | 67.09 | 76.31 | 64.29 | 50.23 | 60.35 |
| NNCLR | 82.80 | 60.22 | 34.85 | 69.20 | 81.39 | 69.42 | 52.73 | 64.37 |
| All4One | **84.37** | 62.12 | 36.32 | **70.10** | 82.40 | 70.37 | 53.73 | 65.63 |
| **DistCLR** | 84.18 | **62.60** | **36.59** | 70.05 | **83.12** | **72.62** | **53.94** | **66.16** |

Table 4: Representation quality and strict CIFAR-10 6-known/4-unknown open-set recognition (OSR) on the four standard splits (Neal et al., 2018; Vaze et al., 2022). All representation models use a ResNet-18 backbone; representation values are means over three pretraining seeds. OSR values are mean ± sample standard deviation across the four splits, each averaged over three runs; unknown classes are excluded before SSL pretraining and probe fitting. MSP: maximum softmax probability; $\mathcal{U}$: DistCLR's fixed-orientation score; MSP+$X$: mean rank of MSP and $X$, no fitted weights. **Bold**: better paired result (left), best score (right).

| Method | kNN@1 | kNN@5 | EffRank |
|---|---|---|---|
| SimCLR | 59.4 | 80.7 | 81.93 |
| **DistCLR-SimCLR** | **60.9** | **83.0** | **82.19** |
| NNCLR | 63.6 | **86.0** | 166.63 |
| **DistCLR-NNCLR** | **64.0** | **86.0** | **187.04** |
| All4One | **66.2** | 87.5 | 145.44 |
| **DistCLR-All4One** | **66.2** | **87.6** | **155.53** |

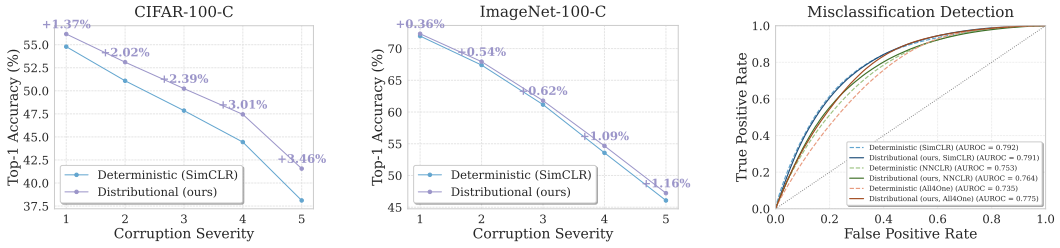
| Model and score | Known acc. | AUROC ↑ | FPR95 ↓ | OSCR ↑ |
|---|---|---|---|---|
| InfoNCE, MSP | $93.68 \pm 2.41$ | $.876 \pm .021$ | $.389 \pm .064$ | $.846 \pm .027$ |
| InfoNCE, energy | $93.68 \pm 2.41$ | $.886 \pm .025$ | $.386 \pm .080$ | $.848 \pm .027$ |
| InfoNCE, MSP+$k$NN | $93.68 \pm 2.41$ | $.889 \pm .029$ | $.371 \pm .085$ | $.851 \pm .028$ |
| DistCLR, MSP | $94.03 \pm 2.25$ | $.881 \pm .022$ | $.385 \pm .075$ | $.852 \pm .025$ |
| DistCLR, MSP+$\mathcal{U}$ | $94.03 \pm 2.25$ | $\mathbf{.905 \pm .011}$ | $\mathbf{.325 \pm .012}$ | $\mathbf{.871 \pm .025}$ |

### 4.3 Semantic Representation Structure

Table 4 shows improvements in neighborhood consistency and semantic organization across the evaluated contrastive frameworks. Across $k$-NN accuracy and effective rank, DistCLR improves or matches every metric for every host framework. Effective rank rises by +20.4 for NNCLR and +10.1 for All4One. In strict CIFAR-10 OSR on the four standard splits, the parameter-free MSP+$\mathcal{U}$ rank average is the best of all tested scores on average, including energy, $k$NN, Mahalanobis, and other fusions (AUROC .905 vs. .889, FPR95 .325 vs. .371 for the best InfoNCE score; Appendix B.4).

### 4.4 Out-of-distribution Generalization

On CIFAR-100-C across five corruption severities (Figure 2(a)), DistCLR maintains an accuracy margin over deterministic SimCLR at every level, with the gap widening monotonically from +1.37% at severity 1 to +3.46% at severity 5. On ImageNet-100-C (Figure 2(b)), at matched clean accuracy, the margin likewise widens from +0.36% to +1.16%, with lower selective-classification risk (AURC) at every severity (Appendix B.5). The learned scale $u_i$ then serves as a per-sample uncertainty signal for misclassification detection (Figure 2(c)): DistCLR improves AUROC on NNCLR (0.753 → 0.764) and All4One (0.735 → 0.775), and matches the strong deterministic SimCLR (0.791 vs. 0.792). The largest gain occurs on All4One, where the deterministic confidence signal is weakest. On CIFAR-100, DistCLR also exceeds the uncertainty-aware methods TaU, MCInfoNCE, and AMCL in both probe accuracy and error AUROC (Appendix B.3).




(a) (b) (c)

Figure 2: Robustness across corruption severities on (a) CIFAR-100-C and (b) ImageNet-100-C, and (c) misclassification detection on CIFAR-100-C. Panels (a) and (b) report means over three pretraining seeds.

Table 5: Ablations. (a) Loss components (means over three pretraining seeds; ticks mark enabled terms, the empty row is fixed InfoNCE). (b) Matched mechanism controls (mean ± sample standard deviation over three runs); error AUROC measures how well the uncertainty score ranks linear-probe mistakes. **Bold**: best score per column.

| $\mathcal{L}_{\mathcal{U}\text{-InfoNCE}}$ | $\mathcal{L}_{\text{prior}}$ | Probe | ARI |
|---|---|---|---|
| | | 65.79 | 16.30 |
| ✓ | | 66.14 | 16.52 |
| | ✓ | 65.40 | 16.48 |
| ✓ | ✓ | **66.80** | **18.48** |

(a)

| Variant | ARI | EffRank | Error AUROC |
|---|---|---|---|
| Fixed InfoNCE | $16.87 \pm 0.81$ | $191.67 \pm 0.69$ | – |
| Pointwise scale | $17.34 \pm 0.06$ | $185.29 \pm 0.05$ | $.540 \pm .001$ |
| Pairwise DistCLR | $\mathbf{18.21 \pm 0.61}$ | $\mathbf{195.49 \pm 0.68}$ | $\mathbf{.749 \pm .002}$ |
| Constant + geometry | $18.16 \pm 1.03$ | $194.86 \pm 0.17$ | $.502 \pm .013$ |
| Shuffled + geometry | $16.80 \pm 0.27$ | $193.56 \pm 1.24$ | $.553 \pm .042$ |

(b)

### 4.5 Ablation Study

Table 5(a) ablates the two loss components. Used separately, $\mathcal{U}$-InfoNCE and the prior remain close to fixed InfoNCE in ARI (16.52 and 16.48 vs. 16.30). Combined, they give the highest ARI and linear-probe accuracy, improving on fixed InfoNCE by 2.18 and 1.01 points (18.48 and 66.80%).

Table 5(b) changes one mechanism at a time under matched optimization: pointwise scaling predicts one candidate-independent scale per view, constant scale with geometry isolates the factor $(1 - s^2)$, and shuffling keeps the learned scale distribution and geometry but breaks the sample-to-scale assignment. Geometry drives most of the structural effect (Figure 3): constant scale with geometry already raises ARI and effective rank over fixed InfoNCE, but detects errors near chance (.502). Correctly assigned learned scales carry the uncertainty signal: pairwise DistCLR attains the highest ARI, effective rank, and error AUROC, improving on shuffled scales with identical geometry by 1.41 ARI, 1.93 effective rank, and .196 error AUROC; its ARI margin over constant scale with geometry is within one standard deviation. Errors are ranked by $1 - u$: the learned scale measures how much softening a view can afford. Views whose positives are easily identified tolerate a larger $u$, whereas partially aligned views are driven toward sharper temperatures (Appendix B.6), so a small $u$ marks a hard view; the learned scale thus ranks uncertainty rather than estimating a calibrated variance. Learned scales also separate errors at matched augmentation disagreement (Appendix B.2).

## 5 Conclusion

We introduced Distributional Contrastive Learning (DistCLR), an ELBO-motivated reformulation of contrastive learning that models both instance representations and their pairwise similarities as latent distributions. A closed-form marginalization of the contrastive softmax over encoder noise yields a heteroscedastic per-pair temperature that sharpens confident pairs and softens ambiguous ones, while a directional prior regularizes the aggregated posterior. DistCLR changes only the pairwise contrastive score, adds negligible overhead, and recovers InfoNCE under homoscedastic uncertainty. Across three contrastive frameworks, it matches or improves classification, with clearer semantic organization, better transfer, and greater corruption robustness on CIFAR-100 and ImageNet-100.

## Ethics Statement

This work uses established public image benchmarks and involves no human participants, private records, or newly collected personal data. The principal risks arise downstream: pretrained representations may inherit demographic or geographic biases from their data, and an uncertainty score may create false confidence outside the evaluated conditions. High-stakes use should therefore include subgroup evaluation, distribution-shift testing, human oversight, and failure procedures. DistCLR adds less than $1\%$ wall-clock overhead relative to deterministic InfoNCE and does not require larger batch sizes or additional augmentation passes.

## Reproducibility Statement

The paper and appendix specify the objective, approximations, architecture, optimization settings, evaluation metrics, and computational overhead. Central mechanism results use three independent full pretraining runs, summarized as mean $\pm$ sample standard deviation in Table 5; OSR results use three independent full pretraining runs on each of the four standard splits, summarized as mean $\pm$ sample standard deviation across splits in Tables 4 and 8. ImageNet-100 corruption results use three pretraining runs on a strict 50-class split (Table 9). For OSR, unknown classes are excluded before SSL loading and probe fitting; the $\mathcal{U}$ score orientation is fixed before evaluation. An anonymized implementation accompanies the submission. The final release will include experiment configurations, split manifests, evaluation scripts, and checkpoint identifiers needed to reproduce the reported tables and figures.

## AI use statement

Generative AI tools assisted with LaTeX formatting, figure and table presentation, language-editing, and retrieval and discovery (e.g., finding related work). The authors are responsible for verifying all AI-assisted text, claims, and artifacts and for the final content of the submission.

## References


Jyoti Aneja, Alex Schwing, Jan Kautz, and Arash Vahdat. A contrastive learning approach for training variational autoencoder priors. *Advances in Neural Information Processing Systems*, 34:480–493, 2021.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *Proceedings of the International Conference on Machine Learning*, pp. 1613–1622. PMLR, 2015.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

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

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*, pp. 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*, pp. 15750–15758, 2021.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9640–9649, 2021.

Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pp. 539–546. IEEE, 2005.

Victor Guilherme Turrisi Da Costa, Enrico Fini, Moin Nabi, Nicu Sebe, and Elisa Ricci. solo-learn: A library of self-supervised methods for visual representation learning. *Journal of Machine Learning Research*, 23(56):1–6, 2022.

Tim R Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M Tomczak. Hyperspherical variational auto-encoders. *arXiv preprint arXiv:1804.00891*, 2018.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux – effortless Bayesian deep learning. In *Advances in Neural Information Processing Systems*, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. Ieee, 2009.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9588–9597, 2021.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *Proceedings of the International Conference on Machine Learning*, pp. 3015–3024. PMLR, 2021.

Imanol G Estepa, Ignacio Sarasúa, Bhalaji Nagarajan, and Petia Radeva. All4one: Symbiotic neighbour contrastive learning via self-attention and redundancy reduction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16243–16253, 2023.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9052–9071, 2024.

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*, pp. 770–778, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

Boris Ivanovic, Karen Leung, Edward Schmerling, and Marco Pavone. Multimodal deep generative models for trajectory prediction: A conditional variational autoencoder approach. *IEEE Robotics and Automation Letters*, 6(2):295–302, 2020.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Michael Kirchhof, Enkelejda Kasneci, and Seong Joon Oh. Probabilistic contrastive learning recovers the correct aleatoric uncertainty of ambiguous inputs. In *Proceedings of the International Conference on Machine Learning*, pp. 17085–17104. PMLR, 2023.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Daesoo Lee and Erlend Aune. Computer vision self-supervised learning methods on time series. *arXiv preprint arXiv:2109.00783*, 2021.

David J. C. MacKay. The evidence framework applied to classification networks. *Neural Computation*, 4(5):720–736, 1992.

Kanti V. Mardia and Peter E. Jupp. *Directional Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Chichester, 2nd edition, 2000. doi: 10.1002/9780470316979.

Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Seong Joon Oh, Kevin Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew Gallagher. Modeling uncertainty with hedged instance embedding. In *Proceedings of the International Conference on Learning Representations*, 2019.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. Probabilistic representations for video contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14711–14721, 2022.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *Proceedings of the International Conference on Machine Learning*, pp. 5171–5180. PMLR, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763. PmLR, 2021.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 1278–1286, 2014.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2007.

Evgenia Rusak, Patrik Reizinger, Attila Juhos, Oliver Bringmann, Roland S. Zimmermann, and Wieland Brendel. InfoNCE: Identifying the gap between theory and practice. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2025.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in Neural Information Processing Systems*, 29, 2016.

Hangke Sui, Yuqing Wang, and Minh N Do. Unicon: Unified framework for efficient contrastive alignment via kernels. In *Proceedings of the International Conference on Learning Representations*, 2026.

Masahiro Suzuki and Yutaka Matsuo. A survey of multimodal deep generative models. *Advanced Robotics*, 36(5-6):261–278, 2022.

Jakub Tomczak and Max Welling. Vae with a vampprior. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 1214–1223, 2018.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *International Conference on Learning Representations*, 2022.

Haoqing Wang, Xun Guo, Zhi-Hong Deng, and Yan Lu. Rethinking minimal sufficient representation in contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16041–16050, 2022.

Lei Wang, Piotr Koniusz, Tom Gedeon, and Liang Zheng. Adaptive multi-head contrastive learning. In *European Conference on Computer Vision*. Springer, 2024.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Ziwen Wang, Jiajun Fan, Thao Nguyen, Heng Ji, and Ge Liu. Variational supervised contrastive learning. In *Proceedings of the International Conference on Neural Information Processing Systems*, 2025.

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, 2018.

Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. In *Proceedings of the International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=CZ8Y3NzuVzO.

Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

Tianhao Yu, Haiyang Cui, Jianan Canal Li, Yunan Luo, Guangde Jiang, and Huimin Zhao. Enzyme function prediction using contrastive learning. *Science*, 379(6639):1358–1363, 2023.

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

Oliver Zhang, Mike Wu, Jasmine Bayrooti, and Noah Goodman. Temperature as uncertainty in contrastive learning. *arXiv preprint arXiv:2110.04403*, 2021.

Mingkai Zheng, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Ressl: Relational self-supervised learning with weak augmentation. *Advances in Neural Information Processing Systems*, 34:2543–2555, 2021.

# Supplementary Material

## A Discussion

### A.1 Limitations and Future Work

Our evidence is currently limited to normalized finite-candidate contrastive objectives for image data. The moderation in Eq. (9) treats logits independently, although all logits of an anchor share its noise (Appendix B.6); moderating positive–negative margins, whose variance includes the shared anchor term, is a natural extension. DistCLR predicts one scale per view and uses a global prior anchor; both choices may be too coarse for modalities with token- or augmentation-specific noise. Future work should test exact directional moments, structured priors, additional open-set datasets, more pretraining seeds, and large-scale Vision Transformers.

### A.2 Broader Impact

DistCLR may help identify ambiguous comparisons and novel inputs when pretrained representations are reused downstream. This ranking signal could support triage and human review in applications such as medical imaging or autonomous perception. It is not a calibrated probability or a standalone safety guarantee; deployment requires task-specific validation under the target distribution. The method adds a small head and avoids extra augmentation passes, limiting its incremental computational cost.

## B Experimental Analysis and Additional Results

### B.1 Implementation Details

**Optimization Details.** All experiments are conducted within the `solo-learn` framework (Da Costa et al., 2022), and unless otherwise specified we adopt the default optimization settings used for NNCLR and All4One in that codebase. Concretely, we train with the LARS optimizer (You et al., 2017) using momentum $0.9$, the linear-scaling learning-rate rule $\text{lr} = lr_{\text{base}}\,|\mathcal{B}|/256$ with cosine decay, and a 10-epoch warm-up phase. Mixed-precision training is enabled, with the variance head evaluated in full precision to keep its output numerically stable. Pretraining durations match the protocols used by NNCLR and All4One on each benchmark: 1000 epochs on CIFAR-10/100, 400 epochs on ImageNet-100, and 1000 epochs on ImageNet-1k. Per-method learning rates, weight decay, temperatures, batch sizes, and device counts are given in the configuration files of the supplementary code; ImageNet-1k runs use 4 NVIDIA H100 GPUs, batch size 4096, and weight decay $10^{-6}$. The variance head itself is a lightweight two-layer MLP with LayerNorm and an exponential output of a softly upper-capped log-variance, mapping projected features to a single positive scalar per sample. The variance head takes the normalized projector output and predicts the directional scale $u_i$ directly. This parameterizes the ratio $\sigma_i^2/\|\boldsymbol{\mu}_i\|^2$ appearing in the derivation without performing an additional division by the projector norm.

**Data Augmentation Strategies.** We use the standard augmentation pipeline shipped with `solo-learn` for self-supervised contrastive learning (random resized crop, horizontal flip, color jitter, grayscale, Gaussian blur, and solarization), keeping it identical to the configurations used by SimCLR, NNCLR and All4One on each dataset to isolate the contribution of the learning objective.

**Positive and Negative Sample Definition.** Following standard self-supervised contrastive learning, positives are two augmented views of the same instance and negatives are drawn from the remainder of the batch, with the queue and nearest-neighbor lookup configurations matching those of the corresponding baseline (NNCLR, All4One) within solo-learn. DistCLR preserves this sampling procedure and changes how each relationship is scored: each pair $(r, t)$ contributes a vMF-based score with its own concentration $1/\upsilon_{r,t}$ derived in closed form from $(\hat{\mathbf{z}}_r, \hat{\mathbf{z}}_t, u_r, u_t)$, replacing the global temperature without altering the underlying batching, queue, or memory-bank machinery. This isolates the effect of similarity uncertainty as a drop-in replacement for the temperature in InfoNCE-family losses.

**Approximation.** In practice, we adopt the small-$\kappa$ expansion $\log C_L(\kappa) \approx -\kappa^2/(2L) + \text{const}$ of the vMF log-normalizer, retaining only the leading $\kappa$-dependent term since the additive constant cancels in the InfoNCE softmax. This quadratic surrogate avoids evaluating the modified Bessel function $I_{L/2-1}(\kappa)$ at every pair, eliminating its associated numerical instabilities while preserving the regularizing behavior of $b_{r,t}$: it still penalizes overconfident concentrations (large $\kappa$, i.e., small $\upsilon_{r,t}$) and maintains the relative ordering of per-pair uncertainties used in Eq. (7).

**Linear evaluation.** Following the standard self-supervised linear-probe protocol (Chen et al., 2020a; Grill et al., 2020), we freeze the pretrained backbone and train a single linear classifier on top of the global-average-pooled features. We report Top-1 and Top-5 accuracy on the test set.

**Transfer learning (Table 3).** We follow the linear-probing protocol of BYOL (Grill et al., 2020), also adopted by NNCLR (Dwibedi et al., 2021), which builds on the benchmark of Kornblith et al. (2019): the ResNet-18 backbone pretrained on ImageNet-100 (400 epochs, as in Table 1) is frozen, features are extracted from the penultimate layer, and an $L_2$-regularized logistic regression is fit per dataset, with the regularization coefficient selected on a held-out validation split. We report top-1 accuracy on CIFAR-10/100 and Food-101, mean-per-class accuracy on Aircraft, Flowers-102 and Oxford-IIIT Pets (which mitigates class imbalance), and 11-point interpolated mAP on Pascal VOC2007.

**Representation quality (Table 4).** We report neighborhood accuracy and effective rank on frozen features: (i) *$k$-NN accuracy at $k \in \{1, 5\}$*, the proportion of test points whose label matches the majority of their $k$ nearest training neighbors under cosine distance, measuring local neighborhood consistency; (ii) *effective rank* of the feature covariance, $\mathrm{ER}(\Sigma) = \exp\big(-\sum_i p_i \log p_i\big)$ with $p_i = \sigma_i / \sum_j \sigma_j$ and $\sigma_i$ the singular values of $\Sigma$, capturing how many directions are effectively used by the encoder (Roy & Vetterli, 2007).

**Open-set recognition (Table 4).** We use the standard CIFAR-10 splits with six known and four unknown classes (Neal et al., 2018; Vaze et al., 2022); the five published lists contain four distinct known-class sets, and we evaluate each distinct set once. Unknown classes are excluded before SSL data loading and linear-probe fitting. We report known-class accuracy, AUROC, FPR95, and OSCR for three independent pretraining runs per split. We use $\mathcal{U}(x) = 1 - u(x)$, where $u(x)$ is the predicted per-view directional scale; larger $\mathcal{U}$ indicates a more likely unknown input. The same transformation is used in all runs. MSP uses $1 - \max_c p(c \mid x)$ as the unknown-class score.

**Robustness and misclassification detection (Fig. 2).** We evaluate on CIFAR-100-C (Hendrycks & Dietterich, 2019), reporting linear-probe Top-1 accuracy averaged over the 19 corruption types at each of the five severity levels; the ImageNet-100-C protocol is given in Appendix B.5. For misclassification detection, we use the score $1 - u_i$, where $u_i$ is the predicted per-view directional scale; larger values indicate a more likely classification error. We report AUROC against the binary correct/incorrect labels of the linear probe.

**Component ablation (Table 5(a)).** Under the same training protocol, we compare fixed InfoNCE, $\mathcal{U}$-InfoNCE alone, the prior alone, and their combination. We report linear-probe accuracy and clustering ARI. ARI measures agreement between $K$-means clusters ($K$ equals the number of classes) and ground-truth labels, corrected for chance.

### B.2 Uncertainty Beyond Augmentation Disagreement

We evaluate three 1,000-epoch SimCLR/ResNet-18 CIFAR-100 checkpoints on the same 2,000 randomly selected test images. Disagreement is one minus the mean pairwise cosine similarity of projected directions across 16 training augmentations per image. A supplementary linear probe uses normalized backbone features, 40,000 training images, and 100 SGD epochs (learning rate $0.1$, momentum $0.9$, cosine decay, batch size $512$). The remaining 10,000 training images determine the learned score's orientation ($1 - u_i$ in all three seeds). Table 6 reports error-detection AUROC overall and averaged within five equal-size disagreement groups. Learned scales distinguish errors even at similar disagreement levels, providing information beyond augmentation consistency.

### B.3 Comparison with Uncertainty-Aware Methods

We compare 1,000-epoch CIFAR-100/ResNet-18 checkpoints using the supplementary probe protocol described above. TaU and MCInfoNCE are adaptations to our training framework; AMCL implements

Table 6: Error detection beyond augmentation disagreement. Within-group values average AUROC over disagreement quintiles. Values are mean ± sample standard deviation over three independent pretraining seeds.

| Score | Overall AUROC | Within-group AUROC |
|---|---|---|
| Augmentation disagreement | $.681 \pm .004$ | $.564 \pm .002$ |
| Learned scale | $.749 \pm .002$ | $.694 \pm .000$ |

the exact InfoNCE formulation in its appendix. Values are mean ± sample standard deviation over three independent pretraining seeds. Error detection uses $1 - \max_c p(c \mid x)$ for every method, where $p(c \mid x)$ is the linear probe's class probability.

Table 7: Uncertainty-aware methods under the supplementary probe protocol. Values are mean ± sample standard deviation over three independent pretraining seeds. MSP denotes maximum softmax probability; larger $1 - \mathrm{MSP}$ indicates a more likely classification error. **Bold**: best score per column.

| Method | Probe accuracy (%) | Error AUROC (MSP) |
|---|---|---|
| DistCLR | $\mathbf{57.91 \pm 0.05}$ | $\mathbf{.813 \pm .000}$ |
| TaU adaptation | $53.83 \pm 0.28$ | $.797 \pm .002$ |
| MCInfoNCE SSL adaptation | $52.94 \pm 0.04$ | $.770 \pm .002$ |
| AMCL, exact InfoNCE | $56.26 \pm 0.23$ | $.779 \pm .000$ |

### B.4 Open-Set Recognition with Post-hoc Baselines

We compare MSP+$\mathcal{U}$ with standard post-hoc unknown scores computed from the same frozen encoders and linear probes: energy (negative log-sum-exp of logits), $k$-NN distance (negative cosine similarity to the 50th nearest known-class training feature), and Mahalanobis distance (class means with a tied covariance), all on normalized backbone features. MSP+$\mathcal{U}$ averages the within-test-set ranks of $1 - \mathrm{MSP}$ and $\mathcal{U}$ with no fitted weights. Scores, $k$, and orientations were fixed before evaluation.

On the four standard CIFAR-10 6/4 splits (Table 8), MSP+$\mathcal{U}$ is the best score on average on every metric and has the smallest spread across splits in AUROC and FPR95. Fusing MSP with energy or $k$-NN in the same way yields smaller gains on either encoder, so the improvement is specific to $\mathcal{U}$, which carries information complementary to classifier confidence. The DistCLR encoder also improves every post-hoc score over the InfoNCE encoder on average.

Table 8: Strict CIFAR-10 OSR on the four standard splits (Neal et al., 2018; Vaze et al., 2022). Each split is the mean of three independent pretraining runs; values are mean ± sample standard deviation across the four splits. Higher scores indicate a more likely unknown input.

| Encoder | Unknown score | AUROC ↑ | FPR95 ↓ | OSCR ↑ |
|---|---|---|---|---|
| InfoNCE | $1-\mathrm{MSP}$ | $.876 \pm .021$ | $.389 \pm .064$ | $.846 \pm .027$ |
| InfoNCE | Energy | $.886 \pm .025$ | $.386 \pm .080$ | $.848 \pm .027$ |
| InfoNCE | $k$-NN ($k = 50$) | $.840 \pm .039$ | $.481 \pm .087$ | $.799 \pm .033$ |
| InfoNCE | Mahalanobis | $.757 \pm .032$ | $.681 \pm .079$ | $.716 \pm .035$ |
| InfoNCE | MSP+Energy (rank mean) | $.885 \pm .023$ | $.384 \pm .070$ | $.851 \pm .026$ |
| InfoNCE | MSP+$k$-NN (rank mean) | $.889 \pm .029$ | $.371 \pm .085$ | $.851 \pm .028$ |
| DistCLR | $1-\mathrm{MSP}$ | $.881 \pm .022$ | $.385 \pm .075$ | $.852 \pm .025$ |
| DistCLR | Energy | $.890 \pm .027$ | $.383 \pm .090$ | $.853 \pm .026$ |
| DistCLR | $k$-NN ($k = 50$) | $.857 \pm .032$ | $.437 \pm .083$ | $.818 \pm .030$ |
| DistCLR | Mahalanobis | $.783 \pm .030$ | $.627 \pm .078$ | $.744 \pm .031$ |
| DistCLR | MSP+Energy (rank mean) | $.889 \pm .024$ | $.382 \pm .084$ | $.856 \pm .025$ |
| DistCLR | MSP+$k$-NN (rank mean) | $.897 \pm .026$ | $.349 \pm .077$ | $.861 \pm .026$ |
| DistCLR | MSP+$\mathcal{U}$ (rank mean) | $\mathbf{.905 \pm .011}$ | $\mathbf{.325 \pm .012}$ | $\mathbf{.871 \pm .025}$ |

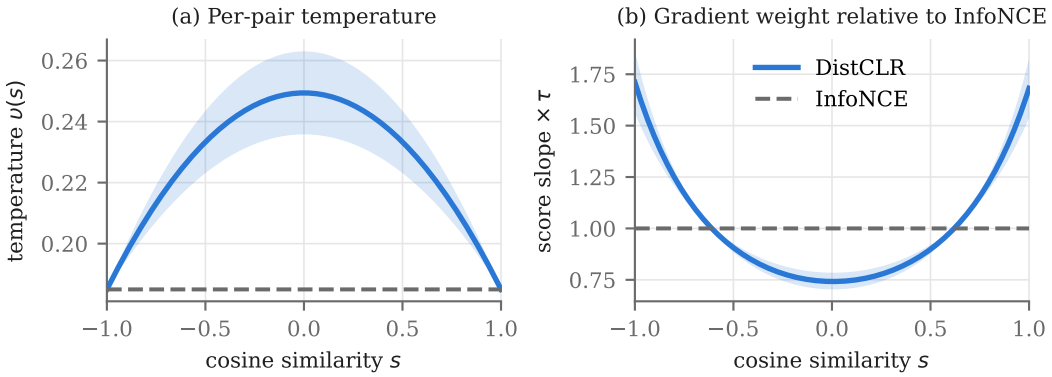



Figure 3: Geometry of the DistCLR critic at the CIFAR-100 SimCLR operating point ($\tau_A = 0.185$, $L = 256$, learned scale $u = 0.032$; band: $\pm 2$ standard deviations of $u$). (a) The per-pair temperature equals $\tau_A$ at $|s| = 1$ and rises for mid-similarity pairs. (b) Slope of the implemented logit with respect to $s$, relative to InfoNCE at temperature $\tau_A$: near-duplicate pairs receive about $1.7\times$ the InfoNCE gradient weight, while near-orthogonal pairs receive about $0.74\times$.

### B.5 Corruption Robustness on ImageNet-100

We pretrain DistCLR-SimCLR and deterministic SimCLR (fixed InfoNCE) on the 50 ImageNet-100 classes with the lowest indices, using the ImageNet-100 recipe of Table 1 (ResNet-18, 400 epochs, batch size 256); the other 50 classes are excluded from pretraining and probe fitting. A linear probe fit on known-class training features is evaluated on the known-class validation images of ImageNet-100-C, i.e., the 15 ImageNet-C corruptions (Hendrycks & Dietterich, 2019) at five severities. Table 9 and Figure 2(b) report three pretraining seeds. At matched clean accuracy (77.5% for both encoders), the DistCLR accuracy margin grows with severity, from $+0.36$ to $+1.16$ points, so the gain under shift is not inherited from clean accuracy. DistCLR also attains higher error AUROC and lower AURC for $1 - \mathrm{MSP}$ at every severity.

Table 9: Covariate shift on ImageNet-100-C (50 known classes, ResNet-18). Known-class accuracy (%, mean $\pm$ sample standard deviation over three pretraining seeds) and the error AUROC and AURC of $1 - \mathrm{MSP}$ (means over the three seeds), each averaged over the 15 corruption types.

| | Accuracy ↑ | | Error AUROC ↑ | | AURC ↓ | |
|---|---|---|---|---|---|---|
| Severity | InfoNCE | DistCLR | InfoNCE | DistCLR | InfoNCE | DistCLR |
| 1 | $71.97 \pm 0.38$ | $72.33 \pm 0.25$ | .869 | .873 | .091 | .088 |
| 2 | $67.40 \pm 0.51$ | $67.94 \pm 0.04$ | .858 | .864 | .124 | .119 |
| 3 | $61.17 \pm 0.31$ | $61.79 \pm 0.12$ | .845 | .849 | .170 | .165 |
| 4 | $53.57 \pm 0.28$ | $54.66 \pm 0.22$ | .829 | .831 | .239 | .230 |
| 5 | $46.06 \pm 0.01$ | $47.22 \pm 0.17$ | .814 | .820 | .311 | .297 |
| Mean | $60.03 \pm 0.29$ | $60.79 \pm 0.06$ | .843 | .848 | .187 | .180 |

### B.6 Learning Signal on the Learned Scale

Write $\ell_j = s_j/\upsilon_j - 1/(2L\upsilon_j^2)$ for the implemented logits, $p^+$ and $p_j$ for the candidate probabilities, and $\upsilon_j = \tau_A + (u_r + u_j)(1 - s_j^2)$. Since $\partial \ell / \partial \upsilon = -(s - 1/(L\upsilon))/\upsilon^2$ and $\partial \upsilon_j / \partial u_r = 1 - s_j^2$, the gradient of $\mathcal{L}^{(r,t)}_{\mathcal{U}\text{-InfoNCE}}$ with respect to the anchor scale is

$$\frac{\partial \mathcal{L}}{\partial u_r} = (1 - p^+)\,\frac{(1 - s^{+2})\big(s^+ - \frac{1}{L\upsilon^+}\big)}{\upsilon^{+2}} \;-\; \sum_{j \in \mathcal{N}} p_j\,\frac{(1 - s_j^2)\big(s_j - \frac{1}{L\upsilon_j}\big)}{\upsilon_j^2}. \tag{B.1}$$

The first term is positive whenever $s^+ > 1/(L\upsilon^+)$, so the positive pair then pushes $u_r$ down; the term vanishes at $s^+ = 1/(L\upsilon^+)$ and at perfect alignment, so it is largest at intermediate alignment.

The second term raises $u_r$ only through negatives with $s_j > 1/(Lv_j)$. The prior sets the average scale, and the positive term lowers it most for partially aligned views, which is consistent with $1 - u$ ranking errors and unknown inputs; this orientation matches the one selected on held-out training data (Appendix B.2) and is fixed before evaluation.

# C Theoretical Analysis

## C.1 ELBO Decomposition

We derive the motivating decomposition in Eq. (3) from the evidence lower bound in Eq. (2). Recall the ELBO:

$$\mathcal{J}_{\text{ELBO}} \;=\; \mathbb{E}_{q_\phi(\mathbf{z}|\mathcal{X})}\big[\log p_\theta(\mathcal{X}|\mathbf{z})\big] \;-\; \mathcal{K}_{\text{ELBO}}, \tag{C.1}$$

with $\mathcal{K}_{\text{ELBO}} = D_{\text{KL}}\big(q_\phi(\mathbf{z}|\mathcal{X}) \,\|\, p(\mathbf{z})\big)$. Applying Bayes' theorem to the likelihood:

$$\log p_\theta(\mathcal{X}|\mathbf{z}) \;=\; \log p_\theta(\mathbf{z}|\mathcal{X}) \;+\; \log p_\theta(\mathcal{X}) \;-\; \log p_\theta(\mathbf{z}). \tag{C.2}$$

We approximate the marginal $p_\theta(\mathbf{z})$ by the aggregated posterior over the induced view-set distribution:

$$p_\theta(\mathbf{z}) \;\approx\; \int p_\theta(\mathbf{z}|\mathcal{X}')\, p_{\mathcal{D}}(\mathcal{X}')\, \mathrm{d}\mathcal{X}' \;=\; \mathbb{E}_{p_{\mathcal{D}}}\big[p_\theta(\mathbf{z}|\mathcal{X}')\big] \;=:\; p_{\mathcal{D}_\theta}(\mathbf{z}). \tag{C.3}$$

The population expectation defines $p_{\mathcal{D}_\theta}$. The corresponding finite-candidate log-average for the pairwise scores is analyzed in Appendix C.7. Substituting Eqs. (C.2)–(C.3) into (C.1):

$$\mathcal{J}_{\text{ELBO}} \;\approx\; \mathbb{E}_q\big[\log p_\theta(\mathbf{z}|\mathcal{X})\big] \;+\; \log p_\theta(\mathcal{X}) \;-\; \mathbb{E}_q\big[\log p_{\mathcal{D}_\theta}(\mathbf{z})\big] \;-\; \mathcal{K}_{\text{ELBO}}. \tag{C.4}$$

Negating and evaluating the expectation at the posterior mean gives the negative ELBO in alignment–uniformity form:

$$-\mathcal{J}_{\text{ELBO}} \;\approx\; \underbrace{-\log p_\theta(\mathbf{z}|\mathcal{X})}_{\text{alignment}} \;+\; \underbrace{\log p_{\mathcal{D}_\theta}(\mathbf{z})}_{\text{uniformity}} + \mathcal{K}_{\text{ELBO}} \;-\; \log p_\theta(\mathcal{X}), \tag{C.5}$$

which is Eq. (3). The alignment term encourages the latent to concentrate near the instance posterior, the uniformity term prevents mode collapse by penalizing high aggregate density, and the prior term regularizes the variational posterior toward the prior. The remaining data-evidence term is omitted by design to obtain a contrastive surrogate (Appendix C.7.5).

## C.2 Pairwise Directional Scoring

The ELBO decomposition motivates alignment and uniformity terms. DistCLR instantiates these terms on projected view pairs using the vMF-based score $\ell_{r,t}$ in Eq. (5). For a candidate set $\mathcal{B}$, define

$$P_\theta(t \mid r, \mathcal{B}) = \frac{\exp(\ell_{r,t})}{\sum_{j\in\mathcal{B}} \exp(\ell_{r,j})}, \qquad \mathcal{L}^{(r,t)}_{\mathcal{U}\text{-InfoNCE}} = -\log P_\theta(t \mid r, \mathcal{B}). \tag{C.6}$$

This is a normalized conditional distribution over candidate indices. The vMF log-density supplies both the directional alignment term and its concentration-dependent correction; the softmax supplies normalization across candidates when concentration varies with the pair. Thus, Eq. (5) is used as a score rather than as a continuous conditional density with candidate-dependent concentration. Under equal temperatures, the corrections cancel and the objective reduces exactly to InfoNCE. For multiple views, the loss averages over ordered positive pairs.

## C.3 Heteroscedastic Self-Normalization Bias

We show that the $\mathcal{U}$-InfoNCE objective (Eq. (7)) incurs the same asymptotic self-normalization bias as standard InfoNCE. Define the per-pair logits

$$\ell^+ := \frac{\hat{\mathbf{z}}_r^\top \hat{\mathbf{z}}^+}{v^+} + b^+, \qquad \ell_j^- := \frac{\hat{\mathbf{z}}_r^\top \hat{\mathbf{z}}_j^-}{v_j^-} + b_j^-, \tag{C.7}$$

where the log-normalizers $b^+, b_j^-$ are defined in Eq. (5). The $\mathcal{U}$-InfoNCE loss can be written as

$$\mathcal{L}_{\mathcal{U}\text{-InfoNCE}} \;=\; -\ell^+ + \log\Big(e^{\ell^+} + \textstyle\sum_{j=1}^{|\mathcal{N}|} e^{\ell_j^-}\Big). \tag{C.8}$$

The *ideal* estimator excluding the positive from the denominator is

$$\tilde{\mathcal{L}} \;=\; -\ell^+ + \log\Big(\textstyle\sum_{j=1}^{|\mathcal{N}|} e^{\ell_j^-}\Big). \tag{C.9}$$

The self-normalization bias is

$$\mathcal{L}_{\mathcal{U}\text{-InfoNCE}} - \tilde{\mathcal{L}} \;=\; \log\left(1 + \frac{e^{\ell^+}}{\sum_{j=1}^{|\mathcal{N}|} e^{\ell_j^-}}\right). \tag{C.10}$$

**Proposition 2** (Heteroscedastic bias rate)**.** *Suppose the negative logits $\{\ell_j^-\}_{j=1}^{|\mathcal{N}|}$ are i.i.d. draws from a distribution with $\mathbb{E}[e^{\ell_j^-}] = \bar{w} < \infty$, $\mathrm{Var}(e^{\ell_j^-}) < \infty$, and $\inf \mathrm{supp}(e^{\ell_j^-}) > 0$. Suppose the positive logit satisfies $e^{\ell^+} \le W_{\max} < \infty$ almost surely. Then:*

$$\mathbb{E}\Big[\mathcal{L}_{\mathcal{U}\text{-InfoNCE}} - \tilde{\mathcal{L}}\Big] \;=\; O\Big(\frac{1}{|\mathcal{N}|}\Big). \tag{C.11}$$

*Proof.* Using $\log(1+x) \le x$ for $x \ge 0$:

$$\mathcal{L}_{\mathcal{U}\text{-InfoNCE}} - \tilde{\mathcal{L}} \;\le\; \frac{e^{\ell^+}}{\sum_{j=1}^{|\mathcal{N}|} e^{\ell_j^-}}. \tag{C.12}$$

By the bounded positive assumption, $e^{\ell^+} \le W_{\max}$ a.s., so it suffices to bound $\mathbb{E}[1/S_{|\mathcal{N}|}]$ where $S_{|\mathcal{N}|} := \sum_{j=1}^{|\mathcal{N}|} e^{\ell_j^-}$. With $\mathrm{Var}(e^{\ell_j^-}) =: v^2 < \infty$, Chebyshev's inequality gives $\Pr(S_{|\mathcal{N}|} < |\mathcal{N}|\bar{w}/2) \le 4v^2/(|\mathcal{N}|\bar{w}^2)$. Splitting:

$$\mathbb{E}\left[\frac{1}{S_{|\mathcal{N}|}}\right] \;\le\; \frac{2}{|\mathcal{N}|\bar{w}} \;+\; \frac{1}{w_{\min}} \cdot \frac{4v^2}{|\mathcal{N}|\bar{w}^2} \;=\; O\Big(\frac{1}{|\mathcal{N}|}\Big), \tag{C.13}$$

where $w_{\min} := \inf \mathrm{supp}(e^{\ell_j^-}) > 0$ (guaranteed since $\upsilon_j^- \ge \tau_A > 0$ and $b_j^-$ is finite). Combining with Eq. (C.12):

$$\mathbb{E}\Big[\mathcal{L}_{\mathcal{U}\text{-InfoNCE}} - \tilde{\mathcal{L}}\Big] \;\le\; W_{\max} \cdot O\Big(\frac{1}{|\mathcal{N}|}\Big) \;=\; O\Big(\frac{1}{|\mathcal{N}|}\Big). \tag{C.14}$$

□

**Remark 2.** *The boundedness conditions hold in practice: the temperatures $\upsilon^+, \upsilon_j^-$ are bounded away from zero by $\tau_A > 0$, the cosine similarities lie in $[-1, 1]$, and the log-normalizers $b^+, b_j^-$ are finite for finite $\kappa$. The heteroscedastic temperatures therefore do not alter the $O(1/|\mathcal{N}|)$ bias rate.*

### C.4 Proof of Proposition 1 (Similarity Variance)

We derive the first-order variance approximation in Eq. (11).

**Setup.** For each view $i \in \{r, t\}$, consider the reparameterized latent $\mathbf{z}_i = \boldsymbol{\mu}_i + \sigma_i \boldsymbol{\varepsilon}_i$ with $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and its projection $\tilde{\mathbf{z}}_i := \mathbf{z}_i/\|\mathbf{z}_i\|$. Write the random cosine as $\tilde{s} = \tilde{\mathbf{z}}_r^\top \tilde{\mathbf{z}}_t$ and its mean-evaluated counterpart as $s = \hat{\mathbf{z}}_r^\top \hat{\mathbf{z}}_t$. Conditioned on a fixed augmentation pair $\mathcal{A} = (a_r, a_t)$, the encoder parameters $(\boldsymbol{\mu}_i, \sigma_i^2)$ are deterministic, so the only source of randomness is $\boldsymbol{\varepsilon}_i$.

**Step 1: Linearization of the projection.** Define $f_i : \mathbb{R}^L \to S^{L-1}$ by $f_i(\mathbf{z}_i) = \mathbf{z}_i/\|\mathbf{z}_i\|$. Its Jacobian at $\boldsymbol{\mu}_i$ is

$$J_i \;:=\; \frac{\partial f_i}{\partial \mathbf{z}_i}\bigg|_{\boldsymbol{\mu}_i} \;=\; \frac{1}{\|\boldsymbol{\mu}_i\|}\big(\mathbf{I} - \hat{\mathbf{z}}_i \hat{\mathbf{z}}_i^\top\big) \;=\; \frac{P_i}{\|\boldsymbol{\mu}_i\|}, \tag{C.15}$$

where $P_i = \mathbf{I} - \hat{\mathbf{z}}_i \hat{\mathbf{z}}_i^\top$ is the orthogonal projector onto the tangent plane at $\hat{\mathbf{z}}_i$. To first order:

$$\tilde{\mathbf{z}}_i \;\approx\; \hat{\mathbf{z}}_i + \frac{\sigma_i}{\|\boldsymbol{\mu}_i\|} P_i \boldsymbol{\varepsilon}_i. \tag{C.16}$$

This approximation has residual $O(\sigma_i^2/\|\boldsymbol{\mu}_i\|^2) = O(u_i)$.

**Step 2: Tangent-plane covariances.** From the linearization (C.16):

$$\mathrm{Cov}(\tilde{\mathbf{z}}_i) \;\approx\; \frac{\sigma_i^2}{\|\boldsymbol{\mu}_i\|^2}\, P_i \;=\; u_i\, P_i, \tag{C.17}$$

$$\mathrm{Cov}(\tilde{\mathbf{z}}_r, \tilde{\mathbf{z}}_t) \;\approx\; \frac{\sigma_r \sigma_t}{\|\boldsymbol{\mu}_r\|\,\|\boldsymbol{\mu}_t\|}\, P_r\, \mathbb{E}[\boldsymbol{\varepsilon}_r \boldsymbol{\varepsilon}_t^\top]\, P_t. \tag{C.18}$$

**Step 3: Gradient of the similarity.** The cosine similarity has tangent-plane gradients:

$$\mathbf{g}_r \;:=\; P_r\, \hat{\mathbf{z}}_t \;=\; \hat{\mathbf{z}}_t - s\, \hat{\mathbf{z}}_r, \qquad \mathbf{g}_t \;:=\; P_t\, \hat{\mathbf{z}}_r \;=\; \hat{\mathbf{z}}_r - s\, \hat{\mathbf{z}}_t, \tag{C.19}$$

where $s = \hat{\mathbf{z}}_r^\top \hat{\mathbf{z}}_t$ is the deterministic mean cosine. Their squared norms are

$$\|\mathbf{g}_r\|^2 \;=\; 1 - 2s^2 + s^2 \;=\; 1 - s^2, \qquad \|\mathbf{g}_t\|^2 \;=\; 1 - s^2. \tag{C.20}$$

**Step 4: Delta-method variance.** By the multivariate delta method:

$$\mathrm{Var}_{\boldsymbol{\varepsilon}}(\tilde{s} \mid \mathcal{A}) \;=\; \mathbf{g}_r^\top \mathrm{Cov}(\tilde{\mathbf{z}}_r)\, \mathbf{g}_r \;+\; \mathbf{g}_t^\top \mathrm{Cov}(\tilde{\mathbf{z}}_t)\, \mathbf{g}_t \;+\; 2\, \mathbf{g}_r^\top \mathrm{Cov}(\tilde{\mathbf{z}}_r, \tilde{\mathbf{z}}_t)\, \mathbf{g}_t \;+\; R, \tag{C.21}$$

where $R$ collects terms beyond the first-order delta approximation.

*First term.* Using Eq. (C.17) and noting $P_r\, \mathbf{g}_r = \mathbf{g}_r$ (since $\hat{\mathbf{z}}_r^\top \mathbf{g}_r = s - s = 0$):

$$\mathbf{g}_r^\top \mathrm{Cov}(\tilde{\mathbf{z}}_r)\, \mathbf{g}_r \;\approx\; u_r\, \|\mathbf{g}_r\|^2 \;=\; u_r\, (1 - s^2). \tag{C.22}$$

*Second term.* Identically, $\mathbf{g}_t^\top \mathrm{Cov}(\tilde{\mathbf{z}}_t)\, \mathbf{g}_t \approx u_t\, (1 - s^2)$.

*Cross term.* Define $\Delta_{r,t} := 2\, \mathbf{g}_r^\top \mathrm{Cov}(\tilde{\mathbf{z}}_r, \tilde{\mathbf{z}}_t)\, \mathbf{g}_t$.

*Higher-order terms.* Terms beyond first order in $u_r, u_t$ are collected in $R$.

Retaining first-order terms gives

$$\mathrm{Var}_{\boldsymbol{\varepsilon}}(\tilde{s} \mid \mathcal{A}) \;\approx\; (u_r + u_t)(1 - s^2) + \Delta_{r,t}, \tag{C.23}$$

which yields Eq. (11) under independent noise. □

**Specialization to independent noise.** When $\boldsymbol{\varepsilon}_r$ and $\boldsymbol{\varepsilon}_t$ are independent (the standard case), $\mathbb{E}[\boldsymbol{\varepsilon}_r \boldsymbol{\varepsilon}_t^\top] = \mathbf{0}$ in Eq. (C.18), whence $\Delta_{r,t} = 0$ and

$$v_{r,t}^{\mathrm{enc}} \;:=\; \mathrm{Var}_{\boldsymbol{\varepsilon}}(\tilde{s} \mid \mathcal{A}) \;\approx\; (u_r + u_t)(1 - s^2). \tag{C.24}$$

### C.5 Mean-Evaluation Consistency

The contrastive logits in Eq. (7) are evaluated at the projected posterior means $\hat{\mathbf{z}}_i := \boldsymbol{\mu}_i / \|\boldsymbol{\mu}_i\|$ instead of reparameterized projections $\tilde{\mathbf{z}}_i := \mathbf{z}_i / \|\mathbf{z}_i\|$ with $\mathbf{z}_i = \boldsymbol{\mu}_i + \sigma_i \boldsymbol{\varepsilon}_i$. We bound the resulting first-moment gap and show it is governed by the same regime as Proposition 1.

Taylor-expanding the projection $\mathbf{z}_i \mapsto \tilde{\mathbf{z}}_i$ around $\boldsymbol{\mu}_i$ (cf. Eq. (C.16)) and using $\mathbb{E}[P_i \boldsymbol{\varepsilon}_i] = \mathbf{0}$:

$$\mathbb{E}_{\boldsymbol{\varepsilon}}[\tilde{\mathbf{z}}_i \mid \mathcal{A}] \;=\; \hat{\mathbf{z}}_i \;+\; O(u_i), \tag{C.25}$$

so for the cosine similarity:

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\big[\tilde{\mathbf{z}}_r^\top \tilde{\mathbf{z}}_t \mid \mathcal{A}\big] \;=\; \hat{\mathbf{z}}_r^\top \hat{\mathbf{z}}_t \;+\; O(u_r + u_t) \;=\; s \;+\; O(u_r + u_t). \tag{C.26}$$

The mean-evaluated similarity is the leading term of the expected cosine at fixed dimension.

### C.6 Computational Cost Analysis

We analyze the overhead of DistCLR relative to deterministic InfoNCE for a mini-batch of $B$ pairs, each with $L$-dimensional representations.

**Standard InfoNCE.** The dominant cost is the $B \times B$ pairwise similarity matrix $S_{ij} = \hat{\mathbf{z}}_r^{(i)\top} \hat{\mathbf{z}}_t^{(j)}$, computed via a matrix multiplication in $O(B^2 L)$.

**DistCLR overhead.** The additional operations beyond standard InfoNCE are:

1. **Directional scale**: the lightweight variance head predicts $u_i$ (the ratio $\sigma_i^2/\|\boldsymbol{\mu}_i\|^2$) directly for each of the $2B$ views (Appendix B.1). Cost: one small MLP forward pass per view, linear in $B$ and negligible relative to the backbone.
2. **Pairwise temperature**: $\upsilon_{ij} = \tau_A + (u_i + u_j)(1 - S_{ij}^2)$ for each of $B^2$ pairs. Cost: $O(B^2)$ elementwise operations reusing the already-computed similarity matrix $S$.
3. **Temperature-scaled logits**: $S_{ij}/\upsilon_{ij}$ requires $O(B^2)$ elementwise divisions.
4. **Log-normalizer**: $b_{ij} = \log C_L(1/\upsilon_{ij})$ for $B^2$ pairs. Direct evaluation would invoke the modified Bessel function $I_{L/2-1}(\kappa)$ at every pair. We instead use the small-$\kappa$ surrogate from Appendix B.1,

$$\log C_L(\kappa) \;\approx\; -\frac{\kappa^2}{2L} + \text{const}, \tag{C.27}$$

which is purely arithmetic and reduces the per-pair evaluation to $O(1)$. The total cost of this step is $O(B^2)$ with small constants.

**Total overhead.** Step 1 is linear in $B$, and Steps 2–4 are $O(B^2)$ elementwise operations; all are dominated by the $O(B^2L)$ similarity computation that is shared with standard InfoNCE. The relative overhead is therefore $O(B^2)/O(B^2L) = O(1/L)$, which is $<1\%$ for typical embedding dimensions $L \geq 128$. We confirm this empirically: on a single NVIDIA A100 GPU with $B = 4096$ and $L = 128$, the per-iteration wall-clock time of DistCLR is within $0.8\%$ of deterministic InfoNCE.

### C.7 Surrogate Approximation Analysis

The full DistCLR objective $\mathcal{L}_{\text{DistCLR}} = \mathcal{L}_{\mathcal{U}\text{-InfoNCE}} + \lambda_{\text{prior}}\mathcal{L}_{\text{prior}}$ (Eq. (15)) is motivated by the negative ELBO. We analyze retained-term approximations; these do not bound the omitted data-evidence term or establish an ELBO guarantee.

#### C.7.1 Source 1: Finite-Candidate Uniformity Estimation

We analyze the finite-candidate estimate of the population log-average score. For a fixed anchor $\hat{\mathbf{z}}_r$ and independently sampled candidate views $\mathcal{X}_j$, define

$$w(\hat{\mathbf{z}}_r, \mathcal{X}') := \exp(\ell(r, \mathcal{X}'))\,, \qquad \bar{w} := \mathbb{E}_{p_{\mathcal{D}}}[w(\hat{\mathbf{z}}_r, \mathcal{X}')]. \tag{C.28}$$

The population uniformity term is $\log \bar{w}$, whereas the contrastive mini-batch estimator uses

$$\widehat{p}_B(\hat{\mathbf{z}}_r) := \frac{1}{B}\sum_{j=1}^{B} w(\hat{\mathbf{z}}_r, \mathcal{X}_j), \qquad \widehat{U}_B(\hat{\mathbf{z}}_r) := \log \widehat{p}_B(\hat{\mathbf{z}}_r). \tag{C.29}$$

Assume $0 < w_{\min} \leq w \leq w_{\max} < \infty$ and $\mathrm{Var}(w) < \infty$. A second-order Taylor expansion of $\log \widehat{p}_B$ around $\bar{w}$ gives

$$\mathbb{E}\big[\widehat{U}_B(\hat{\mathbf{z}}_r)\big] = \log \bar{w} - \frac{\mathrm{Var}(w)}{2B\bar{w}^2} + O(B^{-3/2}), \tag{C.30}$$

so the finite-candidate log-average bias is $O(1/B)$. Separately, Appendix C.3 bounds the effect of including the positive candidate in the normalizer. Heteroscedastic temperatures preserve both rates under the stated assumptions.

#### C.7.2 Source 2: Pairwise Directional Proxy

The pairwise construction defines a normalized conditional distribution over contrastive candidates (Appendix C.2). Its negative log-probability is exactly the contrastive term in Eq. (7); the ELBO decomposition supplies the alignment–uniformity motivation.

### C.7.3 Source 3: Mean Evaluation and Similarity Variance

The third gap comes from evaluating logits at projected posterior means and using the first-order conditional similarity variance to adapt the temperature. Appendix C.5 shows

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\tilde{\mathbf{z}}_r^\top \tilde{\mathbf{z}}_t \mid \mathcal{A}\right] = \hat{\mathbf{z}}_r^\top \hat{\mathbf{z}}_t + O(u_r + u_t), \tag{C.31}$$

and Appendix C.4 gives

$$\mathrm{Var}_{\boldsymbol{\varepsilon}}(\tilde{s} \mid \mathcal{A}) \approx (u_r + u_t)(1 - s^2). \tag{C.32}$$

These are first-order local expansions; the proportionality constant of the moderated temperature is absorbed into the learned scales (Section 3.2).

### C.7.4 Source 4: Shared Augmentation Temperature

The exact augmentation contribution $\upsilon_{r,t}^{\mathrm{A}} = \tau + \frac{\lambda}{2\tau}\mathrm{Var}_{\mathcal{A}}[\mathbb{E}_{\boldsymbol{\varepsilon}}(\tilde{s} \mid \mathcal{A})]$ is pair-dependent and would require repeated marginalization over the augmentation distribution. DistCLR replaces it by a shared scalar $\tau_A$. We treat the resulting pair-wise spread as a modeling assumption:

$$\delta_A(r,t) := \upsilon_{r,t}^{\mathrm{A}} - \tau_A, \qquad |\delta_A(r,t)| \le \epsilon_A. \tag{C.33}$$

Here $\epsilon_A$ is a modeling assumption instead of a quantity computed in closed form; it is small whenever the augmentation-induced similarity spread is approximately stable across pairs.

For a vMF logit $h(s,\upsilon) = s/\upsilon + \log C_L(1/\upsilon)$ with $\upsilon \ge \upsilon_{\min} > 0$, the derivative satisfies

$$\left|\frac{\partial h}{\partial \upsilon}\right| = \frac{|A_L(1/\upsilon) - s|}{\upsilon^2} \le \frac{2}{\upsilon_{\min}^2}, \tag{C.34}$$

where $A_L(\kappa) = I_{L/2}(\kappa)/I_{L/2-1}(\kappa) \in [0,1]$ is the vMF mean-resultant length and $s \in [-1,1]$. Therefore replacing $\upsilon_{r,t}^{\mathrm{A}}$ by $\tau_A$ perturbs each logit by at most $2\epsilon_A/\upsilon_{\min}^2$. Since log-sum-exp is 1-Lipschitz in the $\ell_\infty$ norm, the induced loss perturbation is

$$|\Delta_A \mathcal{L}_{\mathcal{U}\text{-InfoNCE}}| \le \frac{4\epsilon_A}{\upsilon_{\min}^2}, \tag{C.35}$$

where the factor two accounts for perturbing both the positive logit and the denominator logits. Thus the shared augmentation scalar is accurate when the augmentation-induced similarity spread is approximately stable across pairs; the encoder term then captures the remaining pair-specific uncertainty.

### C.7.5 Source 5: Dropped Marginal Term and Scale Regularization

Finally, Eq. (3) contains the term $-\log p_\theta(\mathcal{X})$, which generally depends on model parameters. We omit it by design to define a decoder-free contrastive surrogate. This does not preserve ELBO gradients or provide a likelihood bound. The additive $-\log|\mathcal{B}|$ in Eq. (6) is parameter-independent for fixed batch size and can be dropped without changing gradients.

The implemented prior regularizer retains the variance component of the Gaussian KL, normalized per dimension:

$$\mathcal{L}_{\mathrm{prior}} = \frac{1}{2}\left(\frac{u_{\mathcal{X}}}{\sigma_p^2} - 1 - \log\frac{u_{\mathcal{X}}}{\sigma_p^2}\right), \tag{C.36}$$

matching Eq. (14). This nonnegative penalty is minimized at $u_{\mathcal{X}} = \sigma_p^2$. It regularizes the aggregate scale without a mean penalty and is distinct from the full posterior-to-prior KL $\mathcal{K}_{\mathrm{ELBO}}$ in the motivating decomposition.

### C.7.6 Scope of the Approximation Analysis

The analysis separates finite-candidate estimation, local cosine-moment propagation, and the shared augmentation baseline. The candidate-normalized contrastive objective is defined exactly by Eq. (C.6); its temperature uses the first-order variance in Eq. (11), and its scale penalty is given in Eq. (14). Larger candidate sets improve the finite-sample estimates, while concentration relative to dimension controls the local moment approximations. These component analyses support the construction without requiring an ELBO-gap bound.