

The Geometric Mechanics of Contrastive Representation Learning: Alignment Potentials, Entropic Dispersion, and Cross-Modal Divergence*

Yichao Cai[†] Zhen Zhang^{†,‡} Yuhang Liu^{†,‡} Javen Qinfeng Shi^{†,‡}

[†] *Australian Institute for Machine Learning, Adelaide University*

[‡] *Responsible AI Research Centre*

✉ yichao.cai@adelaide.edu.au

Abstract. While InfoNCE underlies modern contrastive learning, its geometric mechanisms remain under-characterized beyond the canonical alignment–uniformity decomposition. We develop a measure-theoretic framework in which representation measures evolve on a fixed embedding manifold. In the large-batch limit, we prove value and gradient consistency, linking the stochastic objective to explicit deterministic energy landscapes and revealing a geometric bifurcation between unimodal and symmetric multimodal regimes. In the unimodal case, the intrinsic energy is strictly convex and admits a unique Gibbs equilibrium, showing that entropy acts as a tie-breaker within the aligned basin. In the multimodal case, the intrinsic geometry becomes cross-coupled and contains a persistent negative symmetric divergence term: each modality’s marginal reshapes the effective landscape of the other, allowing strong pairwise alignment to coexist with a persistent modality gap. Controlled synthetic experiments and analyses of pretrained CLIP representations support these predictions. Overall, our results shift the analytical lens from pointwise discrimination to population geometry, showing that pairwise alignment alone is insufficient to control cross-modal marginal structure.

1 Introduction

InfoNCE-based contrastive learning is a central objective in self-supervised and multimodal representation learning (Oord et al., 2018; Radford et al., 2021; Jia et al., 2021), yet its representational geometry remains poorly understood. In multimodal systems such as CLIP, training can achieve strong *pairwise* alignment while the modality *marginals* remain separated, an empirically observed phenomenon known as the modality gap (Liang et al., 2022; Levi and Gilboa, 2025), for which a mechanistic explanation is still lacking. Even in the unimodal setting, it remains unclear what population-level energy InfoNCE induces and how its geometry should be understood beyond the alignment–uniformity view (Wang and Isola, 2020). This raises a key geometric question for representation learning: why does matching positive pairs fail to match the induced population distributions? More broadly, what population energies do unimodal and symmetric multimodal InfoNCE induce, and what equilibria does their geometry favor?

Existing theory still lacks a first-principles account of these questions. Density-ratio analyses characterize the optimal critic as pointwise mutual information (Gutmann and Hyvärinen, 2010; Oord et al., 2018), but do not directly describe the deterministic descent directions induced by the softmax denominator or how these directions reshape the induced representations. The alignment–uniformity view (Wang and Isola, 2020) clarifies important asymptotic trade-offs, but leaves open what population structure the InfoNCE objective itself favors. Identifiability results offer recovery guarantees under explicit generative assumptions (Zimmermann et al., 2021; Liu et al., 2026), addressing what is learnable in principle rather than the objective-induced geometry of the learned marginals. This lack of a geometric account is particularly consequential in multimodal settings, where the symmetric InfoNCE objective is formally an average of two directional terms, yet its induced geometry is not merely the simple average of their individual geometries.

We therefore adopt a measure-theoretic viewpoint that makes the population geometry of InfoNCE explicit (see Fig. 1 for an illustration). In the canonical feature normalization setting, we regard the embedding space $\mathcal{Z}$ as a fixed compact manifold equipped with a volume measure μ . Encoders push data distributions forward onto $\mathcal{Z}$, thereby inducing both representation laws and positive-pair laws directly in the embedding space. Under the exponential similarity kernel, the InfoNCE denominator converges, in the large-batch limit, to a population *partition field*: a kernel-averaged functional of the current representation law that determines the softmax gradient. We show that stochastic InfoNCE is consistent with a closed-form deterministic energy over representation measures, both in value and in gradient. This passage from finite-batch objectives to population energies allows us to separate intrinsic geometric effects from parametrization or training artifacts.

*Please cite the ICML 2026 version: <https://openreview.net/forum?id=X9VtjdBlQw>; see App. H for a summary of the notation and expository updates in this preprint version. Code available at: https://github.com/YichaoCai1/InfoNCE_Geometry.

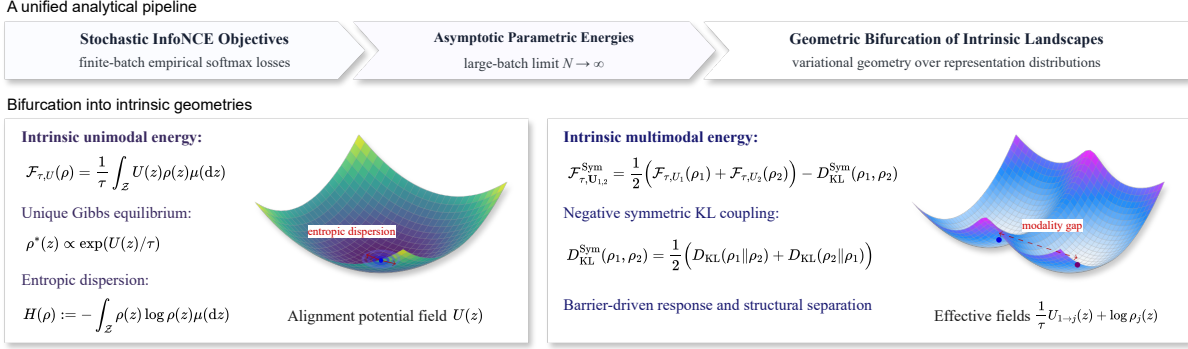


Figure 1: A unified analytical pipeline for contrastive learning geometry. Starting from stochastic InfoNCE losses, the large-batch limit yields deterministic parametric energies, which in turn lift to intrinsic landscapes over representation densities. At this intrinsic level, the geometry bifurcates: in the unimodal case, the landscape is strictly convex, yielding a Gibbs equilibrium; multimodally, a negative symmetric KL coupling induces barrier-driven coordinate responses and can sustain structural separation. The landscape visualizations are illustrative schematics.

From this intrinsic view, unimodal and symmetric multimodal contrastive learning exhibit a fundamental *geometric bifurcation*. In the unimodal regime, the intrinsic population energy is strictly convex and admits a unique Gibbs equilibrium: at low temperature, mass concentrates in highly aligned regions, while entropy controls dispersion *within* those regions. This reframes “uniformity” (Wang and Isola, 2020) not as a global force opposing alignment, but as entropy-driven spreading inside the alignment basin. In contrast, symmetric multimodal InfoNCE induces a cross-coupled geometry: heterogeneous modalities can define different conditional laws and hence different alignment fields, producing a negative symmetric divergence coupling between their marginals. Rather than reducing to a shared unimodal-style equilibrium, each modality enters the effective landscape of the other through a density-dependent barrier, yielding barrier-driven coordinate geometry. Exact marginal coincidence therefore requires additional compatibility between the two directional structures; absent such compatibility, the cross-coupled geometry can sustain a persistent modality gap even when pairwise alignment is strong.

These intrinsic results are not merely formal: under suitable encoder expressivity, the corresponding parametric problems inherit these intrinsic concentration and barrier structures up to controllable KDE mismatch, linking the population-level theory back to encoder parametrizations. We validate the resulting predictions through controlled synthetic experiments and measurements on pretrained CLIP representations. The results show that strong retrieval performance can coexist with a measurable modality gap, and that weakening cross-modal compatibility systematically enlarges this gap even when category-level semantics are preserved. Together, these findings suggest that the modality gap is not merely a failure of pairwise alignment, but reflects a population-level geometric effect; reducing it may therefore require controlling cross-modal divergence beyond positive-pair matching alone.

Contributions. Our main contributions are: (i) large-batch value- and gradient-consistency results connecting stochastic InfoNCE to deterministic population energies; (ii) a characterization of the intrinsic unimodal InfoNCE geometry as a strictly convex population energy with a unique Gibbs equilibrium, together with a reinterpretation of “uniformity” as entropy-driven dispersion within the alignment basin; (iii) a derivation of the multimodal bifurcation, in which a negative symmetric divergence coupling induces barrier-driven coordinate geometry and can sustain a modality gap under conditional heterogeneity; (iv) synthetic and real-data evidence supporting these predictions through controlled experiments and measurements on pretrained OpenCLIP models across CNN- and ViT-based backbones.

2 Problem Setup

We introduce the population-level objects used throughout the analysis for unimodal and symmetric multimodal InfoNCE. For ease of reference, Tab. 1 summarizes the principal notations used throughout the paper.

Data, encoders, and induced laws. Let $\mathcal{P}(\mathcal{S})$ denote the set of probability measures on a measurable space $\mathcal{S}$. In the unimodal setting, $r_{\text{um}} \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$ governs positive observational pairs $(\mathbf{x}, \tilde{\mathbf{x}})$, typically two

views of the same underlying example. We assume that its two $\mathcal{X}$ -marginals coincide, and denote the common marginal by $p_{\mathbf{x}}$. In the multimodal setting, $r_{\text{mm}} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ governs positive cross-modal pairs $(\mathbf{x}, \mathbf{y})$, where $\mathcal{X}$ and $\mathcal{Y}$ are measurable sample spaces; we write $p_{\mathbf{x}}$ and $p_{\mathbf{y}}$ for its $\mathcal{X}$ - and $\mathcal{Y}$ -marginals.

We consider two families of measurable encoders $f_{\theta} : \mathcal{X} \rightarrow \mathcal{Z}$, $\theta \in \Theta$, and $g_{\phi} : \mathcal{Y} \rightarrow \mathcal{Z}$, $\phi \in \Phi$, with compact parameter spaces Θ and Φ , where $\mathcal{Z} \subset \mathbb{R}^n$ is a common representation space. We model $\mathcal{Z}$ as a Riemannian manifold (e.g., $\mathbb{S}^{n-1}$) equipped with a non-atomic Riemannian volume measure μ . These encoders induce the pushforward laws

$$q_{\theta} := (f_{\theta})_{\#} p_{\mathbf{x}}, \quad \pi_{\theta\theta} := (f_{\theta} \times f_{\theta})_{\#} r_{\text{um}}, \quad q_{\phi} := (g_{\phi})_{\#} p_{\mathbf{y}}, \quad \pi_{\theta\phi} := (f_{\theta} \times g_{\phi})_{\#} r_{\text{mm}}.$$

Here, q_{θ} and q_{ϕ} are the encoded marginal laws, while $\pi_{\theta\theta}$ and $\pi_{\theta\phi}$ are the encoded positive-pair laws.

InfoNCE objectives in contrastive learning. We consider InfoNCE objectives based on a symmetric bounded measurable similarity function $\text{sim} : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$ satisfying $\text{sim}(z, w) = \text{sim}(w, z)$, and define the exponential kernel

$$\kappa_{\tau}(z, w) := \exp(\text{sim}(z, w)/\tau), \quad (1)$$

where $\tau > 0$ is the temperature. The unimodal and multimodal objectives are defined as follows.

(i) *Unimodal objective.* For a training batch $\mathcal{B}$ composed of a positive pair $(\mathbf{x}, \tilde{\mathbf{x}}) \sim r_{\text{um}}$ and N negative samples $\{\mathbf{x}'_j\}_{j=1}^N \stackrel{\text{iid}}{\sim} p_{\mathbf{x}}$, the expected finite- N InfoNCE objective is¹

$$\mathcal{L}_{\text{NCE}, N}(\theta) := \mathbb{E}_{\mathcal{B}} \left[-\log \frac{\kappa_{\tau}(f_{\theta}(\mathbf{x}), f_{\theta}(\tilde{\mathbf{x}}))}{\kappa_{\tau}(f_{\theta}(\mathbf{x}), f_{\theta}(\tilde{\mathbf{x}})) + \sum_{j=1}^N \kappa_{\tau}(f_{\theta}(\mathbf{x}), f_{\theta}(\mathbf{x}'_j))} \right]. \quad (2)$$

(ii) *Symmetric multimodal objective.* Given a cross-modal positive pair $(\mathbf{x}, \mathbf{y}) \sim r_{\text{mm}}$ and negatives $\{\mathbf{y}'_j\}_{j=1}^N \stackrel{\text{iid}}{\sim} p_{\mathbf{y}}$, define the expected directional objective

$$\mathcal{L}_{\text{NCE}, N}^{x \rightarrow y}(\theta, \phi) := \mathbb{E}_{\mathcal{B}_{x \rightarrow y}} \left[-\log \frac{\kappa_{\tau}(f_{\theta}(\mathbf{x}), g_{\phi}(\mathbf{y}))}{\kappa_{\tau}(f_{\theta}(\mathbf{x}), g_{\phi}(\mathbf{y})) + \sum_{j=1}^N \kappa_{\tau}(f_{\theta}(\mathbf{x}), g_{\phi}(\mathbf{y}'_j))} \right],$$

where $\mathcal{B}_{x \rightarrow y} := \{\mathbf{x}, \mathbf{y}, \mathbf{y}'_1, \dots, \mathbf{y}'_N\}$. Define $\mathcal{L}_{\text{NCE}, N}^{y \rightarrow x}(\phi, \theta)$ analogously by swapping modalities. The symmetric multimodal InfoNCE objective is

$$\mathcal{L}_{\text{Sym}, N}(\theta, \phi) := \frac{1}{2} (\mathcal{L}_{\text{NCE}, N}^{x \rightarrow y}(\theta, \phi) + \mathcal{L}_{\text{NCE}, N}^{y \rightarrow x}(\phi, \theta)). \quad (3)$$

Population partition fields and smoothed densities. At the population level, the InfoNCE denominator induces kernel-averaged fields on $\mathcal{Z}$. We use these fields to define the corresponding smoothed representation densities. To normalize them, we impose the following condition.

Assumption 2.1 (Constant kernel volume). The representation space $\mathcal{Z}$ is compact with $\mu(\mathcal{Z}) < \infty$. For each $\tau > 0$, there exists $V_{\kappa}(\tau) \in (0, \infty)$ such that

$$\int_{\mathcal{Z}} \kappa_{\tau}(z, w) \mu(dw) = V_{\kappa}(\tau) \quad \text{for all } z \in \mathcal{Z}. \quad (4)$$

Remark 2.1 (When [Asm. 2.1](#) holds). A sufficient condition for [Asm. 2.1](#) is that $\mathcal{Z}$ is a compact homogeneous Riemannian manifold, μ is its invariant volume measure, and κ_{τ} is isotropic, i.e., depends only on geodesic distance. In that case, $V_{\kappa}(\tau, z) := \int_{\mathcal{Z}} \kappa_{\tau}(z, w) \mu(dw)$ is independent of z , so [Asm. 2.1](#) holds exactly.

Definition 2.1 (Population partition fields). Given representation laws q_{θ} and q_{ϕ} , define the population partition fields

$$\Gamma_{\theta, \tau}(z) := \int_{\mathcal{Z}} \kappa_{\tau}(z, w) q_{\theta}(dw), \quad \Gamma_{\phi, \tau}(z) := \int_{\mathcal{Z}} \kappa_{\tau}(z, w) q_{\phi}(dw).$$

¹For analytical clarity, we use an iid negative-sampling formulation. Extending the formal large-batch argument to dependent in-batch negatives requires additional sampling assumptions and is not considered here.

Definition 2.2 (Kernel-smoothed representation density). Under [Asm. 2.1](#), define the kernel-smoothed representation densities $\tilde{\rho}_{\theta,\tau}, \tilde{\rho}_{\phi,\tau} : \mathcal{Z} \rightarrow (0, \infty)$ by

$$\tilde{\rho}_{\theta,\tau}(z) := \frac{\Gamma_{\theta,\tau}(z)}{V_{\kappa}(\tau)}, \quad \tilde{\rho}_{\phi,\tau}(z) := \frac{\Gamma_{\phi,\tau}(z)}{V_{\kappa}(\tau)}.$$

By symmetry of κ_{τ} and [Asm. 2.1](#), both densities integrate to one with respect to μ . Equivalently, $\tilde{\rho}_{\theta,\tau}(z) \mu(dz)$ and $\tilde{\rho}_{\phi,\tau}(z) \mu(dz)$ are the kernel-smoothed probability measures on $\mathcal{Z}$.

Remark 2.2 (Finite- τ smoothing and positivity). For any $\tau > 0$, $\tilde{\rho}_{\theta,\tau}$ is a kernel smoothing of q_{θ} via $\Gamma_{\theta,\tau}$. In particular, it defines a μ -density even when q_{θ} is singular with respect to μ . If $\kappa_{\tau}(z, w)$ is continuous and strictly positive on $\mathcal{Z} \times \mathcal{Z}$, then $\tilde{\rho}_{\theta,\tau}$ is continuous and strictly positive on $\mathcal{Z}$; the same holds for $\tilde{\rho}_{\phi,\tau}$.

Here, $\Gamma_{\theta,\tau}(z)$ is the population negative-partition field associated with a query at z , while $\tilde{\rho}_{\theta,\tau}$ is the corresponding kernel-smoothed density of q_{θ} at scale τ . For i.i.d. negative representation variables $\mathbf{w}_j \sim q_{\theta}$, $\frac{1}{N} \sum_{j=1}^N \kappa_{\tau}(z, \mathbf{w}_j)$ is a Monte Carlo estimator of $\Gamma_{\theta,\tau}(z)$; analogously, $\Gamma_{\phi,\tau}$ arises when negatives are drawn from q_{ϕ} . These population fields, or equivalently their normalized densities under [Asm. 2.1](#), are the quantities that enter the large-batch parametric energies and pointwise gradient limits developed in the following sections.

3 Unimodal InfoNCE: Convex Geometry and Gibbs Equilibrium

This section develops the unimodal baseline that underlies the later multimodal bifurcation. We first show that, in the large-batch regime, unimodal InfoNCE admits an explicit deterministic energy limit depending on θ only through the induced laws $(q_{\theta}, \pi_{\theta\theta})$ on $\mathcal{Z}$ and $\mathcal{Z} \times \mathcal{Z}$. Proofs for all formal statements in this section are deferred to [App. C](#).

3.1 Large-Batch Deterministic Energy

We begin by identifying the population-level energy arising as the large-batch limit of unimodal InfoNCE. The resulting closed-form parametric energy $\mathcal{J}_{\tau}(\theta)$ has a gradient that matches the InfoNCE gradient up to vanishing error as $N \rightarrow \infty$.

Assumption 3.1 (Optimization regularity). Assume the following conditions hold:

(i) (Encoder regularity). For every $\mathbf{x} \in \mathcal{X}$, the map $\theta \mapsto f_{\theta}(\mathbf{x})$ is C^1 on Θ , and its parameter Jacobian is uniformly bounded:²

$$\sup_{\theta \in \Theta} \sup_{\mathbf{x} \in \mathcal{X}} \|J_{\theta} f_{\theta}(\mathbf{x})\|_{\text{op}} < \infty.$$

(ii) (Similarity regularity). The similarity function sim is C^1 on an open neighborhood of $\mathcal{Z} \times \mathcal{Z}$, and its input gradient is uniformly bounded on $\mathcal{Z} \times \mathcal{Z}$:

$$\sup_{(z,w) \in \mathcal{Z} \times \mathcal{Z}} \|\nabla \text{sim}(z, w)\| < \infty.$$

Remark 3.1 (Optimization stability). [Asm. 3.1](#) ensures that InfoNCE gradients remain uniformly controlled over Θ , which is needed to justify exchanging the large-batch limit with differentiation. Further discussion and the formal uniform bounds are given in [App. A.1](#) and [Lem. B.1](#).

For fixed τ , the normalized finite- N negative partition sum concentrates around the population partition field $\Gamma_{\theta,\tau}$ from [Def. 2.1](#). This yields a closed-form population energy with value and gradient consistency in the large-batch limit.

Definition 3.1 (Alignment potential field). Let $\pi_{\theta\theta} \in \mathcal{P}(\mathcal{Z} \times \mathcal{Z})$ be the encoded positive-pair law with first marginal $q_{\theta} \in \mathcal{P}(\mathcal{Z})$. By disintegration, there exists a measurable family of probability measures $\{\nu_{\theta,z}\}_{z \in \mathcal{Z}} \subset \mathcal{P}(\mathcal{Z})$ such that

$$\pi_{\theta\theta}(dz, dw) = q_{\theta}(dz) \nu_{\theta,z}(dw)$$

as measures on $\mathcal{Z} \times \mathcal{Z}$. We define the alignment potential field by

$$U_{\theta}(z) := - \int_{\mathcal{Z}} \text{sim}(z, w) \nu_{\theta,z}(dw), \quad q_{\theta}\text{-a.e. } z. \quad (5)$$

²Throughout, $\|\cdot\|$ denotes the ℓ_2 norm for vectors unless specified; $\|\cdot\|_{\text{op}}$ denotes the induced operator norm for linear maps.

Definition 3.2 (Unimodal parametric energy). Fix $\tau > 0$. Let $\tilde{\rho}_{\theta,\tau}$ be the kernel-smoothed μ -density from Def. 2.2, and let U_θ be the alignment potential field from Def. 3.1. Define the unimodal parametric energy $\mathcal{J}_\tau : \Theta \rightarrow \mathbb{R}$ by

$$\mathcal{J}_\tau(\theta) := \frac{1}{\tau} \int_{\mathcal{Z}} U_\theta(z) q_\theta(dz) - H_\times \text{sim}(q_\theta, \tilde{\rho}_{\theta,\tau}), \quad (6)$$

where $H_\times \text{sim}(q, \tilde{\rho}) := - \int_{\mathcal{Z}} \log \tilde{\rho}(z) q(dz)$ denotes the cross-entropy of the probability measure $q \in \mathcal{P}(\mathcal{Z})$ against a strictly positive density $\tilde{\rho}$.

Theorem 3.1 (Large-batch unimodal consistency). *Consider the unimodal InfoNCE objective $\mathcal{L}_{\text{NCE},N}(\theta)$ in Eq. (2) and the parametric energy $\mathcal{J}_\tau(\theta)$ in Def. 3.2. Assume Asms. 2.1 and 3.1. Then for any fixed $\tau > 0$ and $\theta \in \Theta$, as $N \rightarrow \infty$,*

$$|\mathcal{L}_{\text{NCE},N}(\theta) - \mathcal{J}_\tau(\theta) - \log(NV_\kappa(\tau))| \rightarrow 0, \quad \|\nabla_\theta \mathcal{L}_{\text{NCE},N}(\theta) - \nabla_\theta \mathcal{J}_\tau(\theta)\| \rightarrow 0.$$

Thm. 3.1 shows that, at each fixed parameter value, the expected finite- N InfoNCE objective and its gradient converge to the deterministic energy $\mathcal{J}_\tau$ and its gradient, up to the additive constant $\log(NV_\kappa(\tau))$ in value. Thus, the partition field $\Gamma_{\theta,\tau}$, or equivalently the smoothed density $\tilde{\rho}_{\theta,\tau}$, determines the pointwise large-batch energy and gradient limit.

3.2 Intrinsic Free Energy and Gibbs Equilibrium

The parametric energy $\mathcal{J}_\tau(\theta)$ is generally nonconvex in θ , since the map $\theta \mapsto (q_\theta, \pi_{\theta\theta})$ is implicit. To isolate the intrinsic geometry, we fix a potential field U on $\mathcal{Z}$ and optimize directly over distributions on $\mathcal{Z}$. This yields a strictly convex free-energy functional with a unique Gibbs equilibrium. In the low-temperature regime, $\mathcal{J}_\tau$ approaches this intrinsic landscape up to a controllable KDE mismatch.

Definition 3.3 (Intrinsic unimodal energy). Fix $\tau > 0$ and a Borel measurable potential $U : \mathcal{Z} \rightarrow \mathbb{R}$ that is bounded below μ -a.e. Let

$$\mathcal{D}_\mu(\mathcal{Z}) := \left\{ \rho \geq 0 : \int_{\mathcal{Z}} \rho(z) \mu(dz) = 1 \right\}.$$

We define the intrinsic unimodal energy $\mathcal{F}_{\tau,U} : \mathcal{D}_\mu(\mathcal{Z}) \rightarrow (-\infty, +\infty]$ by

$$\mathcal{F}_{\tau,U}(\rho) := \frac{1}{\tau} \int_{\mathcal{Z}} U(z) \rho(z) \mu(dz) - H(\rho), \quad (7)$$

where $H(\rho) := - \int_{\mathcal{Z}} \rho(z) \log \rho(z) \mu(dz)$ is the differential entropy of μ -density ρ . We adopt $0 \log 0 := 0$, and set $\mathcal{F}_{\tau,U}(\rho) = +\infty$ if either $\int_{\mathcal{Z}} U(z) \rho(z) \mu(dz)$ or $\int_{\mathcal{Z}} \rho(z) \log \rho(z) \mu(dz)$ equals $+\infty$.

Proposition 3.1 (Unique Gibbs equilibrium). *Assume $\mathcal{Z}$ is compact with $\mu(\mathcal{Z}) < \infty$. Fix $\tau > 0$ and let $U : \mathcal{Z} \rightarrow \mathbb{R}$ be Borel measurable and bounded below μ -a.e. Then, $\mathcal{F}_{\tau,U}$ is strictly convex on its effective domain and has the unique minimizer*

$$\rho^*(z) = \frac{\exp(-U(z)/\tau)}{Z_\tau}, \quad Z_\tau := \int_{\mathcal{Z}} \exp\left(-\frac{U(z)}{\tau}\right) \mu(dz). \quad (8)$$

Proposition 3.2 (Low-temperature concentration). *In the setting of Prop. 3.1, fix $\sigma > 0$ and define*

$$\mathcal{W}^\sigma := \left\{ z \in \mathcal{Z} : U(z) \leq \text{ess inf}_\mu U + \sigma \right\}.$$

Then, as $\tau \rightarrow 0^+$, the Gibbs measure $\rho^*(z) \mu(dz)$ concentrates on $\mathcal{W}^\sigma$ in the sense that, for every open $\mathcal{O} \supset \mathcal{W}^\sigma$, $\int_{\mathcal{Z} \setminus \mathcal{O}} \rho^*(z) \mu(dz) \rightarrow 0$.

In the unimodal regime, the intrinsic landscape admits a unique Gibbs equilibrium for each τ , and as $\tau \rightarrow 0^+$, the equilibrium concentrates on near-minimizers of the alignment potential U . Prop. 3.2 formalizes this low-temperature concentration, while Fig. 3a illustrates it by tracking the mass inside a fixed near-minimizer region as τ decreases.

Connecting back to $\mathcal{J}_\tau$: sharp kernels control KDE mismatch. To connect the intrinsic geometry back to the parametric energy, we consider a sharp-kernel regime in which the kernel-smoothed density $\tilde{\rho}_{\theta,\tau}$ approximates the induced representation density ρ_θ .

Assumption 3.2 (Sharp diagonal peak). Assume [Asm. 2.1](#). Let $d_{\mathcal{Z}}(\cdot, \cdot)$ denote the geodesic distance on $\mathcal{Z}$. There exist constants $r > 0$ and $0 < m_1 \leq m_2 < \infty$ such that for all $z \in \mathcal{Z}$ and all $w \in \mathcal{Z}$ satisfying $d_{\mathcal{Z}}(z, w) < r$,

$$-m_2 d_{\mathcal{Z}}(z, w)^2 \leq \text{sim}(z, w) - \text{sim}(z, z) \leq -m_1 d_{\mathcal{Z}}(z, w)^2, \quad \text{and} \quad \arg \max_{w \in \mathcal{Z}} \text{sim}(z, w) = \{z\}, \quad z \in \mathcal{Z}. \quad (9)$$

Remark 3.2. Geometrically, [Asm. 3.2](#) forces the normalized kernel mass to concentrate within an $O(\sqrt{\tau})$ geodesic neighborhood of the diagonal; see [App. A.2](#) for discussion. As a result, $\tilde{\rho}_{\theta,\tau}$ acts as a local smoothing of ρ_θ , which is quantified in the low-temperature regime by [Lem. B.3](#).

Theorem 3.2 (Low-temperature unimodal energy consistency). Assume [Asms. 2.1, 3.1 and 3.2](#). Fix $\theta \in \Theta$ and assume $q_\theta \ll \mu$ with a strictly positive, continuous density ρ_θ , i.e., $\inf_{z \in \mathcal{Z}} \rho_\theta(z) \geq \underline{\rho}_\theta > 0$.³ Define the KDE mismatch $\varepsilon_{\text{kde}}^{(\theta)}(\tau) := \|\tilde{\rho}_{\theta,\tau} - \rho_\theta\|_\infty$. Then there exists $\tau_0(\theta) > 0$ such that

$$|\mathcal{J}_\tau(\theta) - \mathcal{F}_{\tau, U_\theta}(\rho_\theta)| \leq \frac{2\varepsilon_{\text{kde}}^{(\theta)}(\tau)}{\underline{\rho}_\theta}, \quad \forall 0 < \tau \leq \tau_0(\theta),$$

and in particular, as $\tau \rightarrow 0^+$, $|\mathcal{J}_\tau(\theta) - \mathcal{F}_{\tau, U_\theta}(\rho_\theta)| \rightarrow 0$.

Uniformity as a tie-breaker. [Thm. 3.2](#) bridges the intrinsic analysis back to the parametric objective: in the sharp-kernel regime, the parametric energy $\mathcal{J}_\tau(\theta)$ is close to the intrinsic energy $\mathcal{F}_{\tau, U_\theta}(\rho_\theta)$ up to the KDE mismatch between $\tilde{\rho}_{\theta,\tau}$ and ρ_θ . If the encoder family is expressive enough to realize the relevant intrinsic Gibbs equilibrium at the best attainable potential level, then minimizers of $\mathcal{J}_\tau$ inherit the same ground-state concentration behavior. Within an aligned basin where U_θ is already near-minimized, the remaining degree of freedom is how probability mass is distributed, and the entropy term in [Eq. \(7\)](#) selects the most internally dispersed near-optimal configuration. This is the sense in which “uniformity” emerges. [Cor. E.1](#) formalizes the expressivity condition and the resulting parametric inheritance; see [App. E.1](#) for extended formal statements.

☞ **Punchline 1** (Unimodal InfoNCE geometry). The defining feature of unimodal contrastive learning is structural cohesion. Because the modality evaluates its cross-entropy against its *own* smoothed density field, the intrinsic energy is strictly convex. Alignment determines the ground-state basin, while entropy selects a unique, internally dispersed equilibrium within it (see [Fig. 2a](#)).

4 Symmetric Multimodal InfoNCE: Divergence-Coupled Geometry

We analyze the symmetric multimodal InfoNCE objective used in CLIP-style training ([Eq. \(3\)](#)). Unlike the unimodal setting, the geometry is *cross-coupled*: each modality induces the effective field seen by the other. This coupling yields a qualitatively different geometry and can sustain a population-level modality gap. Proofs for all formal statements in this section are deferred to [App. D](#).

4.1 Cross-Coupled Multimodal Deterministic Energy

We now identify the deterministic population energy arising as the large-batch limit of symmetric multimodal InfoNCE. As in the unimodal case, the expected finite- N objective matches a closed-form parametric energy up to an additive large-batch constant in value, with pointwise gradient consistency. The difference is cross-coupling: each modality evaluates its cross-entropy against the other modality’s kernel-smoothed density field.

Assumption 4.1 (Multimodal optimization regularity). In multimodal contrastive learning, assume [Asm. 3.1](#) holds for both encoder families $\{f_\theta\}_{\theta \in \Theta}$ and $\{g_\phi\}_{\phi \in \Phi}$, uniformly over $\Theta \times \Phi$.

Remark 4.1. Under [Asm. 4.1](#), the symmetric InfoNCE objective has a uniformly bounded joint gradient over $\Theta \times \Phi$, which is needed to exchange the large-batch limit with differentiation in the multimodal consistency proof. See [Lem. B.2](#).

³See [App. A.3](#) for a detailed discussion on the strict positivity assumption and the role of density floors.

Definition 4.1 (Directional alignment potential fields). Let $\pi_{\theta\phi}$ denote the encoded positive-pair law on $\mathcal{Z} \times \mathcal{Z}$ with marginals q_θ and q_ϕ . By disintegration, there exist measurable families of probability measures $\{\nu_{\theta\phi,z}\}_{z \in \mathcal{Z}}$ and $\{\nu_{\phi\theta,w}\}_{w \in \mathcal{Z}}$ such that

$$\pi_{\theta\phi}(dz, dw) = q_\theta(dz) \nu_{\theta\phi,z}(dw) = q_\phi(dw) \nu_{\phi\theta,w}(dz),$$

as measures on $\mathcal{Z} \times \mathcal{Z}$. Define the directional alignment potential fields

$$U_{\theta \rightarrow \phi}(z) := - \int_{\mathcal{Z}} \text{sim}(z, w) \nu_{\theta\phi,z}(dw), \quad q_\theta\text{-a.e. } z, \quad U_{\phi \rightarrow \theta}(w) := - \int_{\mathcal{Z}} \text{sim}(z, w) \nu_{\phi\theta,w}(dz), \quad q_\phi\text{-a.e. } w.$$

Definition 4.2 (Symmetric multimodal parametric energy). Fix $\tau > 0$. Recall the kernel-smoothed μ -densities $\tilde{\rho}_{\theta,\tau}, \tilde{\rho}_{\phi,\tau}$ from [Def. 2.2](#). Define the directional multimodal energies by

$$\mathcal{J}_\tau^{x \rightarrow y}(\theta, \phi) := \frac{1}{\tau} \int_{\mathcal{Z}} U_{\theta \rightarrow \phi}(z) q_\theta(dz) - H_\times(q_\theta, \tilde{\rho}_{\phi,\tau}), \quad \mathcal{J}_\tau^{y \rightarrow x}(\phi, \theta) := \frac{1}{\tau} \int_{\mathcal{Z}} U_{\phi \rightarrow \theta}(w) q_\phi(dw) - H_\times(q_\phi, \tilde{\rho}_{\theta,\tau}).$$

and the symmetric multimodal parametric energy by

$$\mathcal{J}_\tau^{\text{Sym}}(\theta, \phi) := \frac{1}{2} (\mathcal{J}_\tau^{x \rightarrow y}(\theta, \phi) + \mathcal{J}_\tau^{y \rightarrow x}(\phi, \theta)). \quad (10)$$

Remark 4.2 (Directional asymmetry). Although both directions are induced by the same pair law $\pi_{\theta\phi}$, the two regular conditional laws $\nu_{\theta\phi,z}$ and $\nu_{\phi\theta,w}$ need not coincide. Consequently, $U_{\theta \rightarrow \phi}$ and $U_{\phi \rightarrow \theta}$ need not coincide: there is no single unimodal-style potential field unless an additional compatibility condition is imposed on the conditional laws.

Theorem 4.1 (Large-batch multimodal consistency). *Consider the symmetric multimodal InfoNCE objective $\mathcal{L}_{\text{Sym},N}(\theta, \phi)$ in [Eq. \(3\)](#) and the energy $\mathcal{J}_\tau^{\text{Sym}}(\theta, \phi)$ in [Def. 4.2](#). Assume [Asms. 2.1](#) and [4.1](#). Then, for any fixed $\tau > 0$ and $(\theta, \phi) \in \Theta \times \Phi$, as $N \rightarrow \infty$,*

$$|\mathcal{L}_{\text{Sym},N}(\theta, \phi) - \mathcal{J}_\tau^{\text{Sym}}(\theta, \phi) - \log(NV_\kappa(\tau))| \rightarrow 0, \quad \|\nabla_{(\theta,\phi)} \mathcal{L}_{\text{Sym},N}(\theta, \phi) - \nabla_{(\theta,\phi)} \mathcal{J}_\tau^{\text{Sym}}(\theta, \phi)\| \rightarrow 0.$$

Remark 4.3. Since the additive $\log(NV_\kappa(\tau))$ is constant in (θ, ϕ) , [Thm. 4.1](#) shows that, at each fixed parameter value, the expected finite- N symmetric InfoNCE objective and its gradient converge to $\mathcal{J}_\tau^{\text{Sym}}$ and its gradient, respectively, up to this additive constant in value. Unlike the unimodal case, the limiting energy is cross-coupled: each modality evaluates its cross-entropy against the other's smoothed density field.

4.2 Cross-Modal Divergence Coupling and Barrier-Driven Coordinate Geometry

We lift it to an intrinsic functional over modality *densities* on $\mathcal{Z}$. Relative to the unimodal free energy, the key new term is a *negative* symmetric divergence coupling. This coupling persists at fixed temperature and qualitatively changes the coordinate-wise geometry.

Definition 4.3 (Intrinsic multimodal energy). Fix $\tau > 0$ and let $U_{1 \rightarrow 2}, U_{2 \rightarrow 1} : \mathcal{Z} \rightarrow \mathbb{R}$ be fixed bounded Borel measurable potential fields. Fix density floors $\underline{\rho}_1, \underline{\rho}_2 > 0$ satisfying $\underline{\rho}_i \mu(\mathcal{Z}) < 1$, and define

$$\mathcal{D}_{\mu, \underline{\rho}_i}(\mathcal{Z}) := \left\{ \rho \in \mathcal{D}_\mu(\mathcal{Z}) \cap L^\infty(\mu) : \rho \geq \underline{\rho}_i \text{ } \mu\text{-a.e.} \right\}.$$

For $(\rho_1, \rho_2) \in \mathcal{D}_{\mu, \underline{\rho}_1}(\mathcal{Z}) \times \mathcal{D}_{\mu, \underline{\rho}_2}(\mathcal{Z})$, define the intrinsic multimodal energy by

$$\mathcal{F}_{\tau, \mathbf{U}_{1,2}}^{\text{Sym}}(\rho_1, \rho_2) := \frac{1}{2} (\mathcal{F}_{\tau, U_{1 \rightarrow 2}}(\rho_1) + \mathcal{F}_{\tau, U_{2 \rightarrow 1}}(\rho_2)) - D_{\text{KL}}^{\text{Sym}}(\rho_1, \rho_2), \quad (11)$$

where $\mathbf{U}_{1,2} := (U_{1 \rightarrow 2}, U_{2 \rightarrow 1})$, $\mathcal{F}_{\tau, U_{1 \rightarrow 2}}$ and $\mathcal{F}_{\tau, U_{2 \rightarrow 1}}$ are the unimodal free energies from [Eq. \(7\)](#), and

$$D_{\text{KL}}^{\text{Sym}}(\rho_1, \rho_2) := \frac{1}{2} (D_{\text{KL}}(\rho_1 \| \rho_2) + D_{\text{KL}}(\rho_2 \| \rho_1)), \quad \text{with} \quad D_{\text{KL}}(\rho_1 \| \rho_2) := \int_{\mathcal{Z}} \rho_1(z) \log \frac{\rho_1(z)}{\rho_2(z)} \mu(dz).$$

Remark 4.4. The constraint $\rho_i \geq \underline{\rho}_i$ is a technical device that rules out degenerate $-\infty$ energies and makes the coordinate-wise geometry explicit. See [App. A.3](#) for a discussion.

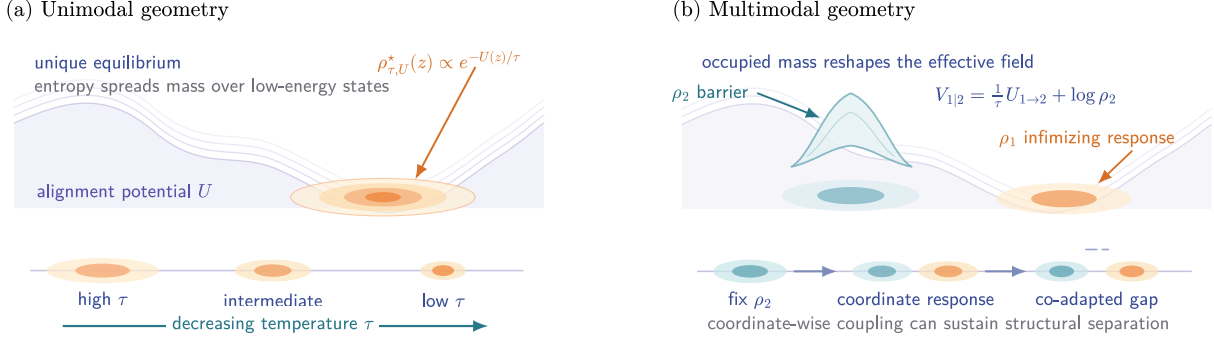


Figure 2: Schematic comparison of the intrinsic unimodal and multimodal geometries. **(a)** In the unimodal problem, alignment and entropy define a free energy with a unique Gibbs equilibrium, $\rho_{\tau,U}^* \propto e^{-U/\tau}$, whose mass increasingly concentrates near low-potential regions as $\tau \downarrow 0$. **(b)** In the multimodal problem, the peer density enters the effective potential through $V_{1|2} = U_{1 \rightarrow 2}/\tau + \log \rho_2$ (and symmetrically for $V_{2|1}$), so regions of high peer density act as barriers. Coordinate-wise minimizing sequences can therefore respond by concentrating mass in low-effective-potential regions away from the peer density, leading to barrier-driven structural separation rather than a single Gibbs equilibrium.

Proposition 4.1 (Barrier coordinate geometry and extremal convergence). *Assume $\mathcal{Z}$ is compact with $\mu(\mathcal{Z}) < \infty$, and let $\mathcal{F}_{\tau, \mathbf{U}_{1,2}}^{\text{Sym}}$ be as in Def. 4.3. Fix $\rho_2 \in \mathcal{D}_{\mu, \rho_2}(\mathcal{Z})$ and define*

$$V_{1|2}(z) := \frac{1}{\tau}U_{1 \rightarrow 2}(z) + \log \rho_2(z).$$

Then, the coordinate map $\rho_1 \mapsto \mathcal{F}_{\tau, \mathbf{U}_{1,2}}^{\text{Sym}}(\rho_1, \rho_2)$ is concave on $\mathcal{D}_{\mu, \rho_1}(\mathcal{Z})$, and its infimum is approached by densities that place the excess mass $M_{\text{ex}}^{(1)} := 1 - \rho_1 \mu(\mathcal{Z})$ on approximate minimizers of $V_{1|2}$ while leaving only the floor ρ_1 elsewhere. An analogous statement for $M_{\text{ex}}^{(2)} := 1 - \rho_2 \mu(\mathcal{Z})$ holds when optimizing in ρ_2 with ρ_1 fixed, with effective potential field $V_{2|1}(z) := \frac{1}{\tau}U_{2 \rightarrow 1}(z) + \log \rho_1(z)$. (The explicit extremal construction is given in App. D.2.)

Remark 4.5 (Barrier-driven coordinate geometry). Unlike the intrinsic unimodal Gibbs-like free energy, the multimodal functional does not admit Gibbs-type coordinate minimization. The term $\log \rho_2$ enters $V_{1|2} = \frac{1}{\tau}U_{1 \rightarrow 2} + \log \rho_2$ as an occupied-mass barrier: regions where modality 2 already assigns high density become more expensive for modality 1, whereas low-density regions of modality 2 are comparatively favored, subject to the alignment potential. Hence the coordinate infimum is approached by extremal rather than Gibbs-like densities: after maintaining the density floor, modality 1 can approach the infimum by concentrating its excess mass near minimizers of $V_{1|2}$, and symmetrically for modality 2. This coordinate-wise avoidance of occupied high-density regions, together with alignment-dependent basin selection, yields a winner-take-all-like extremal geometry.

Theorem 4.2 (Low-temperature multimodal energy consistency). *Assume Asms. 2.1, 3.2 and 4.1. Fix any $(\theta, \phi) \in \Theta \times \Phi$, and assume $q_\theta \ll \mu$, $q_\phi \ll \mu$ with continuous densities ρ_θ, ρ_ϕ bounded below: $\inf_{z \in \mathcal{Z}} \rho_\theta(z) \geq \underline{\rho}_\theta > 0$, $\inf_{z \in \mathcal{Z}} \rho_\phi(z) \geq \underline{\rho}_\phi > 0$. Let $\varepsilon_{\text{kde}}^{(\theta)}(\tau) := \|\tilde{\rho}_{\theta, \tau} - \rho_\theta\|_\infty$ and $\varepsilon_{\text{kde}}^{(\phi)}(\tau) := \|\tilde{\rho}_{\phi, \tau} - \rho_\phi\|_\infty$. Then, there exists $\tau_0(\theta, \phi) > 0$ such that for all $0 < \tau \leq \tau_0(\theta, \phi)$,*

$$\left| \mathcal{J}_\tau^{\text{Sym}}(\theta, \phi) - \mathcal{F}_{\tau, \mathbf{U}_{\theta, \phi}}^{\text{Sym}}(\rho_\theta, \rho_\phi) \right| \leq \frac{\varepsilon_{\text{kde}}^{(\theta)}(\tau) + \varepsilon_{\text{kde}}^{(\phi)}(\tau)}{\min\{\underline{\rho}_\theta, \underline{\rho}_\phi\}},$$

where $\mathcal{F}_{\tau, \mathbf{U}_{\theta, \phi}}^{\text{Sym}}(\rho_\theta, \rho_\phi)$ is the intrinsic multimodal energy from Def. 4.3, evaluated on $\mathcal{D}_{\mu, \rho_\theta}(\mathcal{Z}) \times \mathcal{D}_{\mu, \rho_\phi}(\mathcal{Z})$, with $\mathbf{U}_{\theta, \phi} := (U_{\theta \rightarrow \phi}, U_{\phi \rightarrow \theta})$. In particular,

$$\left| \mathcal{J}_\tau^{\text{Sym}}(\theta, \phi) - \mathcal{F}_{\tau, \mathbf{U}_{\theta, \phi}}^{\text{Sym}}(\rho_\theta, \rho_\phi) \right| \rightarrow 0, \quad \text{as } \tau \rightarrow 0^+.$$

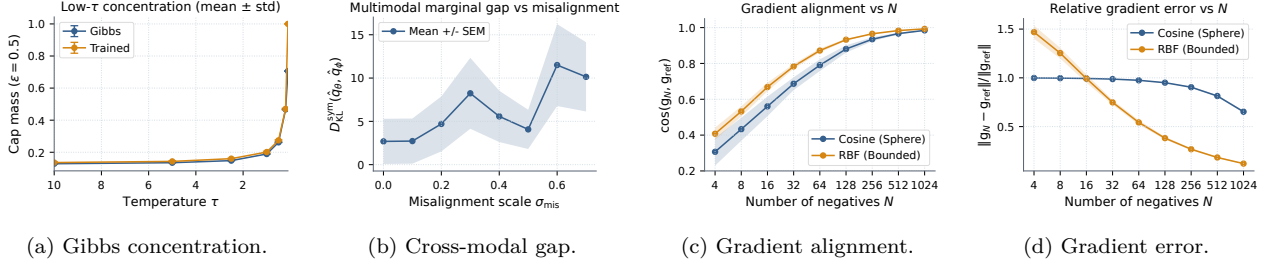


Figure 3: Numerical evaluation of population-geometric predictions and large-batch approximation. (a) In the unimodal setting, the Gibbs equilibrium concentrates increasingly on low-potential regions as $\tau \downarrow 0$. (b) In the multimodal setting, the two representation marginals remain separated, with the gap increasing under stronger cross-modal incompatibility. (c)–(d) Finite-negative InfoNCE gradients approach a large-batch reference as the number of negatives N increases: gradient alignment improves while relative gradient error decreases; Results for the cosine (spherical regime) and RBF (compact Euclidean regime) similarity functions, showing mean $\pm$ standard deviation over 20 seeds.

The limit $\tau \rightarrow 0^+$ controls only the KDE mismatches, driving $\tilde{\rho}_{\cdot,\tau} \rightarrow \rho_{\cdot}$; it does not remove the divergence term $D_{KL}^{Sym}(\rho_\theta, \rho_\phi)$. Because this term enters Def. 4.3 with a negative sign, it remains an explicit separation-inducing contribution as the KDE mismatch vanishes. Fig. 3b illustrates the associated marginal separation under increasing levels of cross-modal pairing misalignment.

Parametric inheritance under model expressivity. Combining Thm. 4.2 with the barrier coordinate geometry in Prop. 4.1 yields a coordinate-wise parametric inheritance statement: under a suitable model-expressivity condition and with one peer encoder fixed, minimizers of the parametric energy concentrate their excess mass near minimizers of the corresponding effective barrier potential. Cor. E.2 formalizes this condition and establishes the symmetric coordinate-wise inheritance result; see App. E.2 for the extended statements.

Punchline 2 (Multimodal InfoNCE geometry). Symmetric multimodal InfoNCE is pairwise attractive yet population-level repulsive. While individual cross-modal pairs align, the cross-coupled geometry can make each modality’s marginal act as a repulsive *population barrier* against the other. In the analyzed regime, this provides a natural geometric mechanism by which a modality gap can persist (see Fig. 2b).

5 Experiments

We test the theory at two levels. First, controlled numerical experiments isolate the core mechanisms predicted by the analysis: large-batch gradient consistency, unimodal Gibbs-type concentration, and structural modality gap under controlled cross-modal misalignment. Second, on MS-COCO (Lin et al., 2014), we test whether the same signatures appear in pretrained CLIP-like models, and whether weakening image–text compatibility systematically enlarges the representational modality gap.

Large-batch gradient consistency. We first test the large-batch consistency predicted by Thm. 3.1. On a synthetic Gaussian-mixture dataset, we compare finite-batch InfoNCE gradients with a high-fidelity large-batch reference under both a spherical cosine regime and a compact Euclidean RBF regime. We vary the number of negatives N and report two diagnostics: gradient alignment with the reference and relative gradient error. As shown in Figs. 3c and 3d, agreement with the large-batch reference improves steadily as N increases in both regimes: alignment rises, while relative error falls toward zero. This matches the consistency prediction of Thm. 3.1 and suggests that the large-batch gradient approximation can already be accurate at moderate N in these controlled settings. Full experimental details are deferred to App. F.1.

Unimodal equilibrium and low- τ concentration. We next test the intrinsic unimodal picture from § 3. On a smooth two-well potential on $\mathbb{S}^2$, we optimize a particle-based surrogate of $\mathcal{F}_{\tau,U}$ and compare the resulting particle distribution with the Gibbs law $\rho_\tau^* \propto \exp(-U/\tau)$ across temperatures. As shown in Fig. 4, the trained particles closely follow the Gibbs trend: at high temperature, both remain diffuse, while at low temperature, both concentrate near the wells. This visual agreement is quantitatively summarized by the cap-mass trend in Fig. 3a, which increases monotonically as τ decreases for both the Gibbs baseline and the trained particle

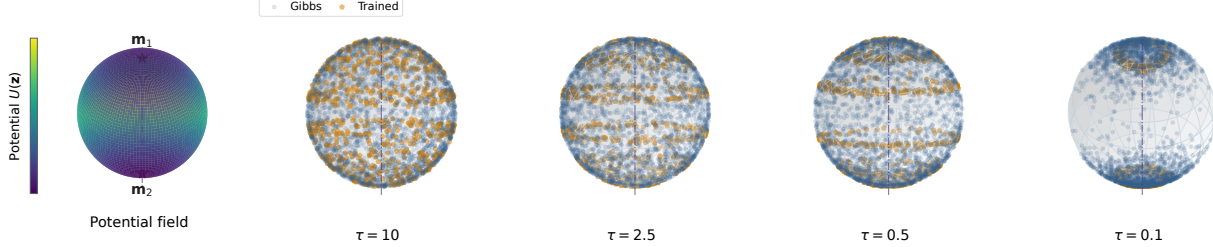


Figure 4: Unimodal potential landscape on $\mathbb{S}^2$ and distributions across τ . Left: the two-well potential U , with minima centers m_1 and m_2 marked. Right: Gibbs samples (blue) and trained particles (orange) for several temperatures τ . At high temperature, both distributions remain diffuse due to the stronger role of entropy; as τ decreases, both concentrate around the low-energy wells.

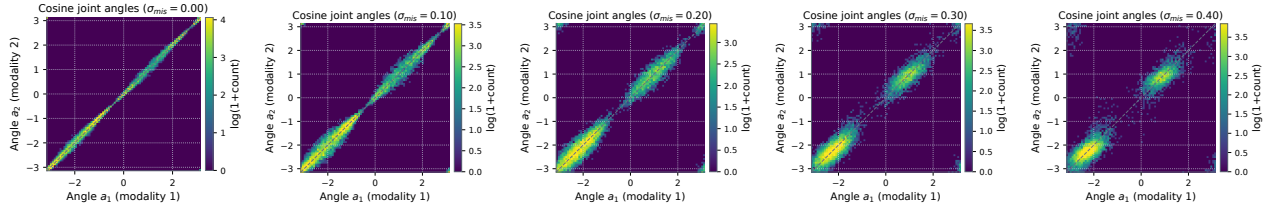


Figure 5: Joint-angle coupling under controlled misalignment. Each plot shows a histogram of joint angles (a_1, a_2) for $\sigma_{\text{mis}} \in \{0, 0.1, \dots, 0.4\}$. When $\sigma_{\text{mis}} = 0$, the coupling is sharply concentrated near the diagonal, indicating near-deterministic pairwise alignment. As $\sigma_{\text{mis}} \uparrow$, the diagonal band broadens and deforms, revealing a noisier cross-modal coupling while preserving the coarse latent modes.

system. Together, these results support the Gibbs-equilibrium and low-temperature concentration picture of the intrinsic theory. Further experimental details are deferred to [App. F.2](#).

Structural modality gap under controlled misalignment. We next test the central multimodal prediction from [§ 4](#) by examining how controlled conditional heterogeneity affects marginal coincidence. We construct a latent-angle experiment in which both modalities share the same underlying semantic variable, while one modality is perturbed by controlled angular misalignment noise σ_{mis} . We train symmetric CLIP-style linear encoders on $\mathbb{S}^1$ and examine both the learned joint coupling and the induced marginal discrepancy. As shown in [Fig. 5](#), increasing σ_{mis} progressively broadens and deforms the diagonal coupling structure, indicating weaker cross-modal alignment despite preserved coarse latent structure. This is accompanied by the increasing marginal discrepancy reported in [Fig. 3b](#). Together, these results are consistent with the theoretical picture that cross-modal heterogeneity can sustain a persistent structural modality gap. Further details are in [App. F.3](#).

MS-COCO studies with pretrained OpenCLIP models. We finally test whether the multimodal mechanism appears in realistic image-text representations on MS-COCO. In an observational study, we evaluate frozen OpenCLIP checkpoints on `val2017` using average Recall@1 together with two geometric diagnostics, the centroid gap $\|\mu_I - \mu_T\|$ and the energy distance between image and text marginals. As shown in [Fig. 6a](#), stronger retrieval does not necessarily imply a smaller modality gap: models with similar AvgR@1 can exhibit substantially different centroid and energy gaps. We then perform a controlled intervention on `train2017` by replacing each caption, with probability p , by a caption from another image sharing at least one COCO category, thereby preserving coarse semantic relatedness while weakening instance-level correspondence. In [Fig. 6b](#), increasing p consistently degrades retrieval and enlarges the centroid gap for both RN50 and ViT-B-16. Together, these results support the prediction that modality gap is not reducible to retrieval quality alone, and that weakening conditional compatibility systematically pushes the image and text marginals further apart. Further details are deferred to [App. F.4](#).

6 Related Work

Contrastive objectives and population geometry. InfoNCE is an instance of noise-contrastive density-ratio estimation, contrasting positives with negatives from a background marginal through softmax normalization

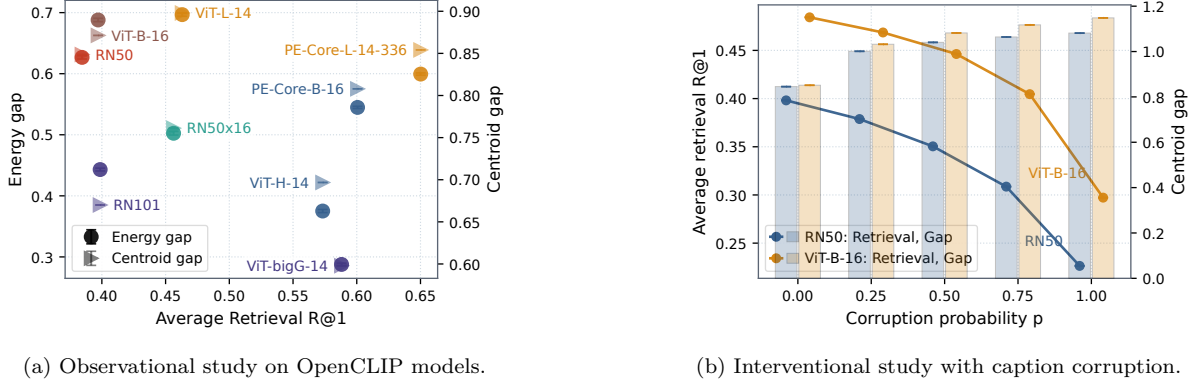


Figure 6: MS-COCO evaluation of the modality-gap mechanism. (a) On frozen OpenCLIP checkpoints, retrieval quality and cross-modal geometric agreement are related but not equivalent. (b) Controlled same-category corruption: Even a semantically plausible weakening of instance-level correspondence systematically enlarges the modality gap while degrading retrieval.

(Gutmann and Hyvärinen, 2010; Mnih and Teh, 2012; Oord et al., 2018). It is often used as a scalable surrogate for mutual-information criteria (Poole et al., 2019; Tschannen et al., 2020), though recent work argues it is better viewed as learning structured density ratios than as an accurate MI estimator (Ryu et al., 2026). A related lens studies population landscapes induced by contrastive objectives, notably through alignment–uniformity on hyperspherical embeddings (Wang and Isola, 2020; Wang and Liu, 2021; Chen et al., 2021; Betser et al., 2026); other analyses show such objectives can admit multiple, sometimes trivial, minimizers, with optimization implicitly selecting structured solutions (Calder and Lee, 2025). We instead prove large-batch gradient consistency and derive deterministic energy landscapes for representations on compact manifolds.

Training dynamics and learning theory. A distinct line of theory studies contrastive learning under stylized data or generative assumptions. Some analyses track feature or weight dynamics under imperfections such as misalignment, imbalance, or augmentation-induced variation, explaining suppression, robustness, or the effects of filtering and pruning (Sun et al., 2025; Liao et al., 2026). Others develop learning-theoretic guarantees under latent-class or multi-view models, and analyze how negative sampling and objective variants affect transfer (Saunshi et al., 2019, 2022; Chuang et al., 2020; Tian et al., 2020; HaoChen and Ma, 2023; Ji et al., 2023); recent CLIP-style analyses further characterize optimal similarity structure and motivate refined objectives (Uesaka et al., 2025; Yoshida et al., 2025). We do not track weight dynamics under a specific latent model, but characterize InfoNCE’s population geometry via large-batch energy limits and variational functionals.

Multimodal representation alignment and modality gap. Empirical work shows that CLIP-style alignment can coexist with a persistent modality gap (Liang et al., 2022; Shi et al., 2023; Levi and Gilboa, 2025; Betser et al., 2025), and recent methods seek to control it through adapters, alternative losses, or explicit regularizers. Yi et al. (2025) shows that unconstrained or cone-constrained optima close the gap, while dimension collapse into distinct hyperplanes leaves a gap equal to their smallest angle; Yin et al. (2026) augment InfoNCE with a kernel-estimated Cauchy–Schwarz divergence to align modality distributions, including unpaired settings; and pairwise sigmoid objectives likewise indicate that cross-modal coupling is sensitive to objective design (Zhai et al., 2023). We provide a geometric mechanism: symmetric InfoNCE induces cross-modal coupling in which each modality’s density shapes the other’s effective field via a persistent negative symmetric divergence term, providing an objective-level mechanism that can sustain a modality gap.

Identifiability in contrastive learning. A complementary line of work asks when contrastive objectives recover latent generative factors under data-generating assumptions. In nonlinear ICA, auxiliary variables such as temporal nonstationarity enable source recovery up to component-wise transformations (Hyvärinen and Morioka, 2016). For invertible-mixing models, global optima can invert the generator up to linear ambiguities, often orthogonal transformations on hyperspherical latents (Zimmermann et al., 2021; Liu et al., 2026). Multi-view models separating invariant content from view-varying style admit recovery of the invariant block up to an invertible mapping (Von Kügelgen et al., 2021; Cai et al., 2024); in multimodal settings with distinct

generative mechanisms, contrastive learning further block-identifies shared factors under nontrivial dependencies (Daunhawer et al., 2023). Broader causal analyses with partial observability and data-misalignment formulations also establish identifiability of shared semantics up to smooth bijections (Yao et al., 2024; Cai et al., 2025). We analyze objective-induced population geometry that standard identifiability analyses do not explain.

7 Conclusion

We developed a population-energy view of InfoNCE that exposes a geometric bifurcation in unimodal and symmetric multimodal contrastive learning. In the unimodal regime, intrinsic energy is convex and Gibbs-like, with entropy selecting an internally dispersed equilibrium within the aligned basin. In the multimodal regime, conditional heterogeneity gives rise to a negative symmetric divergence coupling that provides an explicit separation-inducing contribution and can sustain a persistent modality gap. These results show that pairwise alignment alone is not sufficient to control cross-modal marginal structure and motivate objectives that explicitly regulate cross-modal divergence. More broadly, our analysis suggests that multimodal contrastive learning should be understood not only through matched pairs, but through the coupled population geometry of representation marginals. This shift from pointwise matching to population geometry provides a framework for analyzing and designing future representation-learning objectives. See App. G for further discussion.

Acknowledgements

This project was partially supported by the Responsible AI Research (RAIR) Centre, Australia (Z. Zhang, Y. Liu, and J. Q. Shi). We thank the anonymous reviewers for their constructive feedback.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. While there are broad societal consequences associated with the general advancement of AI, our specific theoretical contributions do not present immediate societal impacts that require dedicated discussion.

References

- Roy Betser, Meir Yossef Levi, and Guy Gilboa. Whitened CLIP as a likelihood surrogate of images and captions. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=FIME06SV71>.
- Roy Betser, Eyal Gofer, Meir Yossef Levi, and Guy Gilboa. InfoNCE induces gaussian distribution. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=B1SH7gNQSq>.
- Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004. URL https://stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf.
- Yichao Cai, Yuhang Liu, Zhen Zhang, and Javen Qinfeng Shi. Clap: Isolating content from style through contrastive learning with augmented prompts. In *European Conference on Computer Vision (ECCV)*, pages 130–147, 2024. URL https://doi.org/10.1007/978-3-031-72664-4_8.
- Yichao Cai, Yuhang Liu, Erdun Gao, Tianjiao Jiang, Zhen Zhang, Anton van den Hengel, and Javen Qinfeng Shi. On the value of cross-modal misalignment in multimodal representation learning. In *Advances in Neural Information Processing Systems*, 2025. URL <https://openreview.net/pdf?id=3KtPuj0w5z>.
- Jeff Calder and Wonjun Lee. Understanding contrastive learning through variational analysis and neural network optimization perspectives. *arXiv preprint arXiv:2503.10812*, 2025. URL <https://arxiv.org/abs/2503.10812>.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. URL <https://proceedings.mlr.press/v119/chen20j.html>.
- Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. *Advances in Neural Information Processing Systems*, 34:11834–11845, 2021. URL <https://openreview.net/forum?id=rYhBGWYm6AU>.

- Ching-Yao Chuang, Joshua Robinson, Lin Yen-Chen, Antonio Torralba, and Stefanie Jegelka. Debaised contrastive learning. In *NeurIPS*, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/63c3ddcc7b23daa1e42dc41f9a44a873-Paper.pdf.
- Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E Vogt. Identifiability results for multimodal contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=U_2kuqoTcB.
- Stefan Elfving, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018. URL <https://doi.org/10.1016/j.neunet.2017.12.012>.
- Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9, pages 297–304. PMLR, 2010. URL <https://proceedings.mlr.press/v9/gutmann10a.html>.
- Jeff Z. HaoChen and Tengyu Ma. A theoretical study of inductive biases in contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=AuEgNLEAmed>.
- D Hendrycks. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. URL <https://arxiv.org/pdf/1606.08415>.
- Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016. URL <https://dl.acm.org/doi/abs/10.5555/3157382.3157519>.
- Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL <https://doi.org/10.5281/zenodo.5143773>.
- Wenlong Ji, Zhun Deng, Ryumei Nakada, James Zou, and Linjun Zhang. The power of contrast for feature learning: A theoretical analysis. *Journal of Machine Learning Research*, 24(330):1–78, 2023. URL <https://jmlr.org/papers/volume24/21-1501/21-1501.pdf>.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. URL <https://proceedings.mlr.press/v139/jia21b/jia21b.pdf>.
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, 2015. URL <https://arxiv.org/pdf/1412.6980>.
- Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819. PMLR, 2020. URL <https://proceedings.mlr.press/v119/lei20c/lei20c.pdf>.
- Meir Yossef Levi and Guy Gilboa. The double-ellipsoid geometry of CLIP. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=QGJ9u9B68Z>.
- Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022. URL <https://openreview.net/forum?id=S7Evzt9uit3>.
- Haixu Liao, Yating Zhou, Songyang Zhang, Meng Wang, and Shuai Zhang. Theoretical analysis of contrastive learning under imbalanced data: From training dynamics to a pruning solution. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=DUXG9E8dEO>.

- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. URL https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48.
- Yuhang Liu, Zhen Zhang, Dong Gong, Erdun Gao, Biwei Huang, Mingming Gong, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Beyond DAGs: A latent partial causal model for multimodal learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=bZqCBgm2N0>.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=Bkg6RiCqY7>.
- Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=B1QRgziT->.
- Andriy Mnih and Yee Whye Teh. A fast and simple algorithm for training neural probabilistic language models. In *29th International Conference on Machine Learning, ICML 2012*, 2012. URL <https://icml.cc/2012/papers/855.pdf>.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. URL <https://arxiv.org/abs/1807.03748>.
- Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 5171–5180. PMLR, 2019. URL <https://proceedings.mlr.press/v97/poole19a.html>.
- 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. URL <https://proceedings.mlr.press/v139/radford21a/radford21a.pdf>.
- Jongha Jon Ryu, Pavan Yeddanapudi, Xiangxiang Xu, and Gregory W. Wornell. Contrastive predictive coding done right for mutual information estimation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=JodkBXWgbA>.
- Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5628–5637, Long Beach, CA, USA, 2019. PMLR. URL <https://proceedings.mlr.press/v97/saunshi19a.html>.
- Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning*, pages 19250–19286. PMLR, 2022. URL <https://proceedings.mlr.press/v162/saunshi22a/saunshi22a.pdf>.
- Peiyang Shi, Michael C. Welle, Mårten Björkman, and Danica Kragic. Towards understanding the modality gap in CLIP. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*, 2023. URL <https://openreview.net/forum?id=8W3KGzw7fNI>.
- Jiawei Sun, Shuai Zhang, Hongkang Li, and Meng Wang. Contrastive learning with data misalignment: Feature purity, training dynamics and theoretical generalization guarantees. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=JM1YzOMDxI>.
- Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision*, pages 776–794. Springer, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123560749.pdf.

- M Tschannen, J Djolonga, P Rubenstein, S Gelly, and M Lucic. On mutual information maximization for representation learning. In *Eighth International Conference on Learning Representations*. OpenReview. net, 2020. URL <https://openreview.net/forum?id=rkxoh24FPH>.
- Toshimitsu Uesaka, Taiji Suzuki, Yuhta Takida, Chieh-Hsin Lai, Naoki Murata, and Yuki Mitsufuji. Weighted point set embedding for multimodal contrastive learning toward optimal similarity metric. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=uSz2K3ORRd>.
- Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021. URL https://openreview.net/forum?id=4pf_p0o0Dt.
- Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/papers/Wang_Understanding_the_Behaviour_of_Contrastive_Loss_CVPR_2021_paper.pdf.
- Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020. URL <https://dl.acm.org/doi/10.5555/3524938.3525859>.
- Dingling Yao, Danru Xu, Sebastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=OGtnhKQJms>.
- Lingjie Yi, Raphael Douady, and Chao Chen. Decipher the modality gap in multimodal contrastive learning: From convergent representations to pairwise alignment. *arXiv preprint arXiv:2510.03268*, 2025. URL <https://arxiv.org/pdf/2510.03268>.
- Wenzhe Yin, Zehao Xiao, Pan Zhou, Shujian Yu, Jiayi Shen, Jan-Jakob Sonke, and Stratis Gavves. Distributional vision-language alignment by cauchy-schwarz divergence. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=UUAjF4xLOe>.
- Naoki Yoshida, Satoshi Hayakawa, Yuhta Takida, Toshimitsu Uesaka, Hiromi Wakaki, and Yuki Mitsufuji. Theoretical refinement of clip by utilizing linear structure of optimal similarity. *arXiv preprint arXiv:2510.15508*, 2025. URL <https://arxiv.org/pdf/2510.15508>.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/papers/Zhai_Sigmoid_Loss_for_Language_Image_Pre-Training_ICCV_2023_paper.pdf.
- Guodong Zhang, Chaoqi Wang, Bowen Xu, and Roger Grosse. Three mechanisms of weight decay regularization. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=B1lz-3Rct7>.
- Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR, 2021. URL <https://proceedings.mlr.press/v139/zimmermann21a/zimmermann21a.pdf>.

Contents of Appendix

A Interpretation of Assumptions	18
A.1 On the Optimization Regularity Conditions	18
A.2 Interpretation of the Sharp Diagonal Peak Regime	19
A.3 On Density Floors and Positivity of the Effective Laws	19
B Lemmas	20
B.1 Uniform Gradient Bounds for InfoNCE Objectives	20
B.2 Sharp-peak Smoothing Implies Vanishing KDE Mismatch	24
C Proofs of Unimodal Formal Statements	26
C.1 Proof of Thm. 3.1	26
C.2 Proof of Prop. 3.1	30
C.3 Proof of Prop. 3.2	32
C.4 Proof of Thm. 3.2	33
D Proofs of Multimodal Formal Statements	35
D.1 Proof of Thm. 4.1	35
D.2 Explicit Extremal Construction and Proof of Prop. 4.1	39
D.3 Proof of Thm. 4.2	43
E Extended Formal Statements	45
E.1 Unimodal Parametric Inheritance and Emergence of Uniformity	45
E.2 Multimodal Parametric Inheritance	48
F Experimental Details	51
F.1 Large-Batch Gradient Consistency Across Similarity Functions	51
F.2 Unimodal Gibbs Equilibrium and Low-Temperature Concentration	52
F.3 Multimodal Structural Gap under Controlled Misalignment	54
F.4 MS-COCO Experiments: Pretrained Gap and Controlled Pair Corruption	57
G Discussion	60
G.1 Limitations and Future Directions	60
G.2 Broader Implications	60
H Changelog	61

Table 1: Core notations used throughout the paper.

Spaces, Measures, Encoders and Laws	
$\mathcal{X}, \mathcal{Y}$	Observation spaces; in the unimodal case both views lie in $\mathcal{X}$.
$\mathcal{Z} \subseteq \mathbb{R}^n$	Common representation space, equipped with a non-atomic Riemannian volume measure μ .
$\mathcal{P}(\mathcal{S})$	Set of probability measures on a measurable space $\mathcal{S}$.
$\mathcal{D}_\mu(\mathcal{Z})$	Set of probability densities on $\mathcal{Z}$ w.r.t. μ .
$r_{\text{um}} \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$	Unimodal positive-pair law governing $(\mathbf{x}, \tilde{\mathbf{x}})$.
$r_{\text{mm}} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$	Multimodal positive-pair law governing $(\mathbf{x}, \mathbf{y})$.
$p_{\mathbf{x}}, p_{\mathbf{y}}$	Marginal data laws on $\mathcal{X}$ and $\mathcal{Y}$, respectively.
Θ, Φ	Compact parameter spaces of the encoder families.
$f_\theta : \mathcal{X} \rightarrow \mathcal{Z}$	Encoder for the first modality.
$g_\phi : \mathcal{Y} \rightarrow \mathcal{Z}$	Encoder for the second modality.
$q_\theta := (f_\theta)_\# p_{\mathbf{x}}$	Encoded marginal law induced by f_θ .
$q_\phi := (g_\phi)_\# p_{\mathbf{y}}$	Encoded marginal law induced by g_ϕ .
$\pi_{\theta\theta} := (f_\theta \times f_\theta)_\# r_{\text{um}}$	Encoded unimodal positive-pair law.
$\pi_{\theta\phi} := (f_\theta \times g_\phi)_\# r_{\text{mm}}$	Encoded multimodal positive-pair law.
InfoNCE Objectives and Kernel Quantities	
$N \in \mathbb{N}_+$	Number of negatives in the InfoNCE denominator.
$\text{sim}(z, w)$	Symmetric similarity function on $\mathcal{Z} \times \mathcal{Z}$.
$\kappa_\tau(z, w) = \exp(\text{sim}(z, w)/\tau)$	Exponential similarity kernel at temperature $\tau > 0$.
$V_\kappa(\tau)$	Kernel volume constant under Asm. 2.1 .
$\mathcal{L}_{\text{NCE}, N}(\theta)$	Expected finite- N unimodal InfoNCE objective.
$\mathcal{L}_{\text{Sym}, N}(\theta, \phi)$	Expected finite- N symmetric multimodal InfoNCE objective.
$\mathcal{L}_{\text{NCE}, N}^{x \rightarrow y}(\theta, \phi), \mathcal{L}_{\text{NCE}, N}^{y \rightarrow x}(\phi, \theta)$	Expected finite- N directional multimodal InfoNCE objectives for $\mathbf{x} \rightarrow \mathbf{y}$ and $\mathbf{y} \rightarrow \mathbf{x}$.
$\Gamma_{\theta, \tau}, \Gamma_{\phi, \tau}$	Population partition fields at τ , induced by q_θ and q_ϕ , respectively.
$\tilde{\rho}_{\theta, \tau}, \tilde{\rho}_{\phi, \tau}$	Kernel-smoothed densities at τ , induced by q_θ and q_ϕ , respectively.
Unimodal Geometry	
$\nu_{\theta, z}$	Conditional law of the positive partner given encoded query z under $\pi_{\theta\theta}$.
$U_\theta : \mathcal{Z} \rightarrow \mathbb{R}$	Unimodal alignment potential field; equivalently, $U_\theta(z)$ is the negative conditional expected similarity at z .
$H_\times \text{sim}(q, \tilde{\rho})$	Cross-entropy of a probability measure q against a positive density $\tilde{\rho}$.
$H(\rho)$	Differential entropy of density ρ .
$\mathcal{J}_\tau(\theta)$	Unimodal parametric energy arising in the pointwise large-batch limit.
$\mathcal{F}_{\tau, U}(\rho)$	Intrinsic unimodal free-energy functional over densities ρ .
Multimodal Geometry	
$\nu_{\theta\phi, z}$	Conditional law of the ϕ -encoded partner given z under $\pi_{\theta\phi}$.
$\nu_{\phi\theta, w}$	Conditional law of the θ -encoded partner given w under $\pi_{\theta\phi}$.
$U_{\theta \rightarrow \phi}(z), U_{\phi \rightarrow \theta}(w)$	Directional alignment potentials seen by θ against ϕ , and conversely.
$\mathcal{J}_\tau^{x \rightarrow y}(\theta, \phi)$	Directional multimodal parametric energy for $x \rightarrow y$.
$\mathcal{J}_\tau^{y \rightarrow x}(\phi, \theta)$	Directional multimodal parametric energy for $y \rightarrow x$.
$\mathcal{J}_\tau^{\text{Sym}}(\theta, \phi)$	Symmetric multimodal parametric energy.
$\mathbf{U}_{1,2} := (U_{1 \rightarrow 2}, U_{2 \rightarrow 1})$	Pair of directional potentials in the intrinsic multimodal functional.
$\mathcal{F}_{\tau, \mathbf{U}_{1,2}}^{\text{Sym}}(\rho_1, \rho_2)$	Intrinsic symmetric multimodal energy.
$D_{\text{KL}}(\rho_1 \parallel \rho_2)$	Kullback–Leibler divergence.
$D_{\text{KL}}^{\text{Sym}}(\rho_1, \rho_2)$	Symmetric KL divergence, $\frac{1}{2}(D_{\text{KL}}(\rho_1 \parallel \rho_2) + D_{\text{KL}}(\rho_2 \parallel \rho_1))$.
$\underline{\rho}_i > 0$	Density floor used to define the feasible class $\mathcal{D}_{\mu, \underline{\rho}_i}(\mathcal{Z})$.
$\mathcal{D}_{\mu, \underline{\rho}}(\mathcal{Z})$	Floored density class $\{\rho \in \mathcal{D}_\mu(\mathcal{Z}) \cap L^\infty(\mu) : \rho \geq \underline{\rho} \text{ } \mu\text{-a.e.}\}$.
$V_{1 2}(z)$	Effective barrier potential for modality 1 given modality 2, $\frac{1}{\tau}U_{1 \rightarrow 2}(z) + \log \rho_2(z)$.
$V_{2 1}(z)$	Effective barrier potential for modality 2 given modality 1, $\frac{1}{\tau}U_{2 \rightarrow 1}(z) + \log \rho_1(z)$.

A Interpretation of Assumptions

A.1 On the Optimization Regularity Conditions

We argue that the regularity conditions in [Asm. 3.1](#) are compatible with standard practice in representation learning. We instantiate these conditions in two dominant geometric regimes: the spherical regime used in modern contrastive learning (e.g., SimCLR ([Chen et al., 2020](#)), CLIP ([Radford et al., 2021](#))) and the Euclidean regime common in classical kernel methods. Both admit the generalized exponential form $\kappa_\tau(z, w) := \exp(\text{sim}(z, w)/\tau)$.

- *Spherical regime (von Mises–Fisher kernel).* Representations are constrained to the hypersphere $\mathbb{S}^{d-1}$ via ℓ_2 normalization, and the similarity function is the cosine similarity $\text{sim}(z, w) = \langle z, w \rangle$:

$$\kappa_\tau^{(\text{vmf})}(z, w) := \exp\left(\frac{\langle z, w \rangle}{\tau}\right), \quad z, w \in \mathbb{S}^{n-1}, \quad (12)$$

- *Euclidean regime (RBF kernel).* The representations lie in a compact subset of $\mathbb{R}^n$, and the similarity function is the negative squared Euclidean distance $\text{sim}(z, w) = -\|z - w\|^2$:

$$\kappa_\tau^{(\text{rbf})}(z, w) := \exp\left(-\frac{\|z - w\|^2}{\tau}\right). \quad (13)$$

On a generic compact subset of $\mathbb{R}^n$, this example satisfies the boundedness and smoothness requirements used in [Asm. 3.1](#), but it does not, in general, satisfy the constant kernel-volume condition in [Asm. 2.1](#) because boundary effects make $V_\tau(z) := \int_{\mathcal{Z}} \kappa_\tau(z, w) \mu(dw)$ depend on z . The condition holds exactly on homogeneous boundaryless domains, such as a flat torus with periodic distance/measure, or after an appropriate boundary-corrected normalization. If [Asm. 2.1](#) is relaxed, the large-batch denominator still converges to the population partition field $\Gamma_{q, \tau}$; however, the clean identity $\Gamma_{q, \tau} = V_\kappa(\tau) \tilde{\rho}_{q, \tau}$ is replaced by a position-dependent normalization. This introduces additional boundary-volume correction terms, e.g., an effective potential contribution $\tau \log V_\tau(z)$, but does not change the large-batch consistency mechanism or the central population-field interpretation.

Encoder regularity. Condition [Asm. 3.1\(i\)](#) requires that for each $\mathbf{x}$, the map $\theta \mapsto f_\theta(\mathbf{x})$ is C^1 and that its parameter Jacobian is uniformly bounded: $\sup_{\theta \in \Theta} \sup_{\mathbf{x} \in \mathcal{X}} \|J_\theta f_\theta(\mathbf{x})\|_{\text{op}} < \infty$. For standard neural encoders built from differentiable primitives (e.g., linear maps, convolutions, normalization layers) and smooth activations (e.g., GELU ([Hendrycks, 2016](#)), SiLU ([Elfwing et al., 2018](#))), the map $\theta \mapsto f_\theta(\mathbf{x})$ is C^1 (or at least piecewise C^1 , which suffices for almost-everywhere analysis). In practice, uniform Jacobian control is promoted by explicit Lipschitz constraints (e.g., spectral normalization, which bounds layer operator norms and hence the network Jacobian via composition ([Miyato et al., 2018](#))) together with standard norm-regularizing training choices such as weight decay ([Loshchilov and Hutter, 2019](#); [Zhang et al., 2019](#)). Accordingly, we interpret compactness of Θ as a mild *localization* assumption: although the ambient parameter space is $\mathbb{R}^p$, successful training typically keeps iterates within a bounded region (often formalized as “SGD iterates belong to a ball” in stability analyses ([Lei and Ying, 2020](#))). One may therefore restrict the analysis to a sufficiently large compact subset Θ_R containing the training trajectory, or equivalently consider projected updates onto Θ_R as standard in constrained optimization ([Boyd and Vandenberghe, 2004](#)).

Similarity regularity. Condition [Asm. 3.1\(ii\)](#) requires that sim is C^1 on an open neighborhood of $\mathcal{Z} \times \mathcal{Z}$ and that $\sup_{(z, w) \in \mathcal{Z} \times \mathcal{Z}} \|\nabla \text{sim}(z, w)\| < \infty$. In the spherical regime, $\text{sim}(z, w) = \langle z, w \rangle$ is C^∞ on $\mathbb{R}^d \times \mathbb{R}^d$, with $\|\nabla_z \text{sim}(z, w)\| = \|w\|$ and $\|\nabla_w \text{sim}(z, w)\| = \|z\|$. Since $\|z\| = \|w\| = 1$, these are uniformly bounded by 1. In the Euclidean regime, $\text{sim}(z, w) = -\|z - w\|^2$ is also C^∞ with gradients scaling linearly with distance. Thus, provided $\mathcal{Z}$ is compact (e.g., via bounded activations like tanh or explicit clipping), $\|\nabla \text{sim}\|$ remains uniformly bounded on $\mathcal{Z} \times \mathcal{Z}$.

Boundedness of the kernel. For any fixed $\tau > 0$, the exponential kernel $\kappa_\tau(z, w) = \exp(\frac{\text{sim}(z, w)}{\tau})$ is strictly positive and bounded. For the spherical regime, $\langle z, w \rangle \in [-1, 1]$ implies $\kappa_\tau^{(\text{vmf})}(z, w) \in [\exp(-1/\tau), \exp(1/\tau)]$. For the RBF regime on a compact domain with diameter $D = \text{diam}(\mathcal{Z})$, we have $\kappa_\tau^{(\text{rbf})}(z, w) \in [\exp(-D^2/\tau), 1]$. Crucially, in both cases, the kernel is bounded away from zero for each fixed $\tau > 0$, ensuring that the logarithmic loss terms are well-defined and Lipschitz continuous.

A.2 Interpretation of the Sharp Diagonal Peak Regime

Geometric Interpretation. **Asm. 3.2** is a uniform non-degeneracy condition on the similarity function $\text{sim}(z, w)$ near the diagonal $w = z$. The inequalities in **Eq. (9)** state that for each anchor z , the function $w \mapsto \text{sim}(z, w)$ has an isolated maximizer at $w = z$ and admits a locally quadratic upper and lower envelope in geodesic distance. For the exponential kernel $\kappa_\tau(z, w) = \exp(\text{sim}(z, w)/\tau)$, this condition implies that the normalized kernel

$$\bar{\kappa}_\tau(z, w) := \frac{\kappa_\tau(z, w)}{\int_{\mathcal{Z}} \kappa_\tau(z, u) \mu(du)}$$

places asymptotically most of its mass within an $O(\sqrt{\tau})$ geodesic neighborhood of z as $\tau \downarrow 0$. Intuitively, $\bar{\kappa}(z, \cdot)$ behaves like a Gaussian bump of bandwidth $\sqrt{\tau}$ centered at z . As a result, the partition field $\Gamma_{\theta, \tau}(z) = \int_{\mathcal{Z}} \kappa_\tau(z, w) q_\theta(dw)$ probes only a local neighborhood of z when τ is small, and the smoothed density $\tilde{\rho}_{\theta, \tau} = \Gamma_{\theta, \tau}/V_\kappa(\tau)$ becomes a controlled local smoothing of ρ_θ .

Examples. For the RBF similarity $\text{sim}(z, w) = -\|z - w\|^2$, **Eq. (9)** holds with equality, so the sharp-diagonal condition in **Asm. 3.2** is satisfied locally. This statement concerns the local quadratic peak condition only. The separate constant-volume condition in **Asm. 2.1** holds exactly for the RBF kernel on homogeneous boundaryless geometries, such as a flat torus with periodic distance/measure, but not on a generic compact Euclidean subset without boundary correction. On the hypersphere with cosine similarity, $1 - \langle z, w \rangle \asymp d_{\mathbb{S}^{n-1}}(z, w)^2$ for nearby points, and the vMF kernel is isotropic on a compact homogeneous manifold, so both the local sharp-peak behavior and the constant-volume condition hold exactly.

A.3 On Density Floors and Positivity of the Effective Laws

Several parts of our analysis involve logarithms of representation densities (e.g., cross-entropies and KL-type terms). In the unimodal intrinsic problem, this is benign: no pointwise density floor is required, and the Gibbs minimizer is strictly positive whenever the potential is finite. In the *multimodal* intrinsic problem, however, the objective *subtracts* a symmetric KL-type divergence, creating an explicit “barrier” mechanism; without positivity constraints, the variational problem becomes ill-posed. This section explains why (i) the *parametric* large-batch objective is automatically interior due to smoothing, and (ii) the intrinsic density floor is a technically convenient proxy for that interior regime.

Smoothed densities. In the large-batch expansion, the quantities that appear are the kernel-smoothed μ -densities $\tilde{\rho}_{\theta, \tau}, \tilde{\rho}_{\phi, \tau}$ from **Def. 2.2**, not the raw pushforwards q_θ, q_ϕ . Because the exponential kernel satisfies $\kappa_\tau(z, w) = \exp(\text{sim}(z, w)/\tau) > 0$ pointwise, we have for every $\tau > 0$ and every probability measure q_θ ,

$$\tilde{\rho}_{\theta, \tau}(z) = \frac{1}{V_\kappa(\tau)} \int_{\mathcal{Z}} \kappa_\tau(z, w) q_\theta(dw) > 0, \quad z \in \mathcal{Z},$$

and analogously for $\tilde{\rho}_{\phi, \tau}$. Moreover, when κ_τ is continuous (as in our standing setting), $\tilde{\rho}_{\theta, \tau}$ is continuous on compact $\mathcal{Z}$ and therefore attains a strictly positive minimum

$$m_{\theta, \tau} := \min_{z \in \mathcal{Z}} \tilde{\rho}_{\theta, \tau}(z) > 0,$$

and similarly $m_{\phi, \tau} > 0$. Thus all log-terms appearing in the parametric energy are well-defined at finite temperature. Importantly, $m_{\theta, \tau}$ depends on (θ, τ) and we do *not* assume a uniform lower bound over Θ (or $\Theta \times \Phi$ in the multimodal case).

From raw parametric densities to smoothed densities. When we compare $\mathcal{J}_\tau(\theta)$ to intrinsic functionals expressed in terms of an unsmoothed density $\rho_\theta := dq_\theta/d\mu$ (when $q_\theta \ll \mu$), the relevant discrepancy is the KDE bias between $\tilde{\rho}_{\theta, \tau}$ and ρ_θ . Under the hypothesis used in **Thm. 3.2**, namely $\inf_{z \in \mathcal{Z}} \rho_\theta(z) \geq \underline{\rho}_\theta > 0$, whenever $\|\tilde{\rho}_{\theta, \tau} - \rho_\theta\|_\infty \leq \frac{1}{2}\underline{\rho}_\theta$, we have

$$\tilde{\rho}_{\theta, \tau}(z) \geq \frac{1}{2}\underline{\rho}_\theta, \quad z \in \mathcal{Z}.$$

The sharp-peak smoothing lemma (**Lem. B.3**) guarantees this condition for all sufficiently small τ under its stated hypotheses. Thus, in the low-temperature comparison the assumed floor on the raw density transfers to the corresponding smoothed density.

Density floor in multimodal analysis. In the multimodal formulation (Def. 4.3), we restrict (ρ_1, ρ_2) to

$$\mathcal{D}_{\mu, \underline{\rho}_i}(\mathcal{Z}) := \left\{ \rho \in \mathcal{D}_\mu(\mathcal{Z}) \cap L^\infty(\mu) : \rho \geq \underline{\rho}_i \text{ } \mu\text{-a.e.} \right\},$$

for fixed $\underline{\rho}_1, \underline{\rho}_2 > 0$ satisfying $\underline{\rho}_i \mu(\mathcal{Z}) < 1$, $i \in \{1, 2\}$. This floor serves two roles.

(i) *Well-posedness under negative symmetric divergence.* Because the multimodal functional contains a negative symmetric divergence coupling, allowing densities to vanish can make the objective unbounded below. Indeed, if $\rho_1 = 0$ on a set $\mathcal{A}$ with $\mu(\mathcal{A}) > 0$ while $\rho_2 > 0$ on $\mathcal{A}$, then $D_{\text{KL}}(\rho_2 \| \rho_1) = +\infty$ and hence the negative symmetric divergence term drives $\mathcal{F}_{\tau, \mathbf{U}_{1,2}}^{\text{Sym}}(\rho_1, \rho_2) = -\infty$. The floor $\rho_i \geq \underline{\rho}_i$ prevents this degeneracy and makes the coordinate variational problem well-defined. Equivalently, one may view the constraint as a hard version of the standard “interior” condition used to keep log-barrier objectives finite.

(ii) *Transparent coordinate geometry.* With the floor, we can decompose

$$\rho_1(z) = \underline{\rho}_1 + \eta_1(z), \quad \eta_1(z) \geq 0, \quad \int_{\mathcal{Z}} \eta_1(z) \mu(dz) = M_{\text{ex}}^{(1)} = 1 - \underline{\rho}_1 \mu(\mathcal{Z}),$$

and analogously for ρ_2 . This isolates a fixed “background” mass from an “excess” budget $M_{\text{ex}}^{(1)}$, making the extremal coordinate behavior in Prop. 4.1 quantitatively explicit. In this sense, the density floor is not an arbitrary restriction: it is the simplest device that (a) mirrors the intrinsic interiority of the parametrically smoothed quantities at finite τ , and (b) exposes the barrier-driven extremal coordinate mechanism in closed form.

B Lemmas

B.1 Uniform Gradient Bounds for InfoNCE Objectives

In this section, we prove uniform gradient bounds for unimodal and multimodal InfoNCE objectives, which are used in the large-batch gradient-consistency proofs of Thms. 3.1 and 4.1.

B.1.1 Unimodal Optimization Stability

Lemma B.1 (Uniform unimodal gradient bound). *Assume Asm. 3.1. Fix any temperature $\tau > 0$ and any batch size $N \in \mathbb{N}_+$. Then the unimodal InfoNCE objective defined in Eq. (2) has a uniformly bounded gradient over the parameter space Θ . Formally, there exists a constant $C(\tau) < \infty$ such that $\|\nabla_\theta \mathcal{L}_{\text{NCE}, N}(\theta)\| \leq C(\tau)$, uniformly for all $\theta \in \Theta$.*

Proof intuition. This proof establishes that the InfoNCE gradients cannot explode, providing a uniform upper bound that is strictly independent of the number of negative samples N . The intuition rests on two core mechanisms: **(i)** Because the representation manifold is compact and both the encoder and similarity functions are sufficiently smooth with bounded Jacobians, the chain rule guarantees that the gradient of any single pairwise similarity score is inherently bounded. **(ii)** The gradient of the log-partition denominator simplifies to a softmax-weighted sum of the pairwise gradients. Because this is a convex combination, the aggregated gradient norm cannot exceed the maximum norm of the individual bounded gradients. Consequently, increasing the number of negatives does not by itself cause the gradient bound to grow with N .

Proof. Step 1: Bounding components in the loss. Since $\mathcal{Z}$ is compact, the product $\mathcal{Z} \times \mathcal{Z}$ is compact. By Asm. 3.1(ii), sim is C^1 on an open neighborhood of $\mathcal{Z} \times \mathcal{Z}$, so ∇sim is continuous on $\mathcal{Z} \times \mathcal{Z}$ and hence attains its maximum norm on this compact set. For a fixed $\tau > 0$, define the temperature-dependent similarity sensitivity bound

$$C_{\text{sim}}(\tau) := \sup_{(z, w) \in \mathcal{Z} \times \mathcal{Z}} \left\| \nabla_{(z, w)} \left(\frac{1}{\tau} \text{sim}(z, w) \right) \right\| = \frac{1}{\tau} \sup_{(z, w) \in \mathcal{Z} \times \mathcal{Z}} \|\nabla \text{sim}(z, w)\| < \infty. \quad (14)$$

Next, by Asm. 3.1(i), the encoder Jacobian is uniformly bounded:

$$C_\Theta := \sup_{\theta \in \Theta} \sup_{x \in \mathcal{X}} \|J_\theta f_\theta(x)\|_{\text{op}} < \infty. \quad (15)$$

For any $x, x' \in \mathcal{X}$, write $z = f_\theta(x)$ and $w = f_\theta(x')$, and denote the corresponding pairwise parameter gradient by

$$\alpha_\theta(z, w) := \nabla_\theta \left(\frac{1}{\tau} \text{sim}(z, w) \right).$$

By the chain rule,

$$\alpha_\theta(z, w) = \nabla_z \left(\frac{1}{\tau} \text{sim}(z, w) \right)^\top J_\theta f_\theta(x) + \nabla_w \left(\frac{1}{\tau} \text{sim}(z, w) \right)^\top J_\theta f_\theta(x').$$

Therefore, by the triangle inequality and sub-multiplicativity,

$$\begin{aligned} \|\alpha_\theta(z, w)\| &\leq \left\| \nabla_z \frac{1}{\tau} \text{sim}(z, w) \right\| \|J_\theta f_\theta(x)\|_{\text{op}} + \left\| \nabla_w \frac{1}{\tau} \text{sim}(z, w) \right\| \|J_\theta f_\theta(x')\|_{\text{op}} \\ &\leq 2C_{\text{sim}}(\tau)C_\Theta. \end{aligned} \tag{16}$$

Thus, the parameter gradient of any pairwise similarity term is uniformly bounded.

Step 2: Gradient of the finite-batch partition function. Fix an anchor $z = f_\theta(x)$, a positive $v = f_\theta(\tilde{x})$, and negatives $\{w_j\}_{j=1}^N$. Define the contrastive set

$$\mathcal{C}(z) := \{v\} \cup \{w_j\}_{j=1}^N.$$

The gradient of the log-partition function

$$\log Z_{\mathcal{B}}(z) = \log \sum_{w \in \mathcal{C}(z)} \exp\left(\frac{\text{sim}(z, w)}{\tau}\right)$$

is given by the softmax-weighted sum of the inner gradients (see e.g. [Boyd and Vandenberghe, 2004](#), p. 74). Since $\kappa_\tau(z, w) = \exp(\text{sim}(z, w)/\tau) > 0$, we may differentiate $\log Z_{\mathcal{B}}$ and write

$$\nabla_\theta \log Z_{\mathcal{B}}(z) = \sum_{w \in \mathcal{C}(z)} \sigma_{\mathcal{B}}(w) \nabla_\theta \log \kappa_\tau(z, w) = \sum_{w \in \mathcal{C}(z)} \sigma_{\mathcal{B}}(w) \alpha_\theta(z, w), \tag{17}$$

where

$$\sigma_{\mathcal{B}}(w) := \frac{\kappa_\tau(z, w)}{Z_{\mathcal{B}}(z)} \quad \text{satisfies} \quad \sigma_{\mathcal{B}}(w) \geq 0, \quad \sum_{w \in \mathcal{C}(z)} \sigma_{\mathcal{B}}(w) = 1.$$

Therefore, using [Eq. \(16\)](#),

$$\begin{aligned} \|\nabla_\theta \log Z_{\mathcal{B}}(z)\| &\leq \sum_{w \in \mathcal{C}(z)} \sigma_{\mathcal{B}}(w) \|\alpha_\theta(z, w)\| \\ &\leq 2C_{\text{sim}}(\tau)C_\Theta \sum_{w \in \mathcal{C}(z)} \sigma_{\mathcal{B}}(w) \\ &= 2C_{\text{sim}}(\tau)C_\Theta. \end{aligned} \tag{18}$$

Step 3: Bounding the loss gradient. For a fixed batch $\mathcal{B}$ and anchor $z = f_\theta(x)$, define the per-batch loss

$$\ell_{\mathcal{B}}(\theta) := -\log \kappa_\tau(z, v) + \log Z_{\mathcal{B}}(z),$$

where $v = f_\theta(\tilde{x})$ is the positive match. Since $\nabla_\theta \log \kappa_\tau(z, v) = \alpha_\theta(z, v)$,

$$\nabla_\theta \ell_{\mathcal{B}}(\theta) = -\alpha_\theta(z, v) + \nabla_\theta \log Z_{\mathcal{B}}(z).$$

By the triangle inequality and [Eqs. \(16\) and \(18\)](#),

$$\|\nabla_\theta \ell_{\mathcal{B}}(\theta)\| \leq \|\alpha_\theta(z, v)\| + \|\nabla_\theta \log Z_{\mathcal{B}}(z)\| \leq 4C_{\text{sim}}(\tau)C_\Theta. \tag{19}$$

Finally, the expected finite- N objective is

$$\mathcal{L}_{\text{NCE}, N}(\theta) = \mathbb{E}_{\mathcal{B}}[\ell_{\mathcal{B}}(\theta)].$$

Since $\ell_{\mathcal{B}}$ is differentiable in θ by [Asm. 3.1](#) for each $\mathcal{B}$ and $\|\nabla_{\theta}\ell_{\mathcal{B}}(\theta)\| \leq 4C_{\text{sim}}(\tau)C_{\Theta}$ provides an integrable dominating function, we may differentiate under the expectation, obtaining

$$\nabla_{\theta}\mathcal{L}_{\text{NCE},N}(\theta) = \mathbb{E}_{\mathcal{B}}[\nabla_{\theta}\ell_{\mathcal{B}}(\theta)].$$

Taking norms and applying Jensen's inequality yields

$$\|\nabla_{\theta}\mathcal{L}_{\text{NCE},N}(\theta)\| \leq \mathbb{E}_{\mathcal{B}}[\|\nabla_{\theta}\ell_{\mathcal{B}}(\theta)\|] \leq 4C_{\text{sim}}(\tau)C_{\Theta}.$$

Thus, for each fixed $\tau > 0$ and any $N \in \mathbb{N}_+$, the gradient of the unimodal InfoNCE objective defined in [Eq. \(2\)](#) is uniformly bounded over Θ by $C(\tau) := 4C_{\text{sim}}(\tau)C_{\Theta}$. $\square$

B.1.2 Multimodal Optimization Stability

Lemma B.2 (Uniform multimodal gradient bound). *Assume [Asm. 4.1](#). Fix any $\tau > 0$ and any $N \in \mathbb{N}_+$. Then the symmetric multimodal InfoNCE objective $\mathcal{L}_{\text{Sym},N}(\theta, \phi)$ defined in [Eq. \(3\)](#) has a uniformly bounded joint gradient over $\Theta \times \Phi$. Formally, there exists a constant $C(\tau) < \infty$ such that*

$$\|\nabla_{(\theta, \phi)}\mathcal{L}_{\text{Sym},N}(\theta, \phi)\| \leq C(\tau), \quad \forall (\theta, \phi) \in \Theta \times \Phi.$$

Proof intuition. The proof extends the unimodal bound to the joint parameter space (Θ, Φ) . For either direction, the softmax-weighted denominator is a convex combination of uniformly bounded pairwise gradients, so the partial gradients with respect to both encoder parameters remain uniformly bounded independently of N . The joint gradient is obtained by concatenating the two parameter blocks, and averaging the two directions preserves a uniform bound.

Proof. Step 1: Uniform bounds on components. We retain the similarity sensitivity bound $C_{\text{sim}}(\tau)$ from [Eq. \(14\)](#). For the encoders, define

$$C_{\Theta} := \sup_{\theta \in \Theta} \sup_{x \in \mathcal{X}} \|J_{\theta}f_{\theta}(x)\|_{\text{op}} < \infty, \quad C_{\Phi} := \sup_{\phi \in \Phi} \sup_{y \in \mathcal{Y}} \|J_{\phi}g_{\phi}(y)\|_{\text{op}} < \infty,$$

which are finite by [Asm. 4.1](#). Define the global encoder constant

$$C_{\text{enc}} := \max\{C_{\Theta}, C_{\Phi}\}.$$

Consider a generic cross-modal representation pair $z = f_{\theta}(x)$ and $w = g_{\phi}(y)$. The gradient of $\log \kappa_{\tau}(z, w)$ with respect to the joint parameter vector (θ, ϕ) splits into block components. By the chain rule and [Eq. \(14\)](#), the partial gradients satisfy

$$\|\nabla_{\theta} \log \kappa_{\tau}(z, w)\| \leq C_{\text{sim}}(\tau)C_{\Theta} \leq C_{\text{sim}}(\tau)C_{\text{enc}}, \quad \|\nabla_{\phi} \log \kappa_{\tau}(z, w)\| \leq C_{\text{sim}}(\tau)C_{\Phi} \leq C_{\text{sim}}(\tau)C_{\text{enc}}. \quad (20)$$

Step 2: Bounding directional partial derivatives. Consider a fixed batch realization $\mathcal{B}$ consisting of a positive pair (x, y) and N negative keys $\{y'_j\}_{j=1}^N$, with the analogous construction in the reverse direction. The symmetric per-batch loss is the average of the two directional losses:

$$\ell_{\text{Sym}}(\mathcal{B}) = \frac{1}{2}(\ell_{\theta \rightarrow \phi}(\mathcal{B}) + \ell_{\phi \rightarrow \theta}(\mathcal{B})).$$

By the triangle inequality,

$$\|\nabla_{(\theta, \phi)}\ell_{\text{Sym}}\| \leq \frac{1}{2}\|\nabla_{(\theta, \phi)}\ell_{\theta \rightarrow \phi}\| + \frac{1}{2}\|\nabla_{(\theta, \phi)}\ell_{\phi \rightarrow \theta}\|.$$

We first analyze the forward component $\ell_{\theta \rightarrow \phi}$. Here, x serves as the anchor, y as the positive key, and $\{y'_j\}_{j=1}^N$ as the negative keys:

$$\ell_{\theta \rightarrow \phi} := -\log \kappa_{\tau}(f_{\theta}(x), g_{\phi}(y)) + \log \left(\kappa_{\tau}(f_{\theta}(x), g_{\phi}(y)) + \sum_{j=1}^N \kappa_{\tau}(f_{\theta}(x), g_{\phi}(y'_j)) \right).$$

The joint gradient vector is the concatenation

$$\nabla_{(\theta, \phi)}\ell_{\theta \rightarrow \phi} = (\nabla_{\theta}\ell_{\theta \rightarrow \phi}^{\top}, \nabla_{\phi}\ell_{\theta \rightarrow \phi}^{\top})^{\top}.$$

We bound the two parameter blocks separately. Let

$$\mathcal{C}_\phi := \{g_\phi(y)\} \cup \{g_\phi(y'_j)\}_{j=1}^N,$$

and denote the softmax weight assigned to any $v \in \mathcal{C}_\phi$ by $\sigma(v)$.

(i) *Gradient with respect to θ (anchor sensitivity).* In this direction, θ governs only the anchor $f_\theta(x)$; the keys in $\mathcal{C}_\phi$ are fixed with respect to θ . Thus,

$$\nabla_\theta \ell_{\theta \rightarrow \phi} = -\nabla_\theta \log \kappa_\tau(f_\theta(x), g_\phi(y)) + \sum_{v \in \mathcal{C}_\phi} \sigma(v) \nabla_\theta \log \kappa_\tau(f_\theta(x), v).$$

Applying the triangle inequality and the uniform sensitivity bound in Eq. (20),

$$\begin{aligned} \|\nabla_\theta \ell_{\theta \rightarrow \phi}\| &\leq \|\nabla_\theta \log \kappa_\tau(f_\theta(x), g_\phi(y))\| + \sum_{v \in \mathcal{C}_\phi} \sigma(v) \|\nabla_\theta \log \kappa_\tau(f_\theta(x), v)\| \\ &\leq C_{\text{sim}}(\tau) C_\Theta + C_{\text{sim}}(\tau) C_\Theta \sum_{v \in \mathcal{C}_\phi} \sigma(v) \\ &= 2C_{\text{sim}}(\tau) C_\Theta, \end{aligned}$$

since $\sum_{v \in \mathcal{C}_\phi} \sigma(v) = 1$.

(ii) *Gradient with respect to ϕ (key sensitivity).* In this direction, ϕ parametrizes every key in the batch, while the anchor $f_\theta(x)$ is fixed with respect to ϕ . Hence,

$$\nabla_\phi \ell_{\theta \rightarrow \phi} = -\nabla_\phi \log \kappa_\tau(f_\theta(x), g_\phi(y)) + \sum_{v \in \mathcal{C}_\phi} \sigma(v) \nabla_\phi \log \kappa_\tau(f_\theta(x), v).$$

Similarly,

$$\begin{aligned} \|\nabla_\phi \ell_{\theta \rightarrow \phi}\| &\leq C_{\text{sim}}(\tau) C_\Phi + C_{\text{sim}}(\tau) C_\Phi \sum_{v \in \mathcal{C}_\phi} \sigma(v) \\ &= 2C_{\text{sim}}(\tau) C_\Phi. \end{aligned}$$

Step 3: Global uniform bound. Combining the two parameter blocks gives

$$\begin{aligned} \|\nabla_{(\theta, \phi)} \ell_{\theta \rightarrow \phi}\| &= \sqrt{\|\nabla_\theta \ell_{\theta \rightarrow \phi}\|^2 + \|\nabla_\phi \ell_{\theta \rightarrow \phi}\|^2} \\ &\leq \sqrt{(2C_{\text{sim}}(\tau) C_\Theta)^2 + (2C_{\text{sim}}(\tau) C_\Phi)^2} \\ &\leq 2\sqrt{2} C_{\text{sim}}(\tau) C_{\text{enc}}. \end{aligned}$$

By symmetry, the same bound applies to the backward loss $\ell_{\phi \rightarrow \theta}$. Therefore,

$$\begin{aligned} \|\nabla_{(\theta, \phi)} \ell_{\text{Sym}}\| &= \frac{1}{2} \|\nabla_{(\theta, \phi)} \ell_{\theta \rightarrow \phi} + \nabla_{(\theta, \phi)} \ell_{\phi \rightarrow \theta}\| \\ &\leq \frac{1}{2} (\|\nabla_{(\theta, \phi)} \ell_{\theta \rightarrow \phi}\| + \|\nabla_{(\theta, \phi)} \ell_{\phi \rightarrow \theta}\|) \\ &\leq 2\sqrt{2} C_{\text{sim}}(\tau) C_{\text{enc}}. \end{aligned}$$

Finally,

$$\mathcal{L}_{\text{Sym}, N}(\theta, \phi) = \mathbb{E}_{\mathcal{B}}[\ell_{\text{Sym}}(\mathcal{B})].$$

Since the per-batch gradient is uniformly bounded, we may differentiate under the expectation, obtaining

$$\nabla_{(\theta, \phi)} \mathcal{L}_{\text{Sym}, N}(\theta, \phi) = \mathbb{E}_{\mathcal{B}}[\nabla_{(\theta, \phi)} \ell_{\text{Sym}}(\mathcal{B})].$$

Hence, by Jensen's inequality,

$$\begin{aligned} \|\nabla_{(\theta, \phi)} \mathcal{L}_{\text{Sym}, N}(\theta, \phi)\| &\leq \mathbb{E}_{\mathcal{B}}[\|\nabla_{(\theta, \phi)} \ell_{\text{Sym}}(\mathcal{B})\|] \\ &\leq 2\sqrt{2} C_{\text{sim}}(\tau) C_{\text{enc}}. \end{aligned}$$

Thus, with

$$C(\tau) := 2\sqrt{2} C_{\text{sim}}(\tau) C_{\text{enc}} < \infty,$$

the gradient is uniformly bounded over $\Theta \times \Phi$. $\square$

B.2 Sharp-peak Smoothing Implies Vanishing KDE Mismatch

The large-batch parametric energy $\mathcal{J}_\tau(\theta)$ depends on the *kernel-smoothed* density $\tilde{\rho}_{\theta,\tau}$ through the cross-entropy term $H_\times(q_\theta, \tilde{\rho}_{\theta,\tau})$, whereas the intrinsic functional $\mathcal{F}_{\tau,U_\theta}$ is expressed in terms of the *true* density ρ_θ and its entropy $H(\rho_\theta)$. To compare $\mathcal{J}_\tau(\theta)$ with $\mathcal{F}_{\tau,U_\theta}(\rho_\theta)$ in [Thm. 3.2](#), we therefore require quantitative control of the bias introduced by smoothing, i.e., that $\tilde{\rho}_{\theta,\tau}$ approximates ρ_θ when the kernel is sufficiently sharp.

Lemma B.3 (Vanishing KDE error under sharp peak). *Assume [Asms. 2.1](#), [3.1](#) and [3.2](#). Fix $\theta \in \Theta$ and assume $q_\theta \ll \mu$ with a continuous density ρ_θ bounded away from zero, i.e.,*

$$\inf_{z \in \mathcal{Z}} \rho_\theta(z) > \underline{\rho}_\theta > 0.$$

Define the KDE mismatch

$$\varepsilon_{\text{kde}}^{(\theta)}(\tau) := \|\tilde{\rho}_{\theta,\tau} - \rho_\theta\|_\infty,$$

where $\tilde{\rho}_{\theta,\tau}$ is the kernel-smoothed density from [Def. 2.2](#). Then, as $\tau \rightarrow 0^+$,

$$\varepsilon_{\text{kde}}^{(\theta)}(\tau) \rightarrow 0.$$

☞ **Proof intuition.** The goal is to show that as $\tau \rightarrow 0^+$, the kernel-smoothed density $\tilde{\rho}_{\theta,\tau}$ converges uniformly to the continuous density ρ_θ over $\mathcal{Z}$. The intuition relies on two properties. (i) The similarity function has a sharp, unique peak along the diagonal. As $\tau \rightarrow 0^+$, the normalized exponential kernel increasingly concentrates its mass near $w = z$, while contributions from points separated from z decay exponentially. (ii) Since ρ_θ is continuous on compact $\mathcal{Z}$, it is uniformly continuous. Thus, within the shrinking neighborhood on which the kernel concentrates, $\rho_\theta(w)$ is uniformly close to $\rho_\theta(z)$, while the contribution from the complement vanishes. Together, these properties imply uniform convergence of the smoothed density to ρ_θ .

Proof. Fix any $\theta \in \Theta$. Let $d_{\mathcal{Z}}(\cdot, \cdot)$ denote the geodesic distance on the compact Riemannian manifold $\mathcal{Z}$ as in [Asm. 3.2](#). Since $\mathcal{Z}$ is compact and ρ_θ is continuous, ρ_θ is bounded. Define the normalized kernel

$$\bar{\kappa}_\tau(z, w) := \frac{\kappa_\tau(z, w)}{V_\kappa(\tau)}.$$

We first show that $\bar{\kappa}_\tau$ concentrates near the diagonal $\{z = w\}$ as $\tau \rightarrow 0^+$, and then establish the claimed sup-norm convergence.

Step 1: Uniform concentration of $\bar{\kappa}_\tau$ near the diagonal. Fix any $\delta \in (0, r)$, where r is from [Asm. 3.2](#), and define

$$\mathcal{A}_\delta := \{(z, w) \in \mathcal{Z} \times \mathcal{Z} : d_{\mathcal{Z}}(z, w) \geq \delta\}.$$

By compactness of $\mathcal{Z}$, the set $\mathcal{A}_\delta$ is compact. Since sim is continuous by [Asm. 3.1](#), consider

$$g(z, w) := \text{sim}(z, z) - \text{sim}(z, w).$$

By [Asm. 3.2](#), for each fixed $z \in \mathcal{Z}$ the map $w \mapsto \text{sim}(z, w)$ has a unique maximizer at $w = z$; hence $g(z, w) > 0$ whenever $d_{\mathcal{Z}}(z, w) \geq \delta$. Therefore, g is strictly positive on $\mathcal{A}_\delta$, and by compactness it attains a strictly positive minimum:

$$m_\delta := \min_{(z, w) \in \mathcal{A}_\delta} g(z, w) > 0.$$

Equivalently, for all $z, w \in \mathcal{Z}$ satisfying $d_{\mathcal{Z}}(z, w) \geq \delta$,

$$\text{sim}(z, w) \leq \text{sim}(z, z) - m_\delta. \tag{21}$$

Fix $\eta \in (0, \delta)$. Let

$$\mathbb{B}_l(z) := \{w \in \mathcal{Z} : d_{\mathcal{Z}}(z, w) < l\}$$

denote an open geodesic ball of radius l . By compactness of $\mathcal{Z}$, the open cover $\{\mathbb{B}_{\eta/2}(z) : z \in \mathcal{Z}\}$ admits a finite subcover. Thus, there exist points $\{z_1, \dots, z_M\} \subset \mathcal{Z}$ such that

$$\mathcal{Z} = \bigcup_{j=1}^M \mathbb{B}_{\eta/2}(z_j).$$

For any $z \in \mathcal{Z}$, choose $i \in \{1, \dots, M\}$ such that $z \in \mathbb{B}_{\eta/2}(z_i)$. For any $w \in \mathbb{B}_{\eta/2}(z_i)$, the triangle inequality gives

$$d_{\mathcal{Z}}(z, w) \leq d_{\mathcal{Z}}(z_i, w) + d_{\mathcal{Z}}(z_i, z) < \eta.$$

Hence

$$\mathbb{B}_{\eta/2}(z_i) \subseteq \mathbb{B}_{\eta}(z).$$

By monotonicity of μ ,

$$\mu(\mathbb{B}_{\eta}(z)) \geq \mu(\mathbb{B}_{\eta/2}(z_i)) \geq \min_{1 \leq j \leq M} \mu(\mathbb{B}_{\eta/2}(z_j)) =: b_{\eta} > 0,$$

where positivity follows from the full support of the Riemannian volume measure μ .

Now choose $\eta \in (0, \delta)$ sufficiently small that

$$m_2 \eta^2 < m_{\delta}, \quad (22)$$

where m_2 is the constant in [Asm. 3.2](#). For $w \in \mathbb{B}_{\eta}(z) \subset \mathbb{B}_r(z)$, the lower quadratic bound in [Asm. 3.2](#) gives

$$\text{sim}(z, w) \geq \text{sim}(z, z) - m_2 d_{\mathcal{Z}}(z, w)^2 \geq \text{sim}(z, z) - m_2 \eta^2.$$

Therefore,

$$\kappa_{\tau}(z, w) \geq \exp(\text{sim}(z, z)/\tau) \exp(-m_2 \eta^2/\tau).$$

Integrating over $\mathbb{B}_{\eta}(z)$ and using $\mu(\mathbb{B}_{\eta}(z)) \geq b_{\eta}$ yields

$$V_{\kappa}(\tau) = \int_{\mathcal{Z}} \kappa_{\tau}(z, w) \mu(dw) \geq b_{\eta} \exp(\text{sim}(z, z)/\tau) \exp(-m_2 \eta^2/\tau). \quad (23)$$

Here, by [Asm. 2.1](#), $V_{\kappa}(\tau)$ is independent of z . The factor $\exp(\text{sim}(z, z)/\tau)$ cancels in the ratio below.

Next, for $w \notin \mathbb{B}_{\delta}(z)$ we have $d_{\mathcal{Z}}(z, w) \geq \delta$, so by [Eq. \(21\)](#),

$$\kappa_{\tau}(z, w) = \exp(\text{sim}(z, w)/\tau) \leq \exp(\text{sim}(z, z)/\tau) \exp(-m_{\delta}/\tau).$$

Therefore, with $V_{\mu} := \mu(\mathcal{Z}) < \infty$,

$$\int_{\mathcal{Z} \setminus \mathbb{B}_{\delta}(z)} \kappa_{\tau}(z, w) \mu(dw) \leq V_{\mu} \exp(\text{sim}(z, z)/\tau) \exp(-m_{\delta}/\tau). \quad (24)$$

Dividing [Eq. \(24\)](#) by [Eq. \(23\)](#) and using [Eq. \(22\)](#) gives, uniformly in z ,

$$\int_{\mathcal{Z} \setminus \mathbb{B}_{\delta}(z)} \bar{\kappa}_{\tau}(z, w) \mu(dw) \leq \frac{V_{\mu}}{b_{\eta}} \exp\left(-\frac{m_{\delta} - m_2 \eta^2}{\tau}\right) \xrightarrow{\tau \rightarrow 0^+} 0. \quad (25)$$

Step 2: Sup-norm convergence of the smoothing. Fix $\varepsilon > 0$. Since ρ_{θ} is continuous on compact $\mathcal{Z}$, it is uniformly continuous. Hence, for fixed $\theta \in \Theta$, choose $\delta \in (0, r)$ such that

$$|\rho_{\theta}(w) - \rho_{\theta}(z)| \leq \varepsilon \quad \text{whenever } d_{\mathcal{Z}}(z, w) < \delta.$$

For any $z \in \mathcal{Z}$,

$$\tilde{\rho}_{\theta, \tau}(z) - \rho_{\theta}(z) = \int_{\mathcal{Z}} \bar{\kappa}_{\tau}(z, w) (\rho_{\theta}(w) - \rho_{\theta}(z)) \mu(dw),$$

and we split the integral over $\mathbb{B}_{\delta}(z)$ and its complement. Using

$$\int_{\mathcal{Z}} \bar{\kappa}_{\tau}(z, w) \mu(dw) = 1,$$

the bound on $\mathbb{B}_{\delta}(z)$, and

$$|\rho_{\theta}(w) - \rho_{\theta}(z)| \leq 2\|\rho_{\theta}\|_{\infty}$$

everywhere, we obtain

$$|\tilde{\rho}_{\theta,\tau}(z) - \rho_{\theta}(z)| \leq \varepsilon + 2\|\rho_{\theta}\|_{\infty} \int_{\mathcal{Z} \setminus \mathbb{B}_{\delta}(z)} \bar{\kappa}_{\tau}(z, w) \mu(dw).$$

Taking the supremum over $z \in \mathcal{Z}$ and invoking [Eq. \(25\)](#) gives

$$\limsup_{\tau \rightarrow 0^+} \|\tilde{\rho}_{\theta,\tau} - \rho_{\theta}\|_{\infty} \leq \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary,

$$\varepsilon_{\text{kde}}^{(\theta)}(\tau) := \|\tilde{\rho}_{\theta,\tau} - \rho_{\theta}\|_{\infty} \xrightarrow{\tau \rightarrow 0^+} 0.$$

□

C Proofs of Unimodal Formal Statements

This section provides rigorous proofs for all formal statements concerning our unimodal analysis. Each statement from the main text is restated for self-containment, accompanied by a proof intuition for providing high-level insights.

C.1 Proof of [Thm. 3.1](#)

Theorem 3.1 (Large-batch unimodal consistency). *Consider the unimodal InfoNCE objective $\mathcal{L}_{\text{NCE},N}(\theta)$ in [Eq. \(2\)](#) and the parametric energy $\mathcal{J}_{\tau}(\theta)$ in [Def. 3.2](#). Assume [Asms. 2.1](#) and [3.1](#). Then for any fixed $\tau > 0$ and $\theta \in \Theta$, as $N \rightarrow \infty$,*

$$|\mathcal{L}_{\text{NCE},N}(\theta) - \mathcal{J}_{\tau}(\theta) - \log(NV_{\kappa}(\tau))| \rightarrow 0, \quad \|\nabla_{\theta} \mathcal{L}_{\text{NCE},N}(\theta) - \nabla_{\theta} \mathcal{J}_{\tau}(\theta)\| \rightarrow 0.$$

☞ **Proof intuition.** This proof establishes that, for each fixed parameter θ and temperature $\tau > 0$, the expected finite- N InfoNCE objective and its gradient converge, up to the additive value constant $\log(NV_{\kappa}(\tau))$, to the deterministic parametric energy $\mathcal{J}_{\tau}(\theta)$ and its gradient as $N \rightarrow \infty$. The intuition has two parts. **(i)** In a finite batch, the positive and negative samples compete inside the denominator. As $N \rightarrow \infty$, the contribution of the single positive term becomes negligible relative to the negative sum, while the normalized negative sum converges to the population kernel average $\Gamma_{\theta,\tau}$. This yields the alignment and cross-entropy terms appearing in $\mathcal{J}_{\tau}$. **(ii)** The same argument applies to the gradient of the log-partition term: the positive contribution vanishes at rate $O(N^{-1})$, while the ratio formed by the negative-sum gradient and the negative partition sum converges to the gradient of the population log-partition field. Uniform gradient bounds then allow this convergence to pass to the expected finite- N objective.

Proof. Consider a random batch $\mathcal{B}$ consisting of a positive pair and N independent negative samples. Let $(\mathbf{x}, \tilde{\mathbf{x}}) \sim r_{\text{um}}$, and let the negative samples $\{\mathbf{x}'_j\}_{j=1}^N$ be drawn i.i.d. from the marginal observation distribution $p_{\mathbf{x}}$, so that

$$\mathcal{B} := \{(\mathbf{x}, \tilde{\mathbf{x}}), \mathbf{x}'_1, \dots, \mathbf{x}'_N\}.$$

Fix any encoder parameter $\theta \in \Theta$, and denote the corresponding random representations by $\mathbf{z} = f_{\theta}(\mathbf{x})$, $\mathbf{v} = f_{\theta}(\tilde{\mathbf{x}})$, and $\mathbf{w}_j = f_{\theta}(\mathbf{x}'_j)$. The corresponding random per-batch unimodal InfoNCE loss is

$$\ell_{\mathcal{B}}(\theta; \mathbf{x}) := -\frac{1}{\tau} \text{sim}(\mathbf{z}, \mathbf{v}) + \log \underbrace{\left(\kappa_{\tau}(\mathbf{z}, \mathbf{v}) + \sum_{j=1}^N \kappa_{\tau}(\mathbf{z}, \mathbf{w}_j) \right)}_{=: \mathbf{Z}_{\mathcal{B}}(\mathbf{z})}, \quad (26)$$

where $\kappa_{\tau}(z, w) = \exp(\text{sim}(z, w)/\tau)$.

Part I. Value consistency. We analyze the asymptotic behavior of the loss as $N \rightarrow \infty$.

Step 1: Isolate the positive bias. First, decompose the partition sum $\mathbf{Z}_{\mathcal{B}}$ inside the logarithm into the positive term and the sum over negatives:

$$\mathbf{Z}_{\mathcal{B}} = \mathbf{P} + \mathbf{S}_N, \quad \mathbf{P} := \kappa_{\tau}(\mathbf{z}, \mathbf{v}), \quad \mathbf{S}_N := \sum_{j=1}^N \kappa_{\tau}(\mathbf{z}, \mathbf{w}_j).$$

The logarithmic term becomes

$$\log Z_{\mathcal{B}} = \log(S_N + P) = \log S_N + \underbrace{\log\left(1 + \frac{P}{S_N}\right)}_{=: \epsilon_N}.$$

Fix $\tau > 0$. Since $\mathcal{Z}$ is compact and sim is continuous on $\mathcal{Z} \times \mathcal{Z}$ by [Asm. 3.1\(ii\)](#), there exist finite constants $\text{sim}_{\min}$ and $\text{sim}_{\max}$ such that $\text{sim}_{\min} \leq \text{sim}(z, w) \leq \text{sim}_{\max}$ for all $(z, w) \in \mathcal{Z} \times \mathcal{Z}$. Consequently,

$$0 < m_{\tau} := \exp(\text{sim}_{\min}/\tau) \leq \kappa_{\tau}(z, w) \leq \exp(\text{sim}_{\max}/\tau) =: M_{\tau} < \infty.$$

Thus, $m_{\tau} \leq P \leq M_{\tau}$ and $Nm_{\tau} \leq S_N \leq NM_{\tau}$. Using $\log(1+x) \leq x$ for $x \geq 0$,

$$0 < \epsilon_N = \log\left(1 + \frac{P}{S_N}\right) \leq \frac{P}{S_N} \leq \frac{M_{\tau}}{Nm_{\tau}} = O_{\tau}(N^{-1}).$$

Therefore, for each fixed $\tau > 0$,

$$\log Z_{\mathcal{B}} = \log S_N + O_{\tau}(N^{-1}).$$

Step 2: Concentration of the negative partition sum. Fix θ and a realization x of the anchor observation, and write $z = f_{\theta}(x)$. For negatives $\mathbf{x}'_j \stackrel{\text{i.i.d.}}{\sim} p_{\mathbf{x}}$, define $\mathbf{w}_j := f_{\theta}(\mathbf{x}'_j)$ and

$$Y_j(z) := \kappa_{\tau}(z, \mathbf{w}_j), \quad S_N(z) := \sum_{j=1}^N Y_j(z).$$

For fixed z , the variables $\{Y_j(z)\}_{j=1}^N$ are i.i.d. with mean

$$\mu_{\kappa}(z) := \mathbb{E}_{\mathbf{w} \sim q_{\theta}}[\kappa_{\tau}(z, \mathbf{w})] = \int_{\mathcal{Z}} \kappa_{\tau}(z, w) q_{\theta}(\mathrm{d}w) \equiv \Gamma_{\theta, \tau}(z) \geq m_{\tau},$$

and finite variance. Moreover, for each fixed $\tau > 0$, κ_{τ} is bounded on $\mathcal{Z} \times \mathcal{Z}$, so $Y_j(z)$ is bounded. By Hoeffding's inequality,

$$\frac{S_N(z)}{N} - \mu_{\kappa}(z) = O_p(N^{-1/2}).$$

Equivalently, writing

$$\frac{S_N(z)}{N\mu_{\kappa}(z)} = 1 + \Delta_N(z), \quad \Delta_N(z) = O_p(N^{-1/2}),$$

and noting $\mu_{\kappa}(z) \geq m_{\tau} > 0$, we have

$$\log S_N(z) = \log(N\mu_{\kappa}(z)) + \log(1 + \Delta_N(z)) = \log N + \log \Gamma_{\theta, \tau}(z) + O_p(N^{-1/2}).$$

Therefore,

$$\log S_N(z) = \log N + \log \Gamma_{\theta, \tau}(z) + O_p(N^{-1/2}).$$

Step 3: Identification of the induced parametric energy. Combining Steps 1–2 gives

$$\log Z_{\mathcal{B}}(\mathbf{z}) = \log N + \log \Gamma_{\theta, \tau}(\mathbf{z}) + o_p(1),$$

where $o_p(1)$ denotes a term vanishing in probability as $N \rightarrow \infty$ for fixed $\tau > 0$. Substituting into [Eq. \(26\)](#) yields

$$\ell_{\mathcal{B}}(\theta; \mathbf{x}) = -\frac{1}{\tau} \text{sim}(\mathbf{z}, \mathbf{v}) + \log N + \log \Gamma_{\theta, \tau}(\mathbf{z}) + o_p(1). \quad (27)$$

Moreover, for fixed $\tau > 0$, the remainder in the logarithmic partition term is uniformly bounded in N : indeed, both S_N/N and $\Gamma_{\theta, \tau}(\mathbf{z})$ lie in $[m_{\tau}, M_{\tau}]$, while the positive-term correction is $O_{\tau}(N^{-1})$. Hence the remainder is uniformly integrable. Since it converges to zero in probability, it also converges to zero in L^1 , and therefore its expectation is $o(1)$.

Taking expectations over the batch construction $(\mathbf{x}, \tilde{\mathbf{x}}) \sim r_{\text{um}}$ and the i.i.d. negatives, and using $\mathcal{L}_{\text{NCE},N}(\theta) = \mathbb{E}_{\mathcal{B}}[\ell_{\mathcal{B}}(\theta; \mathbf{x})]$, we obtain

$$\mathcal{L}_{\text{NCE},N}(\theta) = \mathbb{E}_{(\mathbf{x}, \tilde{\mathbf{x}})} \left[-\frac{1}{\tau} \text{sim}(f_{\theta}(\mathbf{x}), f_{\theta}(\tilde{\mathbf{x}})) \right] + \mathbb{E}_{\mathbf{x}} [\log \Gamma_{\theta,\tau}(f_{\theta}(\mathbf{x}))] + \log N + o(1), \quad (28)$$

where $o(1) \rightarrow 0$ as $N \rightarrow \infty$.

By disintegration of $\pi_{\theta\theta} = (f_{\theta} \times f_{\theta})_{\#} r_{\text{um}}$, define

$$U_{\theta}(z) := - \int_{\mathcal{Z}} \text{sim}(z, w) \nu_{\theta,z}(\mathrm{d}w).$$

Then the law of iterated expectations gives

$$\mathbb{E}_{(\mathbf{x}, \tilde{\mathbf{x}})} \left[-\frac{1}{\tau} \text{sim}(\mathbf{z}, \mathbf{v}) \right] = \frac{1}{\tau} \mathbb{E}_{\mathbf{z} \sim q_{\theta}} [U_{\theta}(\mathbf{z})] = \frac{1}{\tau} \int_{\mathcal{Z}} U_{\theta}(z) q_{\theta}(\mathrm{d}z). \quad (29)$$

Since $\mathbf{z} = f_{\theta}(\mathbf{x}) \sim q_{\theta}$,

$$\mathbb{E}_{\mathbf{x}} [\log \Gamma_{\theta,\tau}(f_{\theta}(\mathbf{x}))] = \mathbb{E}_{\mathbf{z} \sim q_{\theta}} [\log \Gamma_{\theta,\tau}(\mathbf{z})].$$

Under the constant kernel-volume condition [Asm. 2.1](#),

$$\tilde{\rho}_{\theta,\tau}(z) = \frac{\Gamma_{\theta,\tau}(z)}{V_{\kappa}(\tau)}$$

by [Def. 2.2](#), hence

$$\mathbb{E}_{\mathbf{z} \sim q_{\theta}} [\log \Gamma_{\theta,\tau}(\mathbf{z})] = \log V_{\kappa}(\tau) + \mathbb{E}_{\mathbf{z} \sim q_{\theta}} [\log \tilde{\rho}_{\theta,\tau}(\mathbf{z})] = \log V_{\kappa}(\tau) - H_{\times}(q_{\theta}, \tilde{\rho}_{\theta,\tau}), \quad (30)$$

where

$$H_{\times}(q_{\theta}, \tilde{\rho}_{\theta,\tau}) := -\mathbb{E}_{\mathbf{z} \sim q_{\theta}} [\log \tilde{\rho}_{\theta,\tau}(\mathbf{z})].$$

Substituting [Eqs. \(29\)](#) and [\(30\)](#) into [Eq. \(28\)](#) yields

$$\begin{aligned} \mathcal{L}_{\text{NCE},N}(\theta) &= \frac{1}{\tau} \int_{\mathcal{Z}} U_{\theta}(z) q_{\theta}(\mathrm{d}z) - H_{\times}(q_{\theta}, \tilde{\rho}_{\theta,\tau}) + \log(NV_{\kappa}(\tau)) + o(1) \\ &= \mathcal{J}_{\tau}(\theta) + \log(NV_{\kappa}(\tau)) + o(1). \end{aligned}$$

Since $\theta \in \Theta$ was arbitrary, this proves

$$|\mathcal{L}_{\text{NCE},N}(\theta) - \mathcal{J}_{\tau}(\theta) - \log(NV_{\kappa}(\tau))| \longrightarrow 0$$

for each fixed $\theta \in \Theta$ and $\tau > 0$.

Part II. Gradient consistency. We now analyze the gradient of the batch loss $\ell_{\mathcal{B}}(\theta; \mathbf{x})$ for fixed $\tau > 0$ as $N \rightarrow \infty$. Differentiating the finite-batch loss gives

$$\nabla_{\theta} \ell_{\mathcal{B}}(\theta; \mathbf{x}) = -\frac{1}{\tau} \nabla_{\theta} \text{sim}(\mathbf{z}, \mathbf{v}) + \nabla_{\theta} \log Z_{\mathcal{B}}(\mathbf{z}), \quad Z_{\mathcal{B}}(\mathbf{z}) = \mathbf{P} + \mathbf{S}_N, \quad (31)$$

where

$$\mathbf{P} := \kappa_{\tau}(\mathbf{z}, \mathbf{v}), \quad \mathbf{S}_N := \sum_{j=1}^N \kappa_{\tau}(\mathbf{z}, \mathbf{w}_j).$$

Step 1: Reducing the positive term in the log-partition gradient. Let

$$\mathbf{g}_N(\mathbf{z}) := \nabla_{\theta} \log(\mathbf{P} + \mathbf{S}_N), \quad \mathbf{r}_N(\mathbf{z}) := \nabla_{\theta} \log \mathbf{S}_N.$$

Then

$$\begin{aligned}
 \mathbf{g}_N - \mathbf{r}_N &= \nabla_\theta \log(\mathbf{P} + \mathbf{S}_N) - \nabla_\theta \log \mathbf{S}_N \\
 &= \frac{\nabla_\theta \mathbf{P} + \nabla_\theta \mathbf{S}_N}{\mathbf{P} + \mathbf{S}_N} - \frac{\nabla_\theta \mathbf{S}_N}{\mathbf{S}_N} \\
 &= \frac{\nabla_\theta \mathbf{P}}{\mathbf{P} + \mathbf{S}_N} + \nabla_\theta \mathbf{S}_N \left(\frac{\mathbf{S}_N - (\mathbf{P} + \mathbf{S}_N)}{\mathbf{S}_N(\mathbf{P} + \mathbf{S}_N)} \right) \\
 &= \frac{\nabla_\theta \mathbf{P}}{\mathbf{P} + \mathbf{S}_N} - \frac{\mathbf{P}}{\mathbf{S}_N(\mathbf{P} + \mathbf{S}_N)} \nabla_\theta \mathbf{S}_N.
 \end{aligned} \tag{32}$$

Under [Asm. 3.1](#), continuity of sim on the compact domain $\mathcal{Z} \times \mathcal{Z}$ ensures that κ_τ is uniformly bounded. Moreover, expanding $\nabla_\theta \kappa_\tau$ by the chain rule yields a product of the kernel, similarity gradients, and encoder Jacobians, all uniformly bounded by assumption. Thus, for each fixed $\tau > 0$, there exist finite constants M_τ, G_τ such that $0 < \kappa_\tau \leq M_\tau$ and $\|\nabla_\theta \kappa_\tau\| \leq G_\tau$. Since κ_τ is continuous and strictly positive on compact $\mathcal{Z} \times \mathcal{Z}$, it also admits the positive lower bound $m_\tau > 0$. Hence

$$\mathbf{P} \leq M_\tau, \quad \|\nabla_\theta \mathbf{P}\| \leq G_\tau, \quad \mathbf{S}_N \geq Nm_\tau, \quad \|\nabla_\theta \mathbf{S}_N\| \leq NG_\tau.$$

Taking norms in [Eq. \(32\)](#) and using these bounds gives

$$\|\mathbf{g}_N - \mathbf{r}_N\| \leq \frac{G_\tau}{Nm_\tau} + \frac{M_\tau}{(Nm_\tau)^2} NG_\tau = O_\tau(N^{-1}).$$

Therefore,

$$\nabla_\theta \log Z_{\mathcal{B}}(\mathbf{z}) = \nabla_\theta \log \mathbf{S}_N(\mathbf{z}) + O_\tau(N^{-1}). \tag{33}$$

Step 2: Consistency of the negative-sum ratio. Condition on the event that the anchor observation takes the realization x , i.e., $\mathbf{x} = x$, and write $z = f_\theta(x)$. Then

$$\nabla_\theta \log \mathbf{S}_N(z) = \frac{\frac{1}{N} \sum_{j=1}^N \nabla_\theta \kappa_\tau(f_\theta(x), f_\theta(\mathbf{x}'_j))}{\frac{1}{N} \sum_{j=1}^N \kappa_\tau(f_\theta(x), f_\theta(\mathbf{x}'_j))}.$$

Given x , the negatives $\mathbf{x}'_j$ are i.i.d. with law $p_{\mathbf{x}}$, and both $\kappa_\tau(f_\theta(x), f_\theta(\mathbf{x}'_j))$ and its total parameter gradient are uniformly bounded for fixed $\tau > 0$ by [Asm. 3.1](#). Hence, by the law of large numbers,

$$\frac{1}{N} \sum_{j=1}^N \kappa_\tau(f_\theta(x), f_\theta(\mathbf{x}'_j)) \xrightarrow{p} \Gamma_{\theta, \tau}(f_\theta(x)),$$

and

$$\frac{1}{N} \sum_{j=1}^N \nabla_\theta \kappa_\tau(f_\theta(x), f_\theta(\mathbf{x}'_j)) \xrightarrow{p} \mathbb{E}_{\mathbf{x}' \sim p_{\mathbf{x}}} [\nabla_\theta \kappa_\tau(f_\theta(x), f_\theta(\mathbf{x}'))].$$

By the uniform gradient bound, differentiation under the expectation is justified, so

$$\mathbb{E}_{\mathbf{x}' \sim p_{\mathbf{x}}} [\nabla_\theta \kappa_\tau(f_\theta(x), f_\theta(\mathbf{x}'))] = \nabla_\theta \mathbb{E}_{\mathbf{x}' \sim p_{\mathbf{x}}} [\kappa_\tau(f_\theta(x), f_\theta(\mathbf{x}'))] = \nabla_\theta \Gamma_{\theta, \tau}(f_\theta(x)),$$

where the derivative on the right is the total derivative with respect to θ , including the dependence of the anchor representation $f_\theta(x)$ on θ . Since $\Gamma_{\theta, \tau}(f_\theta(x)) \geq m_\tau > 0$, the continuous mapping theorem yields

$$\nabla_\theta \log \mathbf{S}_N(f_\theta(x)) \xrightarrow{p} \frac{\nabla_\theta \Gamma_{\theta, \tau}(f_\theta(x))}{\Gamma_{\theta, \tau}(f_\theta(x))} = \nabla_\theta \log \Gamma_{\theta, \tau}(f_\theta(x)). \tag{34}$$

Combining [Eqs. \(33\)](#) and [\(34\)](#) gives

$$\nabla_\theta \log Z_{\mathcal{B}}(f_\theta(x)) \xrightarrow{p} \nabla_\theta \log \Gamma_{\theta, \tau}(f_\theta(x)). \tag{35}$$

Step 3: Passing to the population objective and identifying $\nabla_\theta \mathcal{J}_\tau$. Recall that

$$\mathcal{L}_{\text{NCE}, N}(\theta) = \mathbb{E}_{\mathcal{B}}[\ell_{\mathcal{B}}(\theta; \mathbf{x})], \quad \nabla_\theta \mathcal{L}_{\text{NCE}, N}(\theta) = \mathbb{E}_{\mathcal{B}}[\nabla_\theta \ell_{\mathcal{B}}(\theta; \mathbf{x})]$$

by [Lem. B.1](#). From [Eqs. \(31\)](#) and [\(35\)](#), we have convergence in probability

$$\nabla_{\theta} \ell_{\mathcal{B}}(\theta; \mathbf{x}) \xrightarrow{P} -\frac{1}{\tau} \nabla_{\theta} \text{sim}(f_{\theta}(\mathbf{x}), f_{\theta}(\tilde{\mathbf{x}})) + \nabla_{\theta} \log \Gamma_{\theta, \tau}(f_{\theta}(\mathbf{x})).$$

Moreover, by [Lem. B.1](#), the per-batch gradients are uniformly bounded for fixed $\tau > 0$ and hence uniformly integrable. Together with the convergence in probability above, this implies convergence in L^1 . Therefore, we may pass to expectations and obtain

$$\nabla_{\theta} \mathcal{L}_{\text{NCE}, N}(\theta) \longrightarrow -\frac{1}{\tau} \mathbb{E}_{r_{\text{um}}} [\nabla_{\theta} \text{sim}(f_{\theta}(\mathbf{x}), f_{\theta}(\tilde{\mathbf{x}}))] + \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} [\nabla_{\theta} \log \Gamma_{\theta, \tau}(f_{\theta}(\mathbf{x}))]. \quad (\star)$$

We now identify the two terms in $(\star)$ as the gradient of the parametric energy $\mathcal{J}_{\tau}(\theta)$. Recall that

$$\int_{\mathcal{Z}} U_{\theta}(z) q_{\theta}(\mathrm{d}z) := \mathbb{E}_{(\mathbf{x}, \tilde{\mathbf{x}}) \sim r_{\text{um}}} [-\text{sim}(f_{\theta}(\mathbf{x}), f_{\theta}(\tilde{\mathbf{x}}))].$$

Under [Asm. 3.1](#), the parameter gradient of the similarity term is uniformly bounded, so differentiation under the expectation gives

$$\nabla_{\theta} \int_{\mathcal{Z}} U_{\theta}(z) q_{\theta}(\mathrm{d}z) = -\mathbb{E}_{r_{\text{um}}} [\nabla_{\theta} \text{sim}(f_{\theta}(\mathbf{x}), f_{\theta}(\tilde{\mathbf{x}}))].$$

Therefore, the first term in $(\star)$ equals

$$\frac{1}{\tau} \nabla_{\theta} \int_{\mathcal{Z}} U_{\theta}(z) q_{\theta}(\mathrm{d}z).$$

By definition,

$$H_{\times}(q_{\theta}, \tilde{\rho}_{\theta, \tau}) := -\mathbb{E}_{\mathbf{z} \sim q_{\theta}} [\log \tilde{\rho}_{\theta, \tau}(\mathbf{z})] = -\mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} [\log \tilde{\rho}_{\theta, \tau}(f_{\theta}(\mathbf{x}))].$$

Under [Asm. 2.1](#),

$$\tilde{\rho}_{\theta, \tau}(z) = \frac{\Gamma_{\theta, \tau}(z)}{V_{\kappa}(\tau)},$$

with $V_{\kappa}(\tau)$ independent of θ by [Def. 2.2](#). Hence

$$\nabla_{\theta} H_{\times}(q_{\theta}, \tilde{\rho}_{\theta, \tau}) = -\mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} [\nabla_{\theta} \log \tilde{\rho}_{\theta, \tau}(f_{\theta}(\mathbf{x}))] = -\mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} [\nabla_{\theta} \log \Gamma_{\theta, \tau}(f_{\theta}(\mathbf{x}))],$$

where differentiation under the expectation is justified by the uniform boundedness established above. Thus, the second term in $(\star)$ equals

$$-\nabla_{\theta} H_{\times}(q_{\theta}, \tilde{\rho}_{\theta, \tau}).$$

Combining the two terms gives

$$\nabla_{\theta} \mathcal{L}_{\text{NCE}, N}(\theta) \longrightarrow \frac{1}{\tau} \nabla_{\theta} \int_{\mathcal{Z}} U_{\theta}(z) q_{\theta}(\mathrm{d}z) - \nabla_{\theta} H_{\times}(q_{\theta}, \tilde{\rho}_{\theta, \tau}) = \nabla_{\theta} \mathcal{J}_{\tau}(\theta).$$

Since $\theta \in \Theta$ was arbitrary, this proves

$$\|\nabla_{\theta} \mathcal{L}_{\text{NCE}, N}(\theta) - \nabla_{\theta} \mathcal{J}_{\tau}(\theta)\| \longrightarrow 0$$

for each fixed $\theta \in \Theta$ and $\tau > 0$. □

C.2 Proof of [Prop. 3.1](#)

Proposition 3.1 (Unique Gibbs equilibrium). *Assume $\mathcal{Z}$ is compact with $\mu(\mathcal{Z}) < \infty$. Fix $\tau > 0$ and let $U : \mathcal{Z} \rightarrow \mathbb{R}$ be Borel measurable and bounded below μ -a.e. Then, $\mathcal{F}_{\tau, U}$ is strictly convex on its effective domain and has the unique minimizer*

$$\rho^*(z) = \frac{\exp(-U(z)/\tau)}{Z_{\tau}}, \quad Z_{\tau} := \int_{\mathcal{Z}} \exp\left(-\frac{U(z)}{\tau}\right) \mu(\mathrm{d}z). \quad (8)$$

☞ **Proof intuition.** The objective of this proof is to show that the representation energy functional admits a unique Gibbs equilibrium. The argument has three steps. First, the space of probability densities is convex. Second, the potential term is linear in the density, while the negative entropy is strictly convex, making the full functional strictly convex on its effective domain. Finally, a variational calculation identifies the Gibbs density as the natural stationary candidate, and rewriting the free-energy gap from this density as a KL divergence verifies its global optimality and uniqueness.

Proof. We analyze the functional $\mathcal{F}_{\tau,U}$ on the space of probability densities

$$\mathcal{D}_\mu(\mathcal{Z}) := \left\{ \rho \in L^1(\mu) : \rho \geq 0 \text{ } \mu\text{-a.e.}, \int_{\mathcal{Z}} \rho(z) \mu(dz) = 1 \right\}.$$

The proof proceeds in three steps: establishing the geometry of the domain, proving strict convexity on the effective domain, and identifying the unique minimizer.

Step 1: Convexity of the domain $\mathcal{D}_\mu(\mathcal{Z})$. Consider any two densities $\rho_1, \rho_2 \in \mathcal{D}_\mu(\mathcal{Z})$ and a mixing coefficient $\lambda \in (0, 1)$. Define $\rho_\lambda := \lambda \rho_1 + (1 - \lambda) \rho_2$. Since $\rho_1, \rho_2 \geq 0$ μ -a.e., we have $\rho_\lambda \geq 0$ μ -a.e. Moreover, by linearity of the integral,

$$\int_{\mathcal{Z}} \rho_\lambda(z) \mu(dz) = \lambda \int_{\mathcal{Z}} \rho_1(z) \mu(dz) + (1 - \lambda) \int_{\mathcal{Z}} \rho_2(z) \mu(dz) = 1.$$

Therefore, $\rho_\lambda \in \mathcal{D}_\mu(\mathcal{Z})$ for every $\lambda \in (0, 1)$, so $\mathcal{D}_\mu(\mathcal{Z})$ is convex.

Step 2: Strict functional convexity. Let $\rho_1, \rho_2 \in \mathcal{D}_\mu(\mathcal{Z})$ be two distinct densities, i.e., $\rho_1 \not\equiv \rho_2$ μ -a.e., and let $\lambda \in (0, 1)$. Define $\rho_\lambda := \lambda \rho_1 + (1 - \lambda) \rho_2$. We consider densities in the effective domain of $\mathcal{F}_{\tau,U}$, where the functional is finite.

The potential-energy term is linear in ρ :

$$\frac{1}{\tau} \int_{\mathcal{Z}} U(z) \rho_\lambda(z) \mu(dz) = \frac{\lambda}{\tau} \int_{\mathcal{Z}} U(z) \rho_1(z) \mu(dz) + \frac{1 - \lambda}{\tau} \int_{\mathcal{Z}} U(z) \rho_2(z) \mu(dz).$$

The negative entropy is

$$-H(\rho) = \int_{\mathcal{Z}} \rho(z) \log \rho(z) \mu(dz),$$

with the convention $0 \log 0 = 0$. Since $\varphi(t) = t \log t$ is strictly convex on $[0, \infty)$, we have

$$\int_{\mathcal{Z}} \varphi(\rho_\lambda(z)) \mu(dz) < \lambda \int_{\mathcal{Z}} \varphi(\rho_1(z)) \mu(dz) + (1 - \lambda) \int_{\mathcal{Z}} \varphi(\rho_2(z)) \mu(dz)$$

whenever ρ_1 and ρ_2 differ on a set of positive μ -measure. Hence $-H$ is strictly convex on its effective domain. Combining the linear potential term with the strictly convex negative entropy gives

$$\mathcal{F}_{\tau,U}(\rho_\lambda) < \lambda \mathcal{F}_{\tau,U}(\rho_1) + (1 - \lambda) \mathcal{F}_{\tau,U}(\rho_2).$$

Therefore, $\mathcal{F}_{\tau,U}$ is strictly convex on its effective domain.

Step 3: Existence and uniqueness of the global minimizer. Since $\mathcal{F}_{\tau,U}$ is strictly convex on its effective domain, any minimizer, if it exists, is unique. For fixed $\tau > 0$, we first derive the candidate minimizer using the method of Lagrange multipliers. Define the Lagrangian

$$\Lambda(\rho, \gamma) := \mathcal{F}_{\tau,U}(\rho) + \gamma \left(\int_{\mathcal{Z}} \rho(z) \mu(dz) - 1 \right).$$

Formally taking the functional derivative with respect to $\rho(z)$ and imposing stationarity gives

$$\frac{\delta \Lambda}{\delta \rho(z)} = \frac{1}{\tau} U(z) + \log \rho(z) + 1 + \gamma = 0.$$

Hence the stationary density must satisfy

$$\log \rho^*(z) = -\frac{U(z)}{\tau} - \gamma - 1,$$

and therefore

$$\rho^*(z) = \frac{1}{Z_\tau} \exp\left(-\frac{U(z)}{\tau}\right), \quad Z_\tau := \int_{\mathcal{Z}} \exp\left(-\frac{U(z)}{\tau}\right) \mu(dz).$$

Since $\text{ess inf}_\mu U > -\infty$ by assumption, there exists $m \in \mathbb{R}$ such that $U(z) \geq m$ for μ -a.e. z . Hence

$$0 < Z_\tau \leq e^{-m/\tau} \mu(\mathcal{Z}) < \infty,$$

so ρ^* is a well-defined probability density in $\mathcal{D}_\mu(\mathcal{Z})$.

It remains to verify that this stationary density is indeed the global minimizer. Since

$$\log \rho^*(z) = -\frac{U(z)}{\tau} - \log Z_\tau,$$

for any ρ in the effective domain of $\mathcal{F}_{\tau,U}$,

$$\begin{aligned} D_{\text{KL}}(\rho \parallel \rho^*) &= \int_{\mathcal{Z}} \rho(z) \log \frac{\rho(z)}{\rho^*(z)} \mu(dz) \\ &= \int_{\mathcal{Z}} \rho(z) \log \rho(z) \mu(dz) + \frac{1}{\tau} \int_{\mathcal{Z}} U(z) \rho(z) \mu(dz) + \log Z_\tau \\ &= \mathcal{F}_{\tau,U}(\rho) + \log Z_\tau. \end{aligned}$$

Equivalently,

$$\mathcal{F}_{\tau,U}(\rho) = -\log Z_\tau + D_{\text{KL}}(\rho \parallel \rho^*).$$

Since $D_{\text{KL}}(\rho \parallel \rho^*) \geq 0$, with equality if and only if $\rho = \rho^*$ μ -a.e., we have

$$\mathcal{F}_{\tau,U}(\rho) \geq -\log Z_\tau = \mathcal{F}_{\tau,U}(\rho^*),$$

with equality if and only if $\rho = \rho^*$ μ -a.e. Therefore, ρ^* is the unique global minimizer of $\mathcal{F}_{\tau,U}$. $\square$

C.3 Proof of Prop. 3.2

Proposition 3.2 (Low-temperature concentration). *In the setting of Prop. 3.1, fix $\sigma > 0$ and define*

$$\mathcal{W}^\sigma := \left\{ z \in \mathcal{Z} : U(z) \leq \text{ess inf}_\mu U + \sigma \right\}.$$

Then, as $\tau \rightarrow 0^+$, the Gibbs measure $\rho^*(z) \mu(dz)$ concentrates on $\mathcal{W}^\sigma$ in the sense that, for every open $\mathcal{O} \supset \mathcal{W}^\sigma$, $\int_{\mathcal{Z} \setminus \mathcal{O}} \rho^*(z) \mu(dz) \rightarrow 0$.

Proof intuition. This proof establishes that as the temperature $\tau \rightarrow 0^+$, the Gibbs distribution concentrates near the lowest-potential regions (the σ -ground state). The argument relies on a direct comparison of exponential decay rates relative to the essential minimum level m . We compare the probability mass between two regions: **(i)** outside the σ -ground state, the potential is at least σ above m , so the corresponding Gibbs weight is bounded by a factor of $\exp(-\sigma/\tau)$; **(ii)** by the definition of the essential infimum, there is a set of positive μ -measure on which the potential lies within $\sigma/2$ of m , providing a denominator contribution bounded below by $\exp(-\sigma/(2\tau))$. Taking the ratio of these bounds yields an exponentially vanishing upper bound on the mass outside the ground-state neighborhood as $\tau \rightarrow 0^+$.

Proof. We analyze the Gibbs-equilibrium measure $\rho^* \mu$ as $\tau \rightarrow 0^+$. Let

$$m := \text{ess inf}_\mu U \quad \text{and} \quad \mathcal{W}^\sigma := \{z \in \mathcal{Z} : U(z) \leq m + \sigma\},$$

for any fixed $\sigma > 0$.

Consider any open set $\mathcal{O} \subset \mathcal{Z}$ such that $\mathcal{O} \supset \mathcal{W}^\sigma$, and denote $\mathcal{O}^c := \mathcal{Z} \setminus \mathcal{O}$. Since $\mathcal{O}$ contains $\mathcal{W}^\sigma$, its complement satisfies

$$\mathcal{O}^c \subset \{z \in \mathcal{Z} : U(z) > m + \sigma\}.$$

Therefore, using the Gibbs density

$$\rho^*(z) = \frac{1}{Z_\tau} \exp\left(-\frac{U(z)}{\tau}\right),$$

we have

$$(\rho^*\mu)(\mathcal{O}^c) = \frac{\int_{\mathcal{O}^c} \exp\left(-\frac{U(z)}{\tau}\right) \mu(dz)}{\int_{\mathcal{Z}} \exp\left(-\frac{U(w)}{\tau}\right) \mu(dw)}.$$

Multiplying numerator and denominator by $\exp(m/\tau)$ yields

$$(\rho^*\mu)(\mathcal{O}^c) = \frac{\int_{\mathcal{O}^c} \exp\left(-\frac{U(z)-m}{\tau}\right) \mu(dz)}{\int_{\mathcal{Z}} \exp\left(-\frac{U(w)-m}{\tau}\right) \mu(dw)}.$$

(i) Numerator bound. For every $z \in \mathcal{O}^c$, we have $U(z) - m > \sigma$, and hence

$$\int_{\mathcal{O}^c} \exp\left(-\frac{U(z)-m}{\tau}\right) \mu(dz) \leq \int_{\mathcal{O}^c} \exp\left(-\frac{\sigma}{\tau}\right) \mu(dz) = \mu(\mathcal{O}^c) \exp\left(-\frac{\sigma}{\tau}\right).$$

(ii) Denominator lower bound. By the definition of $m = \text{ess inf}_\mu U$, for every $\eta > 0$ the set $\{w \in \mathcal{Z} : U(w) < m + \eta\}$ has strictly positive μ -measure. Fix $\eta := \sigma/2$ and define

$$\mathcal{A} := \left\{w \in \mathcal{Z} : U(w) < m + \frac{\sigma}{2}\right\}, \quad \mu(\mathcal{A}) > 0.$$

On $\mathcal{A}$, we have $U(w) - m < \sigma/2$, hence

$$\int_{\mathcal{Z}} \exp\left(-\frac{U(w)-m}{\tau}\right) \mu(dw) \geq \int_{\mathcal{A}} \exp\left(-\frac{U(w)-m}{\tau}\right) \mu(dw) \geq \mu(\mathcal{A}) \exp\left(-\frac{\sigma}{2\tau}\right).$$

Combining the bounds, we obtain

$$(\rho^*\mu)(\mathcal{O}^c) \leq \frac{\mu(\mathcal{O}^c) \exp(-\sigma/\tau)}{\mu(\mathcal{A}) \exp(-\sigma/(2\tau))} = \frac{\mu(\mathcal{O}^c)}{\mu(\mathcal{A})} \exp\left(-\frac{\sigma}{2\tau}\right).$$

Since $\mu(\mathcal{O}^c) \leq \mu(\mathcal{Z}) < \infty$ and $\mu(\mathcal{A}) > 0$ is fixed, the right-hand side converges to 0 as $\tau \rightarrow 0^+$. Hence, for any open neighborhood $\mathcal{O} \supset \mathcal{W}^\sigma$, we have $(\rho^*\mu)(\mathcal{O}^c) \rightarrow 0$, i.e., the Gibbs measure concentrates in any open neighborhood of $\mathcal{W}^\sigma$. $\square$

C.4 Proof of Thm. 3.2

Theorem 3.2 (Low-temperature unimodal energy consistency). *Assume [Asms. 2.1, 3.1 and 3.2](#). Fix $\theta \in \Theta$ and assume $q_\theta \ll \mu$ with a strictly positive, continuous density ρ_θ , i.e., $\inf_{z \in \mathcal{Z}} \rho_\theta(z) \geq \underline{\rho}_\theta > 0$.⁴ Define the KDE mismatch $\varepsilon_{\text{kde}}^{(\theta)}(\tau) := \|\tilde{\rho}_{\theta,\tau} - \rho_\theta\|_\infty$. Then there exists $\tau_0(\theta) > 0$ such that*

$$|\mathcal{J}_\tau(\theta) - \mathcal{F}_{\tau,U_\theta}(\rho_\theta)| \leq \frac{2\varepsilon_{\text{kde}}^{(\theta)}(\tau)}{\underline{\rho}_\theta}, \quad \forall 0 < \tau \leq \tau_0(\theta),$$

and in particular, as $\tau \rightarrow 0^+$, $|\mathcal{J}_\tau(\theta) - \mathcal{F}_{\tau,U_\theta}(\rho_\theta)| \rightarrow 0$.

Proof intuition. The objective of this proof is to establish that at low temperatures, the parametric energy $\mathcal{J}_\tau(\theta)$ approaches the intrinsic thermodynamic free energy $\mathcal{F}_{\tau,U_\theta}(\rho_\theta)$ for each fixed θ . The argument hinges on translating their algebraic difference into an information-theoretic distance and then squeezing that distance to zero. **(i)** We first algebraically decompose the parametric energy and reveal a fundamental identity: the exact gap between the parametric and intrinsic energies is the KL divergence between the induced representation density ρ_θ and its kernel-smoothed counterpart $\tilde{\rho}_{\theta,\tau}$. **(ii)** To prove this KL divergence

⁴See [App. A.3](#) for a detailed discussion on the strict positivity assumption and the role of density floors.

vanishes, we ensure that the logarithmic ratio inside the integral remains controlled. By assumption, the density ρ_θ is strictly bounded away from zero. Because [Lem. B.3](#) guarantees that the smoothed density approaches the true density uniformly as $\tau \rightarrow 0^+$, the smoothed density also remains uniformly bounded away from zero for sufficiently small τ . **(iii)** The logarithm therefore operates in a Lipschitz regime, allowing the KL divergence to be controlled directly by the KDE mismatch. Since this mismatch vanishes as $\tau \rightarrow 0^+$, the parametric and intrinsic energy values agree asymptotically for each fixed θ .

Proof. **Step 1: Well-posedness of the alignment potential term.** Fix any $\theta \in \Theta$. Recall the definition of the alignment potential field ([Def. 3.1](#)):

$$U_\theta(z) := - \int_{\mathcal{Z}} \text{sim}(z, w) \nu_{\theta, z}(\mathrm{d}w) \quad \text{for } q_\theta\text{-a.e. } z.$$

Under [Asm. 3.1](#), the similarity function sim is bounded and Borel measurable, hence U_θ is Borel measurable and bounded, and therefore integrable with respect to q_θ . Moreover, $q_\theta \ll \mu$ with a continuous density ρ_θ bounded away from zero by assumption. Since $\mathcal{Z}$ is compact, ρ_θ is also bounded above, so $\rho_\theta \log \rho_\theta$ is integrable. Thus, the intrinsic representation functional ([Eq. \(7\)](#))

$$\mathcal{F}_{\tau, U_\theta}(\rho_\theta) := \frac{1}{\tau} \int_{\mathcal{Z}} U_\theta(z) \rho_\theta(z) \mu(\mathrm{d}z) - H(\rho_\theta)$$

is well-defined for each $\tau > 0$.

Step 2: Difference reduces to a KL divergence. Recall

$$\mathcal{J}_\tau(\theta) = \frac{1}{\tau} \int_{\mathcal{Z}} U_\theta(z) q_\theta(\mathrm{d}z) - H_\times(q_\theta, \tilde{\rho}_{\theta, \tau}).$$

Using

$$H_\times(q, \tilde{\rho}) = H(\rho) + D_{\text{KL}}(\rho \| \tilde{\rho}), \quad \rho(z) := \frac{q(\mathrm{d}z)}{\mu(\mathrm{d}z)},$$

we obtain

$$\mathcal{J}_\tau(\theta) = \frac{1}{\tau} \int_{\mathcal{Z}} U_\theta(z) \rho_\theta(z) \mu(\mathrm{d}z) - H(\rho_\theta) - D_{\text{KL}}(\rho_\theta \| \tilde{\rho}_{\theta, \tau}),$$

and hence

$$\mathcal{F}_{\tau, U_\theta}(\rho_\theta) - \mathcal{J}_\tau(\theta) = D_{\text{KL}}(\rho_\theta \| \tilde{\rho}_{\theta, \tau}). \tag{36}$$

Step 3: Vanishing KL divergence of the smoothing. Fix any $\theta \in \Theta$. By assumption, the continuous density is uniformly bounded away from zero:

$$\inf_{z \in \mathcal{Z}} \rho_\theta(z) \geq \underline{\rho}_\theta > 0.$$

Recall from [Lem. B.3](#) that the KDE mismatch vanishes in the low-temperature limit:

$$\varepsilon_{\text{kde}}^{(\theta)}(\tau) := \|\tilde{\rho}_{\theta, \tau} - \rho_\theta\|_\infty \longrightarrow 0 \quad \text{as } \tau \rightarrow 0^+.$$

Hence there exists $\tau_0(\theta) > 0$ such that, for all $0 < \tau \leq \tau_0(\theta)$,

$$\varepsilon_{\text{kde}}^{(\theta)}(\tau) \leq \frac{\underline{\rho}_\theta}{2}.$$

It follows that, for every $z \in \mathcal{Z}$,

$$\tilde{\rho}_{\theta, \tau}(z) \geq \rho_\theta(z) - \frac{\underline{\rho}_\theta}{2} \geq \frac{\underline{\rho}_\theta}{2}. \tag{37}$$

On $[\underline{\rho}_\theta/2, \infty)$, the derivative of $\log x$ satisfies $1/x \leq 2/\underline{\rho}_\theta$. Hence, for all $a, b \geq \underline{\rho}_\theta/2$,

$$|\log a - \log b| \leq \frac{2}{\underline{\rho}_\theta} |a - b|.$$

Applying this with $a = \rho_\theta(z)$ and $b = \tilde{\rho}_{\theta,\tau}(z)$ gives

$$\left| \log \frac{\rho_\theta(z)}{\tilde{\rho}_{\theta,\tau}(z)} \right| = |\log \rho_\theta(z) - \log \tilde{\rho}_{\theta,\tau}(z)| \leq \frac{2}{\rho_\theta} |\rho_\theta(z) - \tilde{\rho}_{\theta,\tau}(z)| \leq \frac{2\varepsilon_{\text{kde}}^{(\theta)}(\tau)}{\rho_\theta}. \quad (38)$$

Now

$$D_{\text{KL}}(\rho_\theta \| \tilde{\rho}_{\theta,\tau}) = \int_{\mathcal{Z}} \rho_\theta(z) \log \frac{\rho_\theta(z)}{\tilde{\rho}_{\theta,\tau}(z)} \mu(dz).$$

Using $\int_{\mathcal{Z}} \rho_\theta(z) \mu(dz) = 1$ and [Eq. \(38\)](#),

$$D_{\text{KL}}(\rho_\theta \| \tilde{\rho}_{\theta,\tau}) \leq \int_{\mathcal{Z}} \rho_\theta(z) \left| \log \frac{\rho_\theta(z)}{\tilde{\rho}_{\theta,\tau}(z)} \right| \mu(dz) \leq \frac{2\varepsilon_{\text{kde}}^{(\theta)}(\tau)}{\rho_\theta}. \quad (39)$$

Substituting this into [Eq. \(36\)](#) yields

$$|\mathcal{J}_\tau(\theta) - \mathcal{F}_{\tau, U_\theta}(\rho_\theta)| \leq \frac{2\varepsilon_{\text{kde}}^{(\theta)}(\tau)}{\rho_\theta}.$$

Finally, [Lem. B.3](#) guarantees that $\varepsilon_{\text{kde}}^{(\theta)}(\tau) \rightarrow 0$ as $\tau \rightarrow 0^+$. Consequently,

$$|\mathcal{J}_\tau(\theta) - \mathcal{F}_{\tau, U_\theta}(\rho_\theta)| \rightarrow 0 \quad \text{as } \tau \rightarrow 0^+,$$

which concludes the proof. $\square$

D Proofs of Multimodal Formal Statements

This section provides rigorous proofs for all formal statements concerning our multimodal analysis. Each statement from the main text is restated for self-containment, accompanied by a proof intuition for providing high-level insights.

D.1 Proof of [Thm. 4.1](#)

Theorem 4.1 (Large-batch multimodal consistency). *Consider the symmetric multimodal InfoNCE objective $\mathcal{L}_{\text{Sym},N}(\theta, \phi)$ in [Eq. \(3\)](#) and the energy $\mathcal{J}_\tau^{\text{Sym}}(\theta, \phi)$ in [Def. 4.2](#). Assume [Asms. 2.1](#) and [4.1](#). Then, for any fixed $\tau > 0$ and $(\theta, \phi) \in \Theta \times \Phi$, as $N \rightarrow \infty$,*

$$|\mathcal{L}_{\text{Sym},N}(\theta, \phi) - \mathcal{J}_\tau^{\text{Sym}}(\theta, \phi) - \log(NV_\kappa(\tau))| \rightarrow 0, \quad \|\nabla_{(\theta,\phi)} \mathcal{L}_{\text{Sym},N}(\theta, \phi) - \nabla_{(\theta,\phi)} \mathcal{J}_\tau^{\text{Sym}}(\theta, \phi)\| \rightarrow 0.$$

Proof intuition. This proof extends the large-batch consistency analysis of unimodal InfoNCE to the dual-encoder, cross-modal setting. The logic follows the unimodal proof ([Thm. 3.1](#)) in each direction. **(i)** The symmetric loss is the average of two directional losses. As $N \rightarrow \infty$, the normalized sum of negative keys in each direction converges to the population kernel average of the corresponding target modality. This yields a shared attractive alignment term together with two directional cross-entropy terms. **(ii)** The same large-batch limit applies to the gradients of the two log-partition terms: the contribution of the single positive pair vanishes, while the negative-sum ratios converge to their population counterparts. Uniform gradient bounds then allow these limits to pass to the expected finite- N objective, yielding pointwise value and gradient consistency with the deterministic symmetric energy.

Proof. Consider a random batch $\mathcal{B}$ consisting of a positive pair $(\mathbf{x}, \mathbf{y}) \sim r_{\text{mm}}$ together with N i.i.d. negatives $\{\mathbf{y}'_j\}_{j=1}^N \stackrel{\text{i.i.d.}}{\sim} p_{\mathbf{y}}$ and N i.i.d. negatives $\{\mathbf{x}'_j\}_{j=1}^N \stackrel{\text{i.i.d.}}{\sim} p_{\mathbf{x}}$, all independent conditional on $(\mathbf{x}, \mathbf{y})$. Write the corresponding random representations as

$$\mathbf{z} = f_\theta(\mathbf{x}), \quad \mathbf{w} = g_\phi(\mathbf{y}), \quad \mathbf{w}'_j = g_\phi(\mathbf{y}'_j), \quad \mathbf{z}'_j = f_\theta(\mathbf{x}'_j),$$

and define the exponential kernel

$$\kappa_\tau(z, w) := \exp\left(\frac{\text{sim}(z, w)}{\tau}\right).$$

Denote the directional partition sums by

$$Z_{\mathcal{B}}^{\theta \rightarrow \phi}(\mathbf{z}) := \kappa_{\tau}(\mathbf{z}, \mathbf{w}) + \sum_{j=1}^N \kappa_{\tau}(\mathbf{z}, \mathbf{w}'_j), \quad Z_{\mathcal{B}}^{\phi \rightarrow \theta}(\mathbf{w}) := \kappa_{\tau}(\mathbf{z}, \mathbf{w}) + \sum_{j=1}^N \kappa_{\tau}(\mathbf{z}'_j, \mathbf{w}).$$

The symmetric per-batch loss can be written as

$$\ell_{\mathcal{B}}^{\text{Sym}}(\theta, \phi) = \frac{1}{2} \underbrace{\left[-\frac{1}{\tau} \text{sim}(\mathbf{z}, \mathbf{w}) + \log Z_{\mathcal{B}}^{\theta \rightarrow \phi}(\mathbf{z}) \right]}_{\ell_{\mathcal{B}}^{x \rightarrow y}(\theta, \phi)} + \frac{1}{2} \underbrace{\left[-\frac{1}{\tau} \text{sim}(\mathbf{z}, \mathbf{w}) + \log Z_{\mathcal{B}}^{\phi \rightarrow \theta}(\mathbf{w}) \right]}_{\ell_{\mathcal{B}}^{y \rightarrow x}(\theta, \phi)}. \quad (40)$$

Part I. Value consistency. Fix any $\tau > 0$ and $(\theta, \phi) \in \Theta \times \Phi$. We analyze the directional loss $\mathcal{L}_{\text{NCE}, N}^{x \rightarrow y}(\theta, \phi)$; the reverse direction follows by symmetry. From Eq. (40), the corresponding per-batch directional loss is

$$\ell_{\mathcal{B}}^{x \rightarrow y}(\theta, \phi) := -\frac{1}{\tau} \text{sim}(\mathbf{z}, \mathbf{w}) + \log Z_{\mathcal{B}}^{\theta \rightarrow \phi}(\mathbf{z}). \quad (41)$$

Step 1: Bias–variance control of the log-partition term. The structure of $Z_{\mathcal{B}}^{\theta \rightarrow \phi}$ is identical to the unimodal partition function (Eq. (26)), except that the negatives are drawn from q_{ϕ} rather than q_{θ} . By assumption, $\mathcal{Z}$ is compact and sim is continuous, hence κ_{τ} is bounded on $\mathcal{Z} \times \mathcal{Z}$ for each fixed $\tau > 0$. Let

$$\mathbf{P} := \kappa_{\tau}(\mathbf{z}, \mathbf{w}), \quad \mathbf{S}_N^{\theta \rightarrow \phi}(\mathbf{z}) := \sum_{j=1}^N \kappa_{\tau}(\mathbf{z}, \mathbf{w}'_j), \quad Z_{\mathcal{B}}^{\theta \rightarrow \phi} = \mathbf{P} + \mathbf{S}_N^{\theta \rightarrow \phi}.$$

As in Step 1 of the unimodal proof (Part I of Thm. 3.1), the positive contribution induces a vanishing bias:

$$\log(\mathbf{P} + \mathbf{S}_N^{\theta \rightarrow \phi}) = \log \mathbf{S}_N^{\theta \rightarrow \phi} + \log \left(1 + \frac{\mathbf{P}}{\mathbf{S}_N^{\theta \rightarrow \phi}} \right), \quad \log \left(1 + \frac{\mathbf{P}}{\mathbf{S}_N^{\theta \rightarrow \phi}} \right) = O_{\tau}(N^{-1}).$$

Define the population partition function induced by the target encoder ϕ :

$$\Gamma_{\phi, \tau}(z) := \mathbb{E}_{\mathbf{w} \sim q_{\phi}} [\kappa_{\tau}(z, \mathbf{w})] = \int_{\mathcal{Z}} \kappa_{\tau}(z, w) q_{\phi}(dw).$$

Condition on the anchor observation $\mathbf{x} = x$ and write $z = f_{\theta}(x)$. Since the negatives $\{\mathbf{y}'_j\}_{j=1}^N$ are i.i.d. from $p_{\mathbf{y}}$, the representations $\{\mathbf{w}'_j\}_{j=1}^N$ are i.i.d. with law q_{ϕ} . Hence, by the same concentration argument as in Step 2 of Part I of Thm. 3.1,

$$\frac{1}{N} \mathbf{S}_N^{\theta \rightarrow \phi}(z) = \Gamma_{\phi, \tau}(z) + O_p(N^{-1/2}),$$

which implies, by the same logarithmic expansion used in the unimodal proof,

$$\log \mathbf{S}_N^{\theta \rightarrow \phi}(z) = \log N + \log \Gamma_{\phi, \tau}(z) + O_p(N^{-1/2}).$$

Returning to the random anchor $\mathbf{z}$, combining the above displays yields

$$\ell_{\mathcal{B}}^{x \rightarrow y}(\theta, \phi) = -\frac{1}{\tau} \text{sim}(\mathbf{z}, \mathbf{w}) + \log N + \log \Gamma_{\phi, \tau}(\mathbf{z}) + O_p(N^{-1/2}). \quad (42)$$

Step 2: Identification of the directional energy. Taking expectations of Eq. (42) over the batch construction gives

$$\mathcal{L}_{\text{NCE}, N}^{x \rightarrow y}(\theta, \phi) := \mathbb{E}_{\mathcal{B}} [\ell_{\mathcal{B}}^{x \rightarrow y}(\theta, \phi)] = \mathbb{E} \left[-\frac{1}{\tau} \text{sim}(\mathbf{z}, \mathbf{w}) \right] + \mathbb{E} [\log \Gamma_{\phi, \tau}(\mathbf{z})] + \log N + o(1).$$

Here, for fixed $\tau > 0$, the remainder is uniformly bounded and hence uniformly integrable. Since it converges to zero in probability, it also converges to zero in L^1 , so its expectation is $o(1)$.

Using the disintegration in [Def. 4.1](#), the law of iterated expectations yields

$$\mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim r_{\text{mm}}} \left[-\frac{1}{\tau} \text{sim}(f_{\theta}(\mathbf{x}), g_{\phi}(\mathbf{y})) \right] = \frac{1}{\tau} \mathbb{E}_{\mathbf{z} \sim q_{\theta}} [U_{\theta \rightarrow \phi}(\mathbf{z})] = \frac{1}{\tau} \int_{\mathcal{Z}} U_{\theta \rightarrow \phi}(z) q_{\theta}(\mathrm{d}z).$$

Since $\mathbf{z} = f_{\theta}(\mathbf{x}) \sim q_{\theta}$,

$$\mathbb{E} [\log \Gamma_{\phi, \tau}(\mathbf{z})] = \mathbb{E}_{\mathbf{z} \sim q_{\theta}} [\log \Gamma_{\phi, \tau}(\mathbf{z})].$$

Under the constant kernel-volume condition [Asm. 2.1](#),

$$\tilde{\rho}_{\phi, \tau}(z) = \frac{\Gamma_{\phi, \tau}(z)}{V_{\kappa}(\tau)}$$

by [Def. 2.2](#), and hence

$$\begin{aligned} \mathbb{E}_{\mathbf{z} \sim q_{\theta}} [\log \Gamma_{\phi, \tau}(\mathbf{z})] &= \log V_{\kappa}(\tau) + \mathbb{E}_{\mathbf{z} \sim q_{\theta}} [\log \tilde{\rho}_{\phi, \tau}(\mathbf{z})] \\ &= \log V_{\kappa}(\tau) - H_{\times}(q_{\theta}, \tilde{\rho}_{\phi, \tau}). \end{aligned}$$

Substituting into the preceding expansion gives

$$\begin{aligned} \mathcal{L}_{\text{NCE}, N}^{x \rightarrow y}(\theta, \phi) &= \frac{1}{\tau} \int_{\mathcal{Z}} U_{\theta \rightarrow \phi}(z) q_{\theta}(\mathrm{d}z) - H_{\times}(q_{\theta}, \tilde{\rho}_{\phi, \tau}) + \log(NV_{\kappa}(\tau)) + o(1) \\ &= \mathcal{J}_{\tau}^{x \rightarrow y}(\theta, \phi) + \log(NV_{\kappa}(\tau)) + o(1). \end{aligned}$$

The reverse direction $y \rightarrow x$ follows analogously, yielding

$$\mathcal{L}_{\text{NCE}, N}^{y \rightarrow x}(\phi, \theta) = \mathcal{J}_{\tau}^{y \rightarrow x}(\phi, \theta) + \log(NV_{\kappa}(\tau)) + o(1).$$

Averaging the two directional expansions and using [Eq. \(10\)](#) gives

$$|\mathcal{L}_{\text{Sym}, N}(\theta, \phi) - \mathcal{J}_{\tau}^{\text{Sym}}(\theta, \phi) - \log(NV_{\kappa}(\tau))| \longrightarrow 0$$

as $N \rightarrow \infty$, for each fixed $\tau > 0$ and $(\theta, \phi) \in \Theta \times \Phi$.

Part II. Gradient consistency. Fix any $\tau > 0$ and $(\theta, \phi) \in \Theta \times \Phi$. Differentiating the finite-batch loss in [Eq. \(40\)](#) gives

$$\nabla \ell_{\mathcal{B}}^{\text{Sym}}(\theta, \phi) = -\frac{1}{\tau} \nabla \text{sim}(\mathbf{z}, \mathbf{w}) + \frac{1}{2} \nabla \log Z_{\mathcal{B}}^{\theta \rightarrow \phi}(\mathbf{z}) + \frac{1}{2} \nabla \log Z_{\mathcal{B}}^{\phi \rightarrow \theta}(\mathbf{w}), \quad (43)$$

where $\nabla := \nabla_{(\theta, \phi)}$ denotes the joint gradient.

Step 1: Removing the positive term in each log-partition gradient. Define

$$\mathbf{g}_N^{\theta \rightarrow \phi} := \nabla \log(\mathbf{P} + \mathbf{S}_N^{\theta \rightarrow \phi}), \quad \mathbf{r}_N^{\theta \rightarrow \phi} := \nabla \log \mathbf{S}_N^{\theta \rightarrow \phi},$$

and analogously define $(\mathbf{g}_N^{\phi \rightarrow \theta}, \mathbf{r}_N^{\phi \rightarrow \theta})$. As in [Eq. \(32\)](#), a direct algebraic manipulation yields

$$\mathbf{g}_N^{\theta \rightarrow \phi} - \mathbf{r}_N^{\theta \rightarrow \phi} = \frac{\nabla \mathbf{P}}{\mathbf{P} + \mathbf{S}_N^{\theta \rightarrow \phi}} - \frac{\mathbf{P}}{\mathbf{S}_N^{\theta \rightarrow \phi}(\mathbf{P} + \mathbf{S}_N^{\theta \rightarrow \phi})} \nabla \mathbf{S}_N^{\theta \rightarrow \phi},$$

and the same identity holds with $\theta \rightarrow \phi$ replaced by $\phi \rightarrow \theta$.

By [Asm. 4.1](#), $\mathcal{Z}$ is compact and sim is C^1 with bounded input gradients. Together with the C^1 encoders with uniformly bounded Jacobians, this implies that, for each fixed $\tau > 0$, there exist finite constants

$$0 < m_{\tau} \leq \kappa_{\tau}(\cdot, \cdot) \leq M_{\tau} < \infty, \quad \sup_{(\theta, \phi) \in \Theta \times \Phi} \sup_{\mathcal{B}} \|\nabla \kappa_{\tau}\| \leq G_{\tau} < \infty.$$

Consequently, uniformly over (θ, ϕ) and batch realizations,

$$\mathbf{P} \leq M_{\tau}, \quad \|\nabla \mathbf{P}\| \leq G_{\tau}, \quad \mathbf{S}_N^{\theta \rightarrow \phi} \geq N m_{\tau}, \quad \|\nabla \mathbf{S}_N^{\theta \rightarrow \phi}\| \leq N G_{\tau},$$

and the same bounds hold for $S_N^{\phi \rightarrow \theta}$. Taking norms in the identity above gives

$$\left\| \nabla \log Z_B^{\theta \rightarrow \phi}(\mathbf{z}) - \nabla \log S_N^{\theta \rightarrow \phi}(\mathbf{z}) \right\| = O_\tau(N^{-1}), \quad \left\| \nabla \log Z_B^{\phi \rightarrow \theta}(\mathbf{w}) - \nabla \log S_N^{\phi \rightarrow \theta}(\mathbf{w}) \right\| = O_\tau(N^{-1}). \quad (44)$$

Step 2: Consistency of the negative-sum ratios. We treat the direction $\theta \rightarrow \phi$; the reverse direction is analogous. Condition on the anchor observations $\mathbf{x} = x$ and $\mathbf{y} = y$, and write

$$z = f_\theta(x), \quad w = g_\phi(y).$$

Then

$$\nabla \log S_N^{\theta \rightarrow \phi}(z) = \frac{\frac{1}{N} \sum_{j=1}^N \nabla \kappa_\tau(f_\theta(x), g_\phi(\mathbf{y}'_j))}{\frac{1}{N} \sum_{j=1}^N \kappa_\tau(f_\theta(x), g_\phi(\mathbf{y}'_j))}.$$

Given x , the negatives $\{\mathbf{y}'_j\}_{j=1}^N$ are i.i.d. with law $p_{\mathbf{y}}$. Define

$$\Gamma_{\phi, \tau}(f_\theta(x)) = \mathbb{E}_{\mathbf{y}' \sim p_{\mathbf{y}}} [\kappa_\tau(f_\theta(x), g_\phi(\mathbf{y}'))].$$

Since $\Gamma_{\phi, \tau}(f_\theta(x)) \geq m_\tau > 0$, the denominator is bounded away from zero at the population level. By the law of large numbers and boundedness of κ_τ ,

$$\frac{1}{N} \sum_{j=1}^N \kappa_\tau(f_\theta(x), g_\phi(\mathbf{y}'_j)) \xrightarrow{p} \Gamma_{\phi, \tau}(f_\theta(x)).$$

Similarly, since the total parameter gradient of κ_τ is uniformly bounded,

$$\frac{1}{N} \sum_{j=1}^N \nabla \kappa_\tau(f_\theta(x), g_\phi(\mathbf{y}'_j)) \xrightarrow{p} \mathbb{E}_{\mathbf{y}' \sim p_{\mathbf{y}}} [\nabla \kappa_\tau(f_\theta(x), g_\phi(\mathbf{y}'))].$$

Since the expectation is taken with respect to the parameter-free distribution $p_{\mathbf{y}}$ and the total parameter gradient is uniformly bounded by [Asm. 4.1](#), differentiation under the expectation is justified. Hence

$$\mathbb{E}_{\mathbf{y}' \sim p_{\mathbf{y}}} [\nabla \kappa_\tau(f_\theta(x), g_\phi(\mathbf{y}'))] = \nabla \mathbb{E}_{\mathbf{y}' \sim p_{\mathbf{y}}} [\kappa_\tau(f_\theta(x), g_\phi(\mathbf{y}'))] = \nabla \Gamma_{\phi, \tau}(f_\theta(x)),$$

where the derivative on the right is the total derivative with respect to (θ, ϕ) , including the dependence of the anchor representation $f_\theta(x)$ on θ . The continuous mapping theorem therefore yields

$$\nabla \log S_N^{\theta \rightarrow \phi}(z) \xrightarrow{p} \frac{\nabla \Gamma_{\phi, \tau}(f_\theta(x))}{\Gamma_{\phi, \tau}(f_\theta(x))} = \nabla \log \Gamma_{\phi, \tau}(f_\theta(x)). \quad (45)$$

The reverse direction follows identically. Conditioning on $\mathbf{x} = x$ and $\mathbf{y} = y$ and writing $w = g_\phi(y)$, define

$$\Gamma_{\theta, \tau}(g_\phi(y)) := \mathbb{E}_{\mathbf{x}' \sim p_{\mathbf{x}}} [\kappa_\tau(f_\theta(\mathbf{x}'), g_\phi(y))].$$

Then

$$\nabla \log S_N^{\phi \rightarrow \theta}(w) \xrightarrow{p} \nabla \log \Gamma_{\theta, \tau}(g_\phi(y)). \quad (46)$$

Returning to the random anchors and combining [Eqs. \(44\) to \(46\)](#), as $N \rightarrow \infty$ we obtain

$$\nabla \log Z_B^{\theta \rightarrow \phi}(\mathbf{z}) \xrightarrow{p} \nabla \log \Gamma_{\phi, \tau}(\mathbf{z}), \quad \nabla \log Z_B^{\phi \rightarrow \theta}(\mathbf{w}) \xrightarrow{p} \nabla \log \Gamma_{\theta, \tau}(\mathbf{w}). \quad (47)$$

Step 3: Passing to the population objective and identifying $\nabla_{(\theta, \phi)} \mathcal{J}_\tau^{\text{Sym}}$. By [Lem. B.2](#), $\|\nabla \ell_B^{\text{Sym}}(\theta, \phi)\|$ is uniformly bounded for fixed $\tau > 0$ and hence uniformly integrable. Together with the convergence in probability in [Eq. \(47\)](#), this implies convergence in L^1 , so we may pass to expectations and obtain

$$\begin{aligned} \nabla \mathcal{L}_{\text{Sym}, N}(\theta, \phi) &= \mathbb{E}_B [\nabla \ell_B^{\text{Sym}}(\theta, \phi)] \longrightarrow -\frac{1}{\tau} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim r_{\text{mm}}} [\nabla \text{sim}(f_\theta(\mathbf{x}), g_\phi(\mathbf{y}))] \\ &\quad + \frac{1}{2} \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} [\nabla \log \Gamma_{\phi, \tau}(f_\theta(\mathbf{x}))] \\ &\quad + \frac{1}{2} \mathbb{E}_{\mathbf{y} \sim p_{\mathbf{y}}} [\nabla \log \Gamma_{\theta, \tau}(g_\phi(\mathbf{y}))]. \end{aligned} \quad (48)$$

We now identify the right-hand side as $\nabla \mathcal{J}_\tau^{\text{Sym}}(\theta, \phi)$. First, by the definition of $\pi_{\theta\phi} = (f_\theta \times g_\phi)_\# r_{\text{mm}}$ and the law of iterated expectations applied to [Def. 4.1](#),

$$\int_{\mathcal{Z}} U_{\theta \rightarrow \phi}(z) q_\theta(dz) = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim r_{\text{mm}}} [-\text{sim}(f_\theta(\mathbf{x}), g_\phi(\mathbf{y}))] = \mathbb{E}_{(\mathbf{z}, \mathbf{w}) \sim \pi_{\theta\phi}} [-\text{sim}(\mathbf{z}, \mathbf{w})],$$

and similarly,

$$\int_{\mathcal{Z}} U_{\phi \rightarrow \theta}(w) q_\phi(dw) = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim r_{\text{mm}}} [-\text{sim}(f_\theta(\mathbf{x}), g_\phi(\mathbf{y}))].$$

While these two binding energies coincide after integration, the directional potential fields $U_{\theta \rightarrow \phi}$ and $U_{\phi \rightarrow \theta}$ generally differ pointwise. Under [Asm. 4.1](#), differentiation under the parameter-free expectation is justified, yielding

$$\nabla \int_{\mathcal{Z}} U_{\theta \rightarrow \phi}(z) q_\theta(dz) = \nabla \int_{\mathcal{Z}} U_{\phi \rightarrow \theta}(w) q_\phi(dw) = -\mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim r_{\text{mm}}} [\nabla \text{sim}(f_\theta(\mathbf{x}), g_\phi(\mathbf{y}))].$$

Second, using

$$\log \tilde{\rho}_{\phi, \tau}(z) = \log \Gamma_{\phi, \tau}(z) - \log V_\kappa(\tau), \quad \log \tilde{\rho}_{\theta, \tau}(w) = \log \Gamma_{\theta, \tau}(w) - \log V_\kappa(\tau),$$

with $V_\kappa(\tau)$ independent of (θ, ϕ) by [Asm. 2.1](#), we may write

$$\begin{aligned} \nabla H_\times(q_\theta, \tilde{\rho}_{\phi, \tau}) &= -\nabla \mathbb{E}_{\mathbf{x} \sim p_\mathbf{x}} [\log \tilde{\rho}_{\phi, \tau}(f_\theta(\mathbf{x}))] = -\mathbb{E}_{\mathbf{x} \sim p_\mathbf{x}} [\nabla \log \Gamma_{\phi, \tau}(f_\theta(\mathbf{x}))], \\ \nabla H_\times(q_\phi, \tilde{\rho}_{\theta, \tau}) &= -\nabla \mathbb{E}_{\mathbf{y} \sim p_\mathbf{y}} [\log \tilde{\rho}_{\theta, \tau}(g_\phi(\mathbf{y}))] = -\mathbb{E}_{\mathbf{y} \sim p_\mathbf{y}} [\nabla \log \Gamma_{\theta, \tau}(g_\phi(\mathbf{y}))], \end{aligned}$$

where differentiation under the expectations is justified by the uniform gradient bounds in [Asm. 4.1](#) and [Lem. B.2](#).

Combining these identities with the definition of $\mathcal{J}_\tau^{\text{Sym}}$ in [Def. 4.2](#) shows that the limit in [Eq. \(48\)](#) equals $\nabla \mathcal{J}_\tau^{\text{Sym}}(\theta, \phi)$. Therefore, as $N \rightarrow \infty$, for each fixed $\tau > 0$ and $(\theta, \phi) \in \Theta \times \Phi$, the two parameter components satisfy

$$\|\nabla_\theta \mathcal{L}_{\text{Sym}, N}(\theta, \phi) - \nabla_\theta \mathcal{J}_\tau^{\text{Sym}}(\theta, \phi)\| \rightarrow 0, \quad \|\nabla_\phi \mathcal{L}_{\text{Sym}, N}(\theta, \phi) - \nabla_\phi \mathcal{J}_\tau^{\text{Sym}}(\theta, \phi)\| \rightarrow 0.$$

Consequently, for the joint gradient $\nabla_{(\theta, \phi)} = (\nabla_\theta, \nabla_\phi)$,

$$\begin{aligned} &\left\| \nabla_{(\theta, \phi)} \mathcal{L}_{\text{Sym}, N}(\theta, \phi) - \nabla_{(\theta, \phi)} \mathcal{J}_\tau^{\text{Sym}}(\theta, \phi) \right\| \\ &= \sqrt{\left\| \nabla_\theta \mathcal{L}_{\text{Sym}, N}(\theta, \phi) - \nabla_\theta \mathcal{J}_\tau^{\text{Sym}}(\theta, \phi) \right\|^2 + \left\| \nabla_\phi \mathcal{L}_{\text{Sym}, N}(\theta, \phi) - \nabla_\phi \mathcal{J}_\tau^{\text{Sym}}(\theta, \phi) \right\|^2} \\ &\rightarrow 0. \end{aligned}$$

□

D.2 Explicit Extremal Construction and Proof of [Prop. 4.1](#)

We now give the extremal-density construction underlying [Prop. 4.1](#) and the full proof of the proposition.

Proposition D.1 (Constructive form of the coordinate infimum and extremal convergence). *Assume $\mathcal{Z}$ is compact with $\mu(\mathcal{Z}) < \infty$, and let $\mathcal{F}_{\tau, \mathbf{U}_{1,2}}^{\text{Sym}}$ be as in [Def. 4.3](#). Fix $\rho_2 \in \mathcal{D}_{\mu, \rho_2}(\mathcal{Z})$ and define the effective potential field by*

$$V_{1|2}(z) := \frac{1}{\tau} U_{1,2}(z) + \log \rho_2(z).$$

Then, the coordinate map $\rho_1 \mapsto \mathcal{F}_{\tau, \mathbf{U}_{1,2}}^{\text{Sym}}(\rho_1, \rho_2)$ is concave on $\mathcal{D}_{\mu, \rho_1}(\mathcal{Z})$, and its infimum is approached by sequences that concentrate the excess mass $M_{\text{ex}}^{(1)} := 1 - \rho_1 \mu(\mathcal{Z})$ onto approximate minimizers of $V_{1|2}$, leaving only the floor ρ_1 elsewhere.

Formally, for $\sigma > 0$, let

$$\mathcal{W}_1^\sigma := \{z \in \mathcal{Z} : V_{1|2}(z) \leq \text{ess inf}_\mu V_{1|2} + \sigma\}.$$

Choose measurable sets $\mathcal{S}_1^\sigma \subseteq \mathcal{W}_1^\sigma$ with $0 < \mu(\mathcal{S}_1^\sigma) < \infty$ and $\mu(\mathcal{S}_1^\sigma) \rightarrow 0$ as $\sigma \rightarrow 0$, and define the σ -approximate density

$$\rho_1^{(\sigma)}(z) = \rho_1 + \mathbb{1}_{\mathcal{S}_1^\sigma}(z) \frac{M_{\text{ex}}^{(1)}}{\mu(\mathcal{S}_1^\sigma)}. \quad (49)$$

Then, as $\sigma \rightarrow 0$, the energy of $\rho_1^{(\sigma)}$ converges to the coordinate infimum:

$$\lim_{\sigma \rightarrow 0} \mathcal{F}_{\tau, \mathbf{U}_{1,2}}^{\text{Sym}}(\rho_1^{(\sigma)}, \rho_2) = \inf_{\rho_1 \in \mathcal{D}_{\mu, \underline{\rho}_1}(\mathcal{Z})} \mathcal{F}_{\tau, \mathbf{U}_{1,2}}^{\text{Sym}}(\rho_1, \rho_2). \quad (50)$$

An analogous statement holds when optimizing in ρ_2 with ρ_1 fixed, with respect to the dual effective potential field $V_{2|1}(z) := \frac{1}{\tau}U_{2 \star 1}(z) + \log \rho_1(z)$.

Proof intuition. This proof establishes that when one modality's distribution (ρ_2) is held fixed, an infimizing sequence for the other modality (ρ_1) approaches an extreme, “spiky” configuration. Rather than spreading out, it approaches the minimum allowed density everywhere while concentrating its remaining probability mass near the lowest regions of an effective potential landscape. The logic unfolds in three geometric phases: **(i)** Unlike the unimodal free energy, the cross-entropy term makes the functional *concave* with respect to ρ_1 . This motivates extremal configurations near the boundary of the feasible density class. We derive a lower bound corresponding to a background floor $\underline{\rho}_1$ together with all excess probability mass placed near the essential minimum of the effective potential $V_{1|2}$. **(ii)** We then construct a sequence of feasible densities whose excess mass is concentrated on shrinking subsets of the corresponding low-effective-potential regions. **(iii)** Although this concentration makes the spike increasingly tall, the peer density is bounded, so the additional logarithmic term scales as $\mu(\mathcal{S}_1^\sigma) \log(1/\mu(\mathcal{S}_1^\sigma))$ and therefore vanishes as the support shrinks. The resulting sequence approaches the coordinate infimum while concentrating its excess mass near minimizers of the effective potential.

Proof. Fix $\tau > 0$ and $\rho_2 \in \mathcal{D}_{\mu, \underline{\rho}_2}(\mathcal{Z})$. Throughout, we treat ρ_2 as fixed and analyze ρ_1 ; the argument for ρ_2 is symmetric. Write the coordinate map as

$$\mathcal{F}_1(\rho_1) := \mathcal{F}_{\tau, \mathbf{U}_{1,2}}^{\text{Sym}}(\rho_1, \rho_2), \quad \rho_1 \in \mathcal{D}_{\mu, \underline{\rho}_1}(\mathcal{Z}).$$

Step 1: Concavity of $\rho_1 \mapsto \mathcal{F}_1$ and its lower bound. Using

$$H(\rho_1) = - \int_{\mathcal{Z}} \rho_1(z) \log \rho_1(z) \mu(dz),$$

$$D_{\text{KL}}(\rho_1 \| \rho_2) = \int_{\mathcal{Z}} \rho_1(z) \log \frac{\rho_1(z)}{\rho_2(z)} \mu(dz),$$

and

$$D_{\text{KL}}(\rho_2 \| \rho_1) = \int_{\mathcal{Z}} \rho_2(z) \log \frac{\rho_2(z)}{\rho_1(z)} \mu(dz),$$

we expand the ρ_1 -dependent terms in $\mathcal{F}_1(\rho_1)$:

$$\mathcal{F}_1(\rho_1) = \frac{1}{2} \int_{\mathcal{Z}} \rho_1(z) V_{1|2}(z) \mu(dz) + \frac{1}{2} \int_{\mathcal{Z}} \rho_2(z) \log \rho_1(z) \mu(dz) + C(\rho_2), \quad (51)$$

where

$$V_{1|2}(z) := \frac{1}{\tau} U_{1 \star 2}(z) + \log \rho_2(z),$$

and $C(\rho_2)$ does not depend on ρ_1 .

The map $\rho_1 \mapsto \int_{\mathcal{Z}} \rho_1(z) V_{1|2}(z) \mu(dz)$ is linear, while $\rho_1 \mapsto \int_{\mathcal{Z}} \rho_2(z) \log \rho_1(z) \mu(dz)$ is concave on the positive cone since log is concave and $\rho_2 \geq \underline{\rho}_2 > 0$ μ -a.e. Thus, $\rho_1 \mapsto \mathcal{F}_1(\rho_1)$ is concave on the convex set $\mathcal{D}_{\mu, \underline{\rho}_1}(\mathcal{Z})$. This concave coordinate geometry motivates the boundary-concentrating construction below.

Write any feasible $\rho_1 \in \mathcal{D}_{\mu, \underline{\rho}_1}(\mathcal{Z})$ as

$$\rho_1(z) = \underline{\rho}_1 + \eta_1(z), \quad \eta_1(z) \geq 0.$$

The excess probability mass is

$$M_{\text{ex}}^{(1)} = \int_{\mathcal{Z}} \eta_1(z) \mu(dz) = 1 - \underline{\rho}_1 \mu(\mathcal{Z}) > 0.$$

Let

$$v_* := \operatorname{ess\,inf}_{\mu} V_{1|2}.$$

Since $V_{1|2}(z) \geq v_*$ and $\log \rho_1(z) \geq \log \underline{\rho}_1$ μ -a.e., we have

$$\begin{aligned} \int_{\mathcal{Z}} \rho_1(z) V_{1|2}(z) \mu(dz) &= \underline{\rho}_1 \int_{\mathcal{Z}} V_{1|2}(z) \mu(dz) + \int_{\mathcal{Z}} \eta_1(z) V_{1|2}(z) \mu(dz) \\ &\geq \underline{\rho}_1 \int_{\mathcal{Z}} V_{1|2}(z) \mu(dz) + v_* M_{\text{ex}}^{(1)}, \end{aligned}$$

and

$$\int_{\mathcal{Z}} \rho_2(z) \log \rho_1(z) \mu(dz) \geq \int_{\mathcal{Z}} \rho_2(z) \log \underline{\rho}_1 \mu(dz) = \log \underline{\rho}_1.$$

Plugging these bounds into [Eq. \(51\)](#) yields

$$\mathcal{F}_1(\rho_1) \geq \frac{1}{2} \left(\underline{\rho}_1 \int_{\mathcal{Z}} V_{1|2}(z) \mu(dz) + v_* M_{\text{ex}}^{(1)} \right) + \frac{1}{2} \log \underline{\rho}_1 + C(\rho_2) =: \mathcal{F}_*. \quad (52)$$

Therefore,

$$\inf_{\rho_1 \in \mathcal{D}_{\mu, \underline{\rho}_1}(\mathcal{Z})} \mathcal{F}_1(\rho_1) \geq \mathcal{F}_*.$$

Step 2: Evaluate the coordinate map $\mathcal{F}_1$ at $\rho_1^{(\sigma)}$. For each $\sigma > 0$, define the sublevel set

$$\mathcal{W}_1^\sigma := \{z \in \mathcal{Z} : V_{1|2}(z) \leq v_* + \sigma\}.$$

By the definition of the essential infimum, $\mu(\mathcal{W}_1^\sigma) > 0$. Since μ is a non-atomic Riemannian volume measure, we may choose measurable subsets $\mathcal{S}_1^\sigma \subseteq \mathcal{W}_1^\sigma$ such that $0 < \mu(\mathcal{S}_1^\sigma) < \infty$ and $\mu(\mathcal{S}_1^\sigma) \rightarrow 0$ as $\sigma \rightarrow 0$. Define

$$\rho_1^{(\sigma)}(z) = \underline{\rho}_1 + \mathbb{1}_{\mathcal{S}_1^\sigma}(z) \frac{M_{\text{ex}}^{(1)}}{\mu(\mathcal{S}_1^\sigma)}.$$

Since $\rho_1^{(\sigma)} \geq \underline{\rho}_1$ and

$$\int_{\mathcal{Z}} \rho_1^{(\sigma)}(z) \mu(dz) = \underline{\rho}_1 \mu(\mathcal{Z}) + \frac{M_{\text{ex}}^{(1)}}{\mu(\mathcal{S}_1^\sigma)} \mu(\mathcal{S}_1^\sigma) = \underline{\rho}_1 \mu(\mathcal{Z}) + M_{\text{ex}}^{(1)} = 1,$$

we have $\rho_1^{(\sigma)} \in \mathcal{D}_{\mu, \underline{\rho}_1}(\mathcal{Z})$.

We now evaluate the coordinate map $\mathcal{F}_1$ at $\rho_1^{(\sigma)}$.

(i) *Linear term.* Since $\mathcal{S}_1^\sigma \subseteq \mathcal{W}_1^\sigma$, we have $V_{1|2}(z) \leq v_* + \sigma$ on $\mathcal{S}_1^\sigma$, and hence

$$\begin{aligned} \int_{\mathcal{Z}} \rho_1^{(\sigma)}(z) V_{1|2}(z) \mu(dz) &= \underline{\rho}_1 \int_{\mathcal{Z}} V_{1|2}(z) \mu(dz) + \frac{M_{\text{ex}}^{(1)}}{\mu(\mathcal{S}_1^\sigma)} \int_{\mathcal{S}_1^\sigma} V_{1|2}(z) \mu(dz) \\ &\leq \underline{\rho}_1 \int_{\mathcal{Z}} V_{1|2}(z) \mu(dz) + M_{\text{ex}}^{(1)}(v_* + \sigma). \end{aligned}$$

(ii) *Log term.* Split the domain as $\mathcal{Z} = (\mathcal{Z} \setminus \mathcal{S}_1^\sigma) \cup \mathcal{S}_1^\sigma$. Then

$$\int_{\mathcal{Z}} \rho_2(z) \log \rho_1^{(\sigma)}(z) \mu(dz) = \int_{\mathcal{Z} \setminus \mathcal{S}_1^\sigma} \rho_2(z) \log \underline{\rho}_1 \mu(dz) + \int_{\mathcal{S}_1^\sigma} \rho_2(z) \log \left(\underline{\rho}_1 + \frac{M_{\text{ex}}^{(1)}}{\mu(\mathcal{S}_1^\sigma)} \right) \mu(dz).$$

Define

$$m_\sigma := \int_{\mathcal{S}_1^\sigma} \rho_2(z) \mu(dz).$$

Since $\int_{\mathcal{Z}} \rho_2(z) \mu(dz) = 1$, the first term equals $(1 - m_\sigma) \log \underline{\rho}_1$. Therefore,

$$\int_{\mathcal{Z}} \rho_2(z) \log \rho_1^{(\sigma)}(z) \mu(dz) = \log \underline{\rho}_1 + m_\sigma L_\sigma,$$

where

$$L_\sigma := \log \left(1 + \frac{M_{\text{ex}}^{(1)}}{\underline{\rho}_1 \mu(\mathcal{S}_1^\sigma)} \right).$$

Combining the two terms gives

$$\mathcal{F}_1(\rho_1^{(\sigma)}) \leq \frac{1}{2} \left(\underline{\rho}_1 \int_{\mathcal{Z}} V_{1|2}(z) \mu(dz) + M_{\text{ex}}^{(1)}(v_* + \sigma) \right) + \frac{1}{2} \left(\log \underline{\rho}_1 + m_\sigma L_\sigma \right) + C(\rho_2). \quad (53)$$

Step 3: Conclude convergence. We first show that the additional log-penalty $m_\sigma L_\sigma$ vanishes as $\sigma \rightarrow 0$. Since $\rho_2 \in \mathcal{D}_{\mu, \rho_2}(\mathcal{Z}) \subset L^\infty(\mu)$,

$$m_\sigma = \int_{\mathcal{S}_1^\sigma} \rho_2(z) \mu(dz) \leq \|\rho_2\|_\infty \mu(\mathcal{S}_1^\sigma).$$

Hence

$$0 \leq m_\sigma L_\sigma \leq \|\rho_2\|_\infty \mu(\mathcal{S}_1^\sigma) \log \left(1 + \frac{M_{\text{ex}}^{(1)}}{\underline{\rho}_1 \mu(\mathcal{S}_1^\sigma)} \right).$$

Since $\mu(\mathcal{S}_1^\sigma) \rightarrow 0$ and $t \log(1 + c/t) \rightarrow 0$ as $t \downarrow 0$ for every fixed $c > 0$, it follows that

$$m_\sigma L_\sigma \rightarrow 0 \quad \text{as } \sigma \rightarrow 0.$$

Now subtract $\mathcal{F}_*$ from Eq. (53) to obtain

$$\mathcal{F}_1(\rho_1^{(\sigma)}) - \mathcal{F}_* \leq \frac{1}{2} \left(M_{\text{ex}}^{(1)} \sigma + m_\sigma L_\sigma \right).$$

Letting $\sigma \rightarrow 0$ gives

$$\limsup_{\sigma \rightarrow 0} \mathcal{F}_1(\rho_1^{(\sigma)}) \leq \mathcal{F}_*.$$

Meanwhile, Step 1 showed

$$\inf_{\rho_1 \in \mathcal{D}_{\mu, \underline{\rho}_1}(\mathcal{Z})} \mathcal{F}_1(\rho_1) \geq \mathcal{F}_*.$$

Since every $\rho_1^{(\sigma)}$ is feasible, we also have

$$\inf_{\rho_1 \in \mathcal{D}_{\mu, \underline{\rho}_1}(\mathcal{Z})} \mathcal{F}_1(\rho_1) \leq \mathcal{F}_1(\rho_1^{(\sigma)})$$

for every $\sigma > 0$. Therefore,

$$\inf_{\rho_1 \in \mathcal{D}_{\mu, \underline{\rho}_1}(\mathcal{Z})} \mathcal{F}_1(\rho_1) = \mathcal{F}_*, \quad \lim_{\sigma \rightarrow 0} \mathcal{F}_1(\rho_1^{(\sigma)}) = \inf_{\rho_1 \in \mathcal{D}_{\mu, \underline{\rho}_1}(\mathcal{Z})} \mathcal{F}_1(\rho_1).$$

Thus, the coordinate infimum is approached by boundary-concentrating infimizing sequences whose excess mass lies in shrinking low-effective-potential regions of

$$V_{1|2} = \frac{1}{\tau} U_{1+2} + \log \rho_2.$$

The ρ_2 -coordinate statement follows identically after swapping indices and using

$$V_{2|1} = \frac{1}{\tau} U_{2+1} + \log \rho_1.$$

□

D.3 Proof of Thm. 4.2

Theorem 4.2 (Low-temperature multimodal energy consistency). *Assume Asms. 2.1, 3.2 and 4.1. Fix any $(\theta, \phi) \in \Theta \times \Phi$, and assume $q_\theta \ll \mu$, $q_\phi \ll \mu$ with continuous densities ρ_θ, ρ_ϕ bounded below: $\inf_{z \in \mathcal{Z}} \rho_\theta(z) \geq \underline{\rho}_\theta > 0$, $\inf_{z \in \mathcal{Z}} \rho_\phi(z) \geq \underline{\rho}_\phi > 0$. Let $\varepsilon_{\text{kde}}^{(\theta)}(\tau) := \|\tilde{\rho}_{\theta, \tau} - \rho_\theta\|_\infty$ and $\varepsilon_{\text{kde}}^{(\phi)}(\tau) := \|\tilde{\rho}_{\phi, \tau} - \rho_\phi\|_\infty$. Then, there exists $\tau_0(\theta, \phi) > 0$ such that for all $0 < \tau \leq \tau_0(\theta, \phi)$,*

$$\left| \mathcal{J}_\tau^{\text{Sym}}(\theta, \phi) - \mathcal{F}_{\tau, \mathbf{U}_{\theta, \phi}}^{\text{Sym}}(\rho_\theta, \rho_\phi) \right| \leq \frac{\varepsilon_{\text{kde}}^{(\theta)}(\tau) + \varepsilon_{\text{kde}}^{(\phi)}(\tau)}{\min\{\underline{\rho}_\theta, \underline{\rho}_\phi\}},$$

where $\mathcal{F}_{\tau, \mathbf{U}_{\theta, \phi}}^{\text{Sym}}(\rho_\theta, \rho_\phi)$ is the intrinsic multimodal energy from Def. 4.3, evaluated on $\mathcal{D}_{\mu, \underline{\rho}_\theta}(\mathcal{Z}) \times \mathcal{D}_{\mu, \underline{\rho}_\phi}(\mathcal{Z})$, with $\mathbf{U}_{\theta, \phi} := (U_{\theta \rightarrow \phi}, U_{\phi \rightarrow \theta})$. In particular,

$$\left| \mathcal{J}_\tau^{\text{Sym}}(\theta, \phi) - \mathcal{F}_{\tau, \mathbf{U}_{\theta, \phi}}^{\text{Sym}}(\rho_\theta, \rho_\phi) \right| \rightarrow 0, \quad \text{as } \tau \rightarrow 0^+.$$

☞ **Proof intuition.** This proof establishes that, in the low-temperature limit, the cross-modal parametric energy approaches the intrinsic dual-density free energy for each fixed pair (θ, ϕ) . The argument relies on an algebraic cancellation and the stability of kernel smoothing: **(i)** We first decompose both the parametric and intrinsic energies into their potential, entropy, and cross-modal divergence terms. Their potential and entropy terms cancel in the difference, leaving only discrepancies between KL divergences involving the true marginal densities and their kernel-smoothed counterparts. **(ii)** Because the true densities are bounded away from zero and the smoothed densities converge uniformly to them, all relevant logarithms remain in a Lipschitz regime for sufficiently small τ . This allows the remaining log-ratio perturbations to be bounded directly by the KDE sup-errors of the two modalities. **(iii)** Since both KDE mismatches vanish as $\tau \rightarrow 0^+$, the discrepancy between the parametric and intrinsic energy values also vanishes for each fixed (θ, ϕ) .

Proof. Fix any $(\theta, \phi) \in \Theta \times \Phi$. Under Asm. 4.1, sim is bounded and Borel measurable, hence both cross-modal potentials $U_{\theta \rightarrow \phi}$ and $U_{\phi \rightarrow \theta}$ from Def. 4.1 are Borel measurable and bounded. Moreover, since $\mathcal{Z}$ is compact and ρ_θ, ρ_ϕ are continuous with $\rho_\theta \geq \underline{\rho}_\theta > 0$ and $\rho_\phi \geq \underline{\rho}_\phi > 0$, both densities are bounded above and below. Hence $\log \rho_\theta$ and $\log \rho_\phi$ are bounded, and therefore $D_{\text{KL}}(\rho_\theta \| \rho_\phi)$ and $D_{\text{KL}}(\rho_\phi \| \rho_\theta)$ are finite. Consequently, the multimodal functional $\mathcal{F}_{\tau, \mathbf{U}_{\theta, \phi}}^{\text{Sym}}(\rho_\theta, \rho_\phi)$ is well-defined.

Step 1: Functional discrepancy reduces to log-ratio perturbations. Recall the directional multimodal parametric energies from Def. 4.2:

$$\mathcal{J}_\tau^{x \rightarrow y}(\theta, \phi) = \frac{1}{\tau} \mathcal{U}^{\theta \rightarrow \phi}(q_\theta) - H_\times(q_\theta, \tilde{\rho}_{\phi, \tau}), \quad \mathcal{J}_\tau^{y \rightarrow x}(\phi, \theta) = \frac{1}{\tau} \mathcal{U}^{\phi \rightarrow \theta}(q_\phi) - H_\times(q_\phi, \tilde{\rho}_{\theta, \tau}),$$

and

$$\mathcal{J}_\tau^{\text{Sym}}(\theta, \phi) = \frac{1}{2} (\mathcal{J}_\tau^{x \rightarrow y}(\theta, \phi) + \mathcal{J}_\tau^{y \rightarrow x}(\phi, \theta)).$$

Using

$$H_\times(q_1, \rho_2) = H(\rho_1) + D_{\text{KL}}(\rho_1 \| \rho_2), \quad \rho_1 := \frac{dq_1}{d\mu},$$

we can rewrite

$$\mathcal{J}_\tau^{\text{Sym}}(\theta, \phi) = \frac{1}{2} (\mathcal{F}_{\tau, U_{\theta \rightarrow \phi}}(\rho_\theta) + \mathcal{F}_{\tau, U_{\phi \rightarrow \theta}}(\rho_\phi)) - \frac{1}{2} (D_{\text{KL}}(\rho_\theta \| \tilde{\rho}_{\phi, \tau}) + D_{\text{KL}}(\rho_\phi \| \tilde{\rho}_{\theta, \tau})). \quad (54)$$

On the other hand, by Def. 4.3, with

$$D_{\text{KL}}^{\text{Sym}}(\rho_\theta, \rho_\phi) := \frac{1}{2} (D_{\text{KL}}(\rho_\theta \| \rho_\phi) + D_{\text{KL}}(\rho_\phi \| \rho_\theta)),$$

we have

$$\mathcal{F}_{\tau, \mathbf{U}_{\theta, \phi}}^{\text{Sym}}(\rho_\theta, \rho_\phi) = \frac{1}{2} (\mathcal{F}_{\tau, U_{\theta \rightarrow \phi}}(\rho_\theta) + \mathcal{F}_{\tau, U_{\phi \rightarrow \theta}}(\rho_\phi)) - D_{\text{KL}}^{\text{Sym}}(\rho_\theta, \rho_\phi). \quad (55)$$

Subtracting Eq. (54) from Eq. (55), the potential and entropy terms cancel, giving the exact identity

$$\mathcal{F}_{\tau, \mathbf{U}_{\theta, \phi}}^{\text{Sym}}(\rho_{\theta}, \rho_{\phi}) - \mathcal{J}_{\tau}^{\text{Sym}}(\theta, \phi) = \frac{1}{2} \left(D_{\text{KL}}(\rho_{\theta} \parallel \tilde{\rho}_{\phi, \tau}) - D_{\text{KL}}(\rho_{\theta} \parallel \rho_{\phi}) + D_{\text{KL}}(\rho_{\phi} \parallel \tilde{\rho}_{\theta, \tau}) - D_{\text{KL}}(\rho_{\phi} \parallel \rho_{\theta}) \right). \quad (56)$$

For any μ -densities p, r_1, r_2 with $r_1, r_2 > 0$ μ -a.e., the elementary difference identity

$$D_{\text{KL}}(p \parallel r_1) - D_{\text{KL}}(p \parallel r_2) = \int_{\mathcal{Z}} p(z) \log \frac{r_2(z)}{r_1(z)} \mu(dz)$$

holds. Since $\kappa_{\tau} > 0$, the smoothed densities $\tilde{\rho}_{\theta, \tau}$ and $\tilde{\rho}_{\phi, \tau}$ are strictly positive. Applying the identity with $(p, r_1, r_2) = (\rho_{\theta}, \tilde{\rho}_{\phi, \tau}, \rho_{\phi})$ and $(p, r_1, r_2) = (\rho_{\phi}, \tilde{\rho}_{\theta, \tau}, \rho_{\theta})$ gives

$$\mathcal{F}_{\tau, \mathbf{U}_{\theta, \phi}}^{\text{Sym}}(\rho_{\theta}, \rho_{\phi}) - \mathcal{J}_{\tau}^{\text{Sym}}(\theta, \phi) = \frac{1}{2} \left(\int_{\mathcal{Z}} \rho_{\theta}(z) \log \frac{\rho_{\phi}(z)}{\tilde{\rho}_{\phi, \tau}(z)} \mu(dz) + \int_{\mathcal{Z}} \rho_{\phi}(z) \log \frac{\rho_{\theta}(z)}{\tilde{\rho}_{\theta, \tau}(z)} \mu(dz) \right). \quad (57)$$

Step 2: Nonasymptotic bound by KDE sup-errors. By the lower bounds $\rho_{\theta} \geq \underline{\rho}_{\theta}$ and $\rho_{\phi} \geq \underline{\rho}_{\phi}$ and **Lem. B.3** applied separately to ρ_{θ} and ρ_{ϕ} , there exists $\tau_0(\theta, \phi) > 0$ such that, for all $0 < \tau \leq \tau_0(\theta, \phi)$,

$$\varepsilon_{\text{kde}}^{(\theta)}(\tau) := \|\tilde{\rho}_{\theta, \tau} - \rho_{\theta}\|_{\infty} \leq \frac{\underline{\rho}_{\theta}}{2}, \quad \varepsilon_{\text{kde}}^{(\phi)}(\tau) := \|\tilde{\rho}_{\phi, \tau} - \rho_{\phi}\|_{\infty} \leq \frac{\underline{\rho}_{\phi}}{2}.$$

Hence, for every $z \in \mathcal{Z}$ and all such τ ,

$$\tilde{\rho}_{\theta, \tau}(z) \geq \frac{\underline{\rho}_{\theta}}{2}, \quad \tilde{\rho}_{\phi, \tau}(z) \geq \frac{\underline{\rho}_{\phi}}{2}. \quad (58)$$

On $[\underline{\rho}_{\phi}/2, \infty)$, the map $\log(\cdot)$ is Lipschitz with constant $2/\underline{\rho}_{\phi}$. Therefore, for every $z \in \mathcal{Z}$,

$$|\log \rho_{\phi}(z) - \log \tilde{\rho}_{\phi, \tau}(z)| \leq \frac{2}{\underline{\rho}_{\phi}} |\rho_{\phi}(z) - \tilde{\rho}_{\phi, \tau}(z)| \leq \frac{2 \varepsilon_{\text{kde}}^{(\phi)}(\tau)}{\underline{\rho}_{\phi}}.$$

Equivalently,

$$\left| \log \frac{\rho_{\phi}(z)}{\tilde{\rho}_{\phi, \tau}(z)} \right| \leq \frac{2 \varepsilon_{\text{kde}}^{(\phi)}(\tau)}{\underline{\rho}_{\phi}}. \quad (59)$$

Analogously,

$$\left| \log \frac{\rho_{\theta}(z)}{\tilde{\rho}_{\theta, \tau}(z)} \right| \leq \frac{2 \varepsilon_{\text{kde}}^{(\theta)}(\tau)}{\underline{\rho}_{\theta}}. \quad (60)$$

Using Eq. (57), together with

$$\int_{\mathcal{Z}} \rho_{\theta}(z) \mu(dz) = \int_{\mathcal{Z}} \rho_{\phi}(z) \mu(dz) = 1,$$

we obtain

$$\begin{aligned} \left| \mathcal{F}_{\tau, \mathbf{U}_{\theta, \phi}}^{\text{Sym}}(\rho_{\theta}, \rho_{\phi}) - \mathcal{J}_{\tau}^{\text{Sym}}(\theta, \phi) \right| &\leq \frac{1}{2} \left(\int_{\mathcal{Z}} \rho_{\theta}(z) \left| \log \frac{\rho_{\phi}(z)}{\tilde{\rho}_{\phi, \tau}(z)} \right| \mu(dz) + \int_{\mathcal{Z}} \rho_{\phi}(z) \left| \log \frac{\rho_{\theta}(z)}{\tilde{\rho}_{\theta, \tau}(z)} \right| \mu(dz) \right) \\ &\leq \frac{1}{2} \left(\left\| \log \frac{\rho_{\phi}}{\tilde{\rho}_{\phi, \tau}} \right\|_{\infty} + \left\| \log \frac{\rho_{\theta}}{\tilde{\rho}_{\theta, \tau}} \right\|_{\infty} \right) \\ &\leq \frac{1}{2} \left(\frac{2 \varepsilon_{\text{kde}}^{(\phi)}(\tau)}{\underline{\rho}_{\phi}} + \frac{2 \varepsilon_{\text{kde}}^{(\theta)}(\tau)}{\underline{\rho}_{\theta}} \right) \\ &\leq \frac{\varepsilon_{\text{kde}}^{(\theta)}(\tau) + \varepsilon_{\text{kde}}^{(\phi)}(\tau)}{\min\{\underline{\rho}_{\theta}, \underline{\rho}_{\phi}\}}. \end{aligned}$$

This proves the claimed nonasymptotic bound for all $0 < \tau \leq \tau_0(\theta, \phi)$.

Step 3: Vanishing discrepancy as $\tau \rightarrow 0^+$. By [Lem. B.3](#) applied separately to ρ_θ and ρ_ϕ under [Asm. 3.2](#),

$$\varepsilon_{\text{kde}}^{(\theta)}(\tau) \rightarrow 0, \quad \varepsilon_{\text{kde}}^{(\phi)}(\tau) \rightarrow 0 \quad \text{as } \tau \rightarrow 0^+.$$

Therefore, the bound in Step 2 implies

$$\left| \mathcal{J}_\tau^{\text{Sym}}(\theta, \phi) - \mathcal{F}_{\tau, \mathbf{U}_{\theta, \phi}}^{\text{Sym}}(\rho_\theta, \rho_\phi) \right| \rightarrow 0 \quad \text{as } \tau \rightarrow 0^+,$$

for every fixed $(\theta, \phi) \in \Theta \times \Phi$. $\square$

E Extended Formal Statements

This section provides extended formal statements by connecting the intrinsic geometric insights to parameterized training objectives, when the encoder families are sufficiently expressive.

E.1 Unimodal Parametric Inheritance and Emergence of Uniformity

This subsection formalizes the intuition used in [§ 3.2](#): unimodal InfoNCE inherits the intrinsic Gibbs geometry when (i) the kernel smoothing bias is small and (ii) the encoder family is expressive enough to (approximately) realize the intrinsic Gibbs equilibrium at the best attainable potential level.

Recap of objects. Fix $\tau > 0$. For each $\theta \in \Theta$, recall the induced representation law $q_\theta \in \mathcal{P}(\mathcal{Z})$ and encoded positive-pair law $\pi_{\theta\theta} \in \mathcal{P}(\mathcal{Z} \times \mathcal{Z})$. Under the same setting of [Thm. 3.2](#), assume $q_\theta \ll \mu$ and write $\rho_\theta(z) := q_\theta(dz)/\mu(dz)$. Let U_θ be the alignment potential field induced by $\pi_{\theta\theta}$ ([Def. 3.1](#)). The intrinsic functional is $\mathcal{F}_{\tau, U_\theta}(\rho)$ ([Eq. \(7\)](#)), and the parametric energy is $\mathcal{J}_\tau(\theta)$ ([Def. 3.2](#)).

Free-energy gap equals KL divergence. From [Prop. 3.1](#), for any fixed potential U and temperature τ , define the partition function

$$Z_\tau := \int_{\mathcal{Z}} \exp(-U(z)/\tau) \mu(dz), \quad (61)$$

and the associated Gibbs density

$$\rho^*(z) := \frac{\exp(-U(z)/\tau)}{Z_\tau}. \quad (62)$$

Then, for any $\rho \in \mathcal{D}_\mu(\mathcal{Z})$ with $\rho \log(\rho/\rho^*) \in L^1(\mu)$,

$$\mathcal{F}_{\tau, U}(\rho) + \log Z_\tau = D_{\text{KL}}(\rho \parallel \rho^*) \geq 0. \quad (63)$$

In particular, ρ^* is the unique minimizer of $\mathcal{F}_{\tau, U}$, and the free-energy suboptimality controls proximity to the Gibbs equilibrium via KL divergence.

Remark E.1 (Proof of [Eq. \(63\)](#)). Expanding

$$D_{\text{KL}}(\rho \parallel \rho^*) = \int_{\mathcal{Z}} \rho(z) \log \frac{\rho(z)}{\rho^*(z)} \mu(dz)$$

and substituting $\log \rho^* = -(U/\tau) - \log Z_\tau$ yields [Eq. \(63\)](#).

Expressivity at the best attainable potential level. The intrinsic minimum of $\mathcal{F}_{\tau, U_\theta}$ equals $-\log Z_{\theta, \tau}$ by [Eq. \(63\)](#), where

$$Z_{\theta, \tau} := \int_{\mathcal{Z}} \exp(-U_\theta(z)/\tau) \mu(dz)$$

is the normalization constant of the Gibbs density. Whenever the minimum is attained, define the *best attainable potential level* at temperature τ by choosing any

$$\hat{\theta}_\tau \in \arg \min_{\theta \in \Theta} (-\log Z_{\theta, \tau}). \quad (64)$$

Assumption E.1 (Expressivity at the best level). There exists a function $\varepsilon_{\text{real}}(\tau) \rightarrow 0$ as $\tau \rightarrow 0^+$ such that, for $\hat{\theta}_\tau$ in [Eq. \(64\)](#),

$$\mathcal{F}_{\tau, U_{\hat{\theta}_\tau}}(\rho_{\hat{\theta}_\tau}) + \log Z_{\hat{\theta}_\tau, \tau} \leq \varepsilon_{\text{real}}(\tau). \quad (65)$$

Equivalently, at the best attainable potential level, the encoder-induced density $\rho_{\hat{\theta}_\tau}$ attains the intrinsic minimum of $\mathcal{F}_{\tau, U_{\hat{\theta}_\tau}}$ up to error $\varepsilon_{\text{real}}(\tau)$.

Ground-state inheritance under encoder expressivity. Under [Asm. 3.2](#) and mild regularity conditions, [Thm. 3.2](#) bounds the gap between $\mathcal{J}_\tau(\theta)$ and $\mathcal{F}_{\tau, U_\theta}(\rho_\theta)$ by the KDE mismatch $\varepsilon_{\text{kde}}^{(\theta)}(\tau)$. By [Lem. B.3](#), this mismatch vanishes for each fixed θ . Since the relevant optimizers may depend on τ , we state the following inheritance result along optimizing sequences for which the corresponding energy approximation errors vanish at the evaluated temperatures.

Corollary E.1 (Unimodal parametric inheritance along optimizing sequences). *For each $\tau > 0$, assume the following minimizers exist, and let*

$$\theta_\tau^* \in \arg \min_{\theta \in \Theta} \mathcal{J}_\tau(\theta), \quad \hat{\theta}_\tau \in \arg \min_{\theta \in \Theta} (-\log Z_{\theta, \tau}).$$

Assume the conditions of [Thm. 3.2](#) hold for θ_τ^* and $\hat{\theta}_\tau$ for all sufficiently small $\tau > 0$. Assume additionally that the expressivity condition in [Asm. E.1](#) holds along the sequence $\hat{\theta}_\tau$, and that the energy approximation errors satisfy

$$\eta_\tau(\theta_\tau^*) + \eta_\tau(\hat{\theta}_\tau) \rightarrow 0 \quad \text{as } \tau \rightarrow 0^+, \quad \text{where } \eta_\tau(\theta) := |\mathcal{J}_\tau(\theta) - \mathcal{F}_{\tau, U_\theta}(\rho_\theta)|.$$

Further, let $\rho_{\theta_\tau^*}^{*, \tau}$ denote the Gibbs density associated with $U_{\theta_\tau^*}$ at temperature τ :

$$\rho_{\theta_\tau^*}^{*, \tau}(z) := \frac{\exp(-U_{\theta_\tau^*}(z)/\tau)}{Z_{\theta_\tau^*, \tau}}, \quad Z_{\theta_\tau^*, \tau} := \int_{\mathcal{Z}} \exp(-U_{\theta_\tau^*}(z')/\tau) \mu(dz').$$

Then,

$$D_{\text{KL}}(\rho_{\theta_\tau^*} \| \rho_{\theta_\tau^*}^{*, \tau}) \rightarrow 0 \quad \text{and} \quad \|\rho_{\theta_\tau^*} - \rho_{\theta_\tau^*}^{*, \tau}\|_{\text{TV}} \rightarrow 0.$$

Consequently, the learned representation measures inherit any low-temperature concentration property of the associated Gibbs sequence. In particular, if for some fixed $\sigma > 0$ and some family of open sets $\mathcal{O}_\tau \subset \mathcal{Z}$ satisfying

$$\mathcal{O}_\tau \supset \mathcal{W}_{\theta_\tau^*}^\sigma := \left\{ z \in \mathcal{Z} : U_{\theta_\tau^*}(z) \leq \text{ess inf}_\mu U_{\theta_\tau^*} + \sigma \right\},$$

the Gibbs sequence satisfies $(\rho_{\theta_\tau^*}^{*, \tau} \mu)(\mathcal{Z} \setminus \mathcal{O}_\tau) \rightarrow 0$, then

$$q_{\theta_\tau^*}(\mathcal{Z} \setminus \mathcal{O}_\tau) \rightarrow 0.$$

☞ **Proof intuition.** The goal is to show that a sequence of parametric minimizers θ_τ^* tracks the Gibbs equilibrium induced by its own alignment potential as $\tau \rightarrow 0^+$. The proof has three steps. **(i)** For any fixed induced potential field U_θ , the intrinsic free-energy gap equals the KL divergence from the corresponding Gibbs density. **(ii)** By the expressivity condition, the comparator $\hat{\theta}_\tau$ can approach the intrinsic optimum at the best attainable potential level. Since θ_τ^* minimizes the parametric energy $\mathcal{J}_\tau$, and since the sequence-level energy approximation errors vanish for both θ_τ^* and $\hat{\theta}_\tau$, the intrinsic free-energy gap of $\rho_{\theta_\tau^*}$ also vanishes. Hence $D_{\text{KL}}(\rho_{\theta_\tau^*} \| \rho_{\theta_\tau^*}^{*, \tau}) \rightarrow 0$. **(iii)** Pinsker’s inequality gives convergence in total variation, so any low-temperature concentration property of the Gibbs sequence transfers to the learned representation measures.

Proof. Step 1: Well-defined “best achievable” equilibrium. We first isolate the intrinsic quantity that governs the “best achievable” equilibrium value at temperature τ . Fix any $\theta \in \Theta$ and $\tau > 0$. By the standing setup, $\mathcal{Z}$ is compact with $\mu(\mathcal{Z}) < \infty$; and by [Asm. 3.1](#), $\text{sim} : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$ is continuous. Hence, by the extreme value theorem, sim is bounded on $\mathcal{Z} \times \mathcal{Z}$:

$$\text{Sim}_{\max} := \max_{z, w \in \mathcal{Z}} \text{sim}(z, w) < \infty.$$

The induced alignment potential field ([Def. 3.1](#)) is

$$U_\theta(z) = - \int_{\mathcal{Z}} \text{sim}(z, w) \nu_{\theta, z}(dw),$$

hence

$$U_\theta(z) \geq -\text{Sim}_{\max} \quad \text{for all } z \in \mathcal{Z}.$$

Therefore, for any fixed $\tau > 0$,

$$0 < \exp\left(-\frac{U_\theta(z)}{\tau}\right) \leq \exp\left(\frac{\text{Sim}_{\max}}{\tau}\right), \quad z \in \mathcal{Z},$$

and thus

$$0 < Z_{\theta,\tau} := \int_{\mathcal{Z}} \exp\left(-\frac{U_\theta(z)}{\tau}\right) \mu(dz) \leq \mu(\mathcal{Z}) \exp\left(\frac{\text{Sim}_{\max}}{\tau}\right) < \infty.$$

Thus, the representation energy

$$\mathcal{F}_{\tau,U_\theta}(\rho) = \frac{1}{\tau} \int_{\mathcal{Z}} U_\theta(z) \rho(z) \mu(dz) - H(\rho)$$

admits the unique Gibbs minimizer $\rho_\theta^{*,\tau}(z) = \frac{\exp(-U_\theta(z)/\tau)}{Z_{\theta,\tau}}$ by [Prop. 3.1](#). The corresponding minimum is

$$\inf_{\rho \in \mathcal{D}_\mu(\mathcal{Z})} \mathcal{F}_{\tau,U_\theta}(\rho) = \mathcal{F}_{\tau,U_\theta}(\rho_\theta^{*,\tau}) = -\log Z_{\theta,\tau}. \quad (66)$$

We interpret $-\log Z_{\theta,\tau}$ as the lowest representation energy level induced by the potential field U_θ , which is the best intrinsic value attainable by any density in $\mathcal{D}_\mu(\mathcal{Z})$ once U_θ is fixed.

Step 2: Representation energy gap equals a KL divergence. Fix θ and τ . Recall that, by [Prop. 3.1](#), the unique Gibbs equilibrium is

$$\rho_\theta^{*,\tau}(z) = \frac{\exp(-U_\theta(z)/\tau)}{Z_{\theta,\tau}},$$

so

$$\log \rho_\theta^{*,\tau}(z) = -\frac{U_\theta(z)}{\tau} - \log Z_{\theta,\tau}.$$

Applying [Eq. \(63\)](#) with $\rho = \rho_\theta$ and $\rho^* = \rho_\theta^{*,\tau}$ yields

$$D_{\text{KL}}(\rho_\theta \| \rho_\theta^{*,\tau}) = \mathcal{F}_{\tau,U_\theta}(\rho_\theta) + \log Z_{\theta,\tau}.$$

Finally, by [Eq. \(66\)](#),

$$\inf_{\rho \in \mathcal{D}_\mu(\mathcal{Z})} \mathcal{F}_{\tau,U_\theta}(\rho) = \mathcal{F}_{\tau,U_\theta}(\rho_\theta^{*,\tau}) = -\log Z_{\theta,\tau},$$

hence

$$\mathcal{F}_{\tau,U_\theta}(\rho_\theta) - \inf_{\rho \in \mathcal{D}_\mu(\mathcal{Z})} \mathcal{F}_{\tau,U_\theta}(\rho) = \mathcal{F}_{\tau,U_\theta}(\rho_\theta) + \log Z_{\theta,\tau} = D_{\text{KL}}(\rho_\theta \| \rho_\theta^{*,\tau}). \quad (67)$$

Step 3: Transfer intrinsic optimality to the optimizing sequence. For each $\tau > 0$, let

$$\theta_\tau^* \in \arg \min_{\theta \in \Theta} \mathcal{J}_\tau(\theta), \quad \hat{\theta}_\tau \in \arg \min_{\theta \in \Theta} (-\log Z_{\theta,\tau}).$$

Define

$$\eta_\tau(\theta) := |\mathcal{J}_\tau(\theta) - \mathcal{F}_{\tau,U_\theta}(\rho_\theta)|.$$

By the sequence condition in the statement,

$$\eta_\tau(\theta_\tau^*) + \eta_\tau(\hat{\theta}_\tau) \rightarrow 0.$$

By optimality,

$$\mathcal{J}_\tau(\theta_\tau^*) \leq \mathcal{J}_\tau(\hat{\theta}_\tau),$$

hence

$$\begin{aligned} \mathcal{F}_{\tau,U_{\theta_\tau^*}}(\rho_{\theta_\tau^*}) &\leq \mathcal{J}_\tau(\theta_\tau^*) + \eta_\tau(\theta_\tau^*) \\ &\leq \mathcal{J}_\tau(\hat{\theta}_\tau) + \eta_\tau(\theta_\tau^*) \\ &\leq \mathcal{F}_{\tau,U_{\hat{\theta}_\tau}}(\rho_{\hat{\theta}_\tau}) + \eta_\tau(\theta_\tau^*) + \eta_\tau(\hat{\theta}_\tau). \end{aligned}$$

Adding $\log Z_{\theta_\tau^*, \tau}$ and using

$$\log Z_{\theta_\tau^*, \tau} \leq \log Z_{\hat{\theta}_\tau, \tau}$$

by the definition of $\hat{\theta}_\tau$, we obtain

$$\mathcal{F}_{\tau, U_{\theta_\tau^*}}(\rho_{\theta_\tau^*}) + \log Z_{\theta_\tau^*, \tau} \leq \mathcal{F}_{\tau, U_{\hat{\theta}_\tau}}(\rho_{\hat{\theta}_\tau}) + \log Z_{\hat{\theta}_\tau, \tau} + \eta_\tau(\theta_\tau^*) + \eta_\tau(\hat{\theta}_\tau).$$

Using [Eq. \(65\)](#) along the sequence $\hat{\theta}_\tau$ gives

$$\mathcal{F}_{\tau, U_{\theta_\tau^*}}(\rho_{\theta_\tau^*}) + \log Z_{\theta_\tau^*, \tau} \leq \varepsilon_{\text{real}}(\tau) + \eta_\tau(\theta_\tau^*) + \eta_\tau(\hat{\theta}_\tau).$$

By [Eq. \(67\)](#), applied at $\theta = \theta_\tau^*$,

$$D_{\text{KL}}(\rho_{\theta_\tau^*} \| \rho_{\theta_\tau^*}^{*, \tau}) \leq \varepsilon_{\text{real}}(\tau) + \eta_\tau(\theta_\tau^*) + \eta_\tau(\hat{\theta}_\tau) \xrightarrow{\tau \rightarrow 0^+} 0.$$

Therefore, by Pinsker's inequality,

$$\|\rho_{\theta_\tau^*} - \rho_{\theta_\tau^*}^{*, \tau}\|_{\text{TV}} \leq \sqrt{\frac{1}{2} D_{\text{KL}}(\rho_{\theta_\tau^*} \| \rho_{\theta_\tau^*}^{*, \tau})} \xrightarrow{\tau \rightarrow 0^+} 0. \quad (68)$$

Now let $\mathcal{O}_\tau$ be any family of open sets satisfying

$$\mathcal{O}_\tau \supset \mathcal{W}_{\theta_\tau^*}^\sigma \quad \text{and} \quad (\rho_{\theta_\tau^*}^{*, \tau} \mu)(\mathcal{Z} \setminus \mathcal{O}_\tau) \rightarrow 0.$$

By the variational characterization of total variation, for any measurable $\mathcal{A} \subseteq \mathcal{Z}$,

$$|q_{\theta_\tau^*}(\mathcal{A}) - (\rho_{\theta_\tau^*}^{*, \tau} \mu)(\mathcal{A})| \leq \|\rho_{\theta_\tau^*} - \rho_{\theta_\tau^*}^{*, \tau}\|_{\text{TV}}.$$

Taking $\mathcal{A} = \mathcal{Z} \setminus \mathcal{O}_\tau$ gives

$$q_{\theta_\tau^*}(\mathcal{Z} \setminus \mathcal{O}_\tau) \leq (\rho_{\theta_\tau^*}^{*, \tau} \mu)(\mathcal{Z} \setminus \mathcal{O}_\tau) + \|\rho_{\theta_\tau^*} - \rho_{\theta_\tau^*}^{*, \tau}\|_{\text{TV}} \xrightarrow{\tau \rightarrow 0^+} 0.$$

This proves the claimed inherited concentration. $\square$

Remark E.2 (Why this implies uniformity is a tie-breaker). Inside the aligned basin

$$\{z : U_\theta(z) \approx \text{ess inf}_\mu U_\theta\},$$

the potential term in $\mathcal{F}_{\tau, U_\theta}$ is nearly constant. In that regime, minimizing $\mathcal{F}_{\tau, U_\theta}$ reduces approximately to maximizing entropy subject to the remaining constraints imposed by the encoder family, so the selected configuration is the most internally dispersed distribution compatible with being well-aligned. This is the precise sense in which “uniformity” in unimodal contrastive learning is not a competing global force but an entropic tie-breaker within the alignment basin.

E.2 Multimodal Parametric Inheritance

This subsection formalizes the sense in which the practical multimodal energy $\mathcal{J}_\tau^{\text{Sym}}(\theta, \phi)$ inherits the intrinsic barrier-driven coordinate geometry of $\mathcal{F}_{\tau, \mathbf{U}}^{\text{Sym}}$ when the encoder family is sufficiently expressive. Concretely, fixing one modality (say ϕ), we show that if the image encoder family can realize (up to a vanishing gap) the intrinsic coordinate infimum against the peer density, then any parametric minimizer concentrates its *excess* mass near minimizers of the effective barrier potential $V_{\theta|\phi} = \frac{1}{\tau} U_{\theta \rightarrow \phi} + \log \rho_\phi$.

Background-plus-excess decomposition. Given a density floor $\underline{\rho}_i > 0$ with $\underline{\rho}_i \mu(\mathcal{Z}) < 1$, any feasible density can be decomposed as

$$\rho_i(z) = \underline{\rho}_i + M_{\text{ex}}^{(i)} \bar{\rho}_i(z), \quad \text{with } M_{\text{ex}}^{(i)} := 1 - \underline{\rho}_i \mu(\mathcal{Z}), \quad \text{satisfying } \bar{\rho}_i \geq 0, \quad \int_{\mathcal{Z}} \bar{\rho}_i(z) \mu(\text{d}z) = 1.$$

Here $\underline{\rho}_i$ plays the role of a fixed background mass and $\bar{\rho}_i = (\rho_i - \underline{\rho}_i)/M_{\text{ex}}^{(i)}$ is the normalized distribution of the remaining excess budget $M_{\text{ex}}^{(i)}$.

Corollary E.2 (Multimodal parametric inheritance under model expressivity). *Assume the setting of Thm. 4.2. Fix a peer encoder $\phi \in \Phi$ and, for each $\tau > 0$, assume a minimizer exists and let $\theta_\tau^* \in \arg \min_{\theta \in \Theta} \mathcal{J}_\tau^{\text{Sym}}(\theta, \phi)$. Write the density floor parameterized by θ_τ^* as $\underline{\rho}_{\theta_\tau^*}$ with $\underline{\rho}_{\theta_\tau^*} \mu(\mathcal{Z}) < 1$, and define the set of strictly positive μ -densities bounded below by $\underline{\rho}_{\theta_\tau^*}$ as*

$$\mathcal{D}_{\mu, \underline{\rho}_{\theta_\tau^*}}(\mathcal{Z}) := \{\rho \in \mathcal{D}_\mu(\mathcal{Z}) \cap L^\infty(\mu) : \rho \geq \underline{\rho}_{\theta_\tau^*} \mu\text{-a.e.}\},$$

and assume that the corresponding excess budget remains uniformly nondegenerate for all sufficiently small τ ,

$$M_{\text{ex}}^{(\theta_\tau^*)} := 1 - \underline{\rho}_{\theta_\tau^*} \mu(\mathcal{Z}) \geq m_{\text{ex}}, \quad \text{for some } m_{\text{ex}} > 0.$$

Define the normalized excess density as $\bar{\rho}_{\theta_\tau^*}(z) := \rho_{\theta_\tau^*}(z) - \underline{\rho}_{\theta_\tau^*} / M_{\text{ex}}^{(\theta_\tau^*)}$. Further, assume the encoder family $\{f_\theta\}_\Theta$ is sufficiently expressive: there exists $\varepsilon_{\text{real}}(\tau) \rightarrow 0$ as $\tau \downarrow 0^+$ such that

$$\mathcal{F}_{\tau, \mathbf{U}_{\theta_\tau^*, \phi}}^{\text{Sym}}(\rho_{\theta_\tau^*}, \rho_\phi) \leq \inf_{\rho \in \mathcal{D}_{\mu, \underline{\rho}_{\theta_\tau^*}}(\mathcal{Z})} \mathcal{F}_{\tau, \mathbf{U}_{\theta_\tau^*, \phi}}^{\text{Sym}}(\rho, \rho_\phi) + \varepsilon_{\text{real}}(\tau). \quad (69)$$

Let $V_{\theta_\tau^*|\phi}(z) := \frac{1}{\tau} U_{\theta_\tau^*|\phi}(z) + \log \rho_\phi(z)$ be the effective potential field parameterized by (θ_τ^*, ϕ) . For $\sigma > 0$, define

$$\mathcal{W}_{\theta_\tau^*}^\sigma := \{z \in \mathcal{Z} : V_{\theta_\tau^*|\phi}(z) \leq \text{ess inf}_\mu V_{\theta_\tau^*|\phi} + \sigma\}.$$

Then, for every $\sigma > 0$,

$$\bar{\rho}_{\theta_\tau^*} \mu(\mathcal{Z} \setminus \mathcal{W}_{\theta_\tau^*}^\sigma) \leq \frac{2\varepsilon_{\text{real}}(\tau)}{\sigma M_{\text{ex}}^{(\theta_\tau^*)}}. \quad (70)$$

Moreover, if $\sigma = \sigma(\tau) \downarrow 0$ satisfies $\varepsilon_{\text{real}}(\tau)/\sigma \rightarrow 0$, then

$$\bar{\rho}_{\theta_\tau^*} \mu(\mathcal{Z} \setminus \mathcal{W}_{\theta_\tau^*}^{\sigma(\tau)}) \rightarrow 0.$$

A symmetric statement holds for optimizing in ϕ with θ fixed.

Proof intuition. The proof proceeds by decoupling the induced density $\rho_{\theta_\tau^*}$ into a fixed background floor $\underline{\rho}_{\theta_\tau^*}$ and a free excess mass $M_{\text{ex}}^{(\theta_\tau^*)}$. By the expressivity assumption, the parameterized model can closely approach the coordinate infimum of the intrinsic functional. We construct a sequence of infimizing competitors that concentrate all their excess mass inside shrinking sublevel sets of the effective barrier potential $V_{\theta_\tau^*|\phi}$, driving the cross-entropy term toward its lower bound $\log \underline{\rho}_{\theta_\tau^*}$. By comparing the energy of the parameterized density against these competitors, we show that any excess mass placed outside the sublevel set $\mathcal{W}_{\theta_\tau^*}^\sigma$ incurs a linear energy penalty. The expressivity gap $\varepsilon_{\text{real}}(\tau)$ bounds this penalty, directly restricting the fraction of excess mass that can leak outside the effective-potential sublevel set.

Proof. We provide the derivation for modality $\mathcal{X}$ with parameter $\theta \in \Theta$. The corresponding result for $\phi \in \Phi$ follows by interchanging modalities and notation.

Fix $\phi \in \Phi$ and $\tau > 0$, and let $\theta_\tau^* \in \arg \min_{\theta \in \Theta} \mathcal{J}_\tau^{\text{Sym}}(\theta, \phi)$. Since $\rho_{\theta_\tau^*} \geq \underline{\rho}_{\theta_\tau^*}$ μ -a.e., $M_{\text{ex}}^{(\theta_\tau^*)} = 1 - \underline{\rho}_{\theta_\tau^*} \mu(\mathcal{Z}) > 0$, and $\int_{\mathcal{Z}} \rho_{\theta_\tau^*}(z) \mu(dz) = 1$, we have $\bar{\rho}_{\theta_\tau^*} \geq 0$ μ -a.e. and

$$\int_{\mathcal{Z}} \bar{\rho}_{\theta_\tau^*}(z) \mu(dz) = \frac{1 - \underline{\rho}_{\theta_\tau^*} \mu(\mathcal{Z})}{M_{\text{ex}}^{(\theta_\tau^*)}} = 1,$$

hence $\bar{\rho}_{\theta_\tau^*} \mu$ is a valid probability measure.

Define the effective potential field and its essential infimum

$$V_{\theta_\tau^*|\phi}(z) = \frac{1}{\tau} U_{\theta_\tau^*|\phi}(z) + \log \rho_\phi(z), \quad v_* := \text{ess inf}_\mu V_{\theta_\tau^*|\phi},$$

and, for $\sigma > 0$, define the sublevel set

$$\mathcal{W}_{\theta_\tau^*}^\sigma = \{z \in \mathcal{Z} : V_{\theta_\tau^*|\phi}(z) \leq v_* + \sigma\}.$$

Step 1: A quantitative implication of near-optimality. Consider the coordinate functional with the peer encoder ϕ fixed:

$$\mathcal{F}_\tau(\rho) := \mathcal{F}_{\tau, \mathbf{U}_{\theta_\tau^*}, \phi}^{\text{Sym}}(\rho, \rho_\phi), \quad \rho \in \mathcal{D}_{\mu, \rho_{\theta_\tau^*}}(\mathcal{Z}).$$

As made explicit in the proof of [Prop. 4.1](#), by expanding $\mathcal{F}_{\tau, \mathbf{U}_{\theta_\tau^*}, \phi}^{\text{Sym}}$ and collecting ρ -dependent terms, $\mathcal{F}_\tau$ can be written in the form

$$\mathcal{F}_\tau(\rho) = \frac{1}{2} \int_{\mathcal{Z}} V_{\theta_\tau^*|\phi}(z) \rho(z) \mu(dz) + \frac{1}{2} \int_{\mathcal{Z}} \rho_\phi(z) \log \rho(z) \mu(dz) + \text{const}_\tau(\phi), \quad (71)$$

where $\text{const}_\tau(\phi)$ is independent of ρ . By the floor constraint $\rho \geq \underline{\rho}_{\theta_\tau^*}$ μ -a.e., the monotonicity of the logarithm, and the fact that ρ_ϕ is a probability density, we have

$$\int_{\mathcal{Z}} \rho_\phi(z) \log \rho(z) \mu(dz) \geq \int_{\mathcal{Z}} \rho_\phi(z) \log \underline{\rho}_{\theta_\tau^*} \mu(dz) = \log \underline{\rho}_{\theta_\tau^*}.$$

Hence, for any $\rho \in \mathcal{D}_{\mu, \rho_{\theta_\tau^*}}(\mathcal{Z})$,

$$\mathcal{F}_\tau(\rho) \geq \frac{1}{2} \int_{\mathcal{Z}} V_{\theta_\tau^*|\phi}(z) \rho(z) \mu(dz) + \frac{1}{2} \log \underline{\rho}_{\theta_\tau^*} + \text{const}_\tau(\phi). \quad (72)$$

Let

$$\mathcal{F}_\tau^* := \inf_{\rho \in \mathcal{D}_{\mu, \rho_{\theta_\tau^*}}(\mathcal{Z})} \mathcal{F}_\tau(\rho).$$

Applying [Prop. 4.1](#) with density floor $\underline{\rho} = \underline{\rho}_{\theta_\tau^*}$ to $\mathcal{F}_\tau$ yields, for each $\delta > 0$, an infimizing competitor $\rho^{(\delta)} \in \mathcal{D}_{\mu, \rho_{\theta_\tau^*}}(\mathcal{Z})$ whose excess mass is supported on a shrinking set $\mathcal{S}^\delta \subseteq \mathcal{W}_{\theta_\tau^*}^\delta$ with $\mu(\mathcal{S}^\delta) \rightarrow 0$ as $\delta \rightarrow 0$, such that $\mathcal{F}_\tau(\rho^{(\delta)}) \rightarrow \mathcal{F}_\tau^*$ as $\delta \rightarrow 0$.

Moreover, since $\rho^{(\delta)} = \underline{\rho}_{\theta_\tau^*}$ on $\mathcal{Z} \setminus \mathcal{S}^\delta$ and ρ_ϕ is continuous on compact $\mathcal{Z}$, we have

$$\begin{aligned} \int_{\mathcal{Z}} \rho_\phi(z) \log \rho^{(\delta)}(z) \mu(dz) &= \log \underline{\rho}_{\theta_\tau^*} + \int_{\mathcal{S}^\delta} \rho_\phi(z) \log \left(1 + \frac{M_{\text{ex}}^{(\theta_\tau^*)}}{\underline{\rho}_{\theta_\tau^*} \mu(\mathcal{S}^\delta)} \right) \mu(dz) \\ &\leq \log \underline{\rho}_{\theta_\tau^*} + \|\rho_\phi\|_\infty \mu(\mathcal{S}^\delta) \log \left(1 + \frac{c}{\mu(\mathcal{S}^\delta)} \right) \\ &= \log \underline{\rho}_{\theta_\tau^*} + o_\delta(1), \end{aligned}$$

where $o_\delta(1) \rightarrow 0$ as $\delta \rightarrow 0$ because $\mu(\mathcal{S}^\delta) \log(1/\mu(\mathcal{S}^\delta)) \rightarrow 0$. In addition, since $\rho^{(\delta)} - \underline{\rho}_{\theta_\tau^*}$ places all excess mass inside $\{z \in \mathcal{Z} : V_{\theta_\tau^*|\phi}(z) \leq v_* + \delta\}$,

$$\begin{aligned} \int_{\mathcal{Z}} V_{\theta_\tau^*|\phi}(z) \rho^{(\delta)}(z) \mu(dz) &= \underline{\rho}_{\theta_\tau^*} \int_{\mathcal{Z}} V_{\theta_\tau^*|\phi}(z) \mu(dz) + M_{\text{ex}}^{(\theta_\tau^*)} \int_{\mathcal{Z}} V_{\theta_\tau^*|\phi}(z) \bar{\rho}^{(\delta)}(z) \mu(dz) \\ &\leq \underline{\rho}_{\theta_\tau^*} \int_{\mathcal{Z}} V_{\theta_\tau^*|\phi}(z) \mu(dz) + M_{\text{ex}}^{(\theta_\tau^*)} (v_* + \delta), \end{aligned}$$

where $\bar{\rho}^{(\delta)} := \rho^{(\delta)} - \underline{\rho}_{\theta_\tau^*} / M_{\text{ex}}^{(\theta_\tau^*)}$. Plugging these bounds into [Eq. \(71\)](#) and letting $\delta \rightarrow 0$ yields

$$\mathcal{F}_\tau^* \leq \frac{1}{2} \left(\underline{\rho}_{\theta_\tau^*} \int_{\mathcal{Z}} V_{\theta_\tau^*|\phi}(z) \mu(dz) + M_{\text{ex}}^{(\theta_\tau^*)} v_* \right) + \frac{1}{2} \log \underline{\rho}_{\theta_\tau^*} + \text{const}_\tau(\phi). \quad (73)$$

Now apply [Eq. \(72\)](#) to $\rho = \rho_{\theta_\tau^*}$ and subtract [Eq. \(73\)](#). Using

$$\int_{\mathcal{Z}} V_{\theta_\tau^*|\phi}(z) \rho_{\theta_\tau^*}(z) \mu(dz) = \underline{\rho}_{\theta_\tau^*} \int_{\mathcal{Z}} V_{\theta_\tau^*|\phi}(z) \mu(dz) + M_{\text{ex}}^{(\theta_\tau^*)} \int_{\mathcal{Z}} V_{\theta_\tau^*|\phi}(z) \bar{\rho}_{\theta_\tau^*}(z) \mu(dz),$$

we obtain

$$\mathcal{F}_\tau(\rho_{\theta_\tau^*}) - \mathcal{F}_\tau^* \geq \frac{M_{\text{ex}}^{(\theta_\tau^*)}}{2} \int_{\mathcal{Z}} (V_{\theta_\tau^*|\phi}(z) - v_*) \bar{\rho}_{\theta_\tau^*}(z) \mu(dz). \quad (74)$$

Step 2: Use expressivity and convert to leakage. By the expressivity assumption in Eq. (69),

$$\mathcal{F}_\tau(\rho_{\theta_\tau^*}) \leq \mathcal{F}_\tau^* + \varepsilon_{\text{real}}(\tau).$$

Combining this with Eq. (74) gives

$$\int_{\mathcal{Z}} (V_{\theta_\tau^*|\phi}(z) - v_*) \bar{\rho}_{\theta_\tau^*}(z) \mu(dz) \leq \frac{2\varepsilon_{\text{real}}(\tau)}{M_{\text{ex}}^{(\theta_\tau^*)}}. \quad (75)$$

Finally, since $V_{\theta_\tau^*|\phi}(z) - v_* \geq \sigma$ μ -a.e. on $\mathcal{Z} \setminus \mathcal{W}_{\theta_\tau^*}^\sigma$, we have

$$\begin{aligned} \int_{\mathcal{Z}} (V_{\theta_\tau^*|\phi}(z) - v_*) \bar{\rho}_{\theta_\tau^*}(z) \mu(dz) &\geq \int_{\mathcal{Z} \setminus \mathcal{W}_{\theta_\tau^*}^\sigma} (V_{\theta_\tau^*|\phi}(z) - v_*) \bar{\rho}_{\theta_\tau^*}(z) \mu(dz) \\ &\geq \sigma \int_{\mathcal{Z} \setminus \mathcal{W}_{\theta_\tau^*}^\sigma} \bar{\rho}_{\theta_\tau^*}(z) \mu(dz) \\ &= \sigma \bar{\rho}_{\theta_\tau^*} \mu(\mathcal{Z} \setminus \mathcal{W}_{\theta_\tau^*}^\sigma), \end{aligned}$$

and therefore, by Eq. (75),

$$\bar{\rho}_{\theta_\tau^*} \mu(\mathcal{Z} \setminus \mathcal{W}_{\theta_\tau^*}^\sigma) \leq \frac{2\varepsilon_{\text{real}}(\tau)}{\sigma M_{\text{ex}}^{(\theta_\tau^*)}},$$

which is exactly Eq. (70). The concluding asymptotic statement follows immediately: by the uniform lower bound $M_{\text{ex}}^{(\theta_\tau^*)} \geq m_{\text{ex}} > 0$, if $\sigma = \sigma(\tau) \downarrow 0$ and $\varepsilon_{\text{real}}(\tau)/\sigma(\tau) \rightarrow 0$, then $\bar{\rho}_{\theta_\tau^*} \mu(\mathcal{Z} \setminus \mathcal{W}_{\theta_\tau^*}^{\sigma(\tau)}) \rightarrow 0$ as $\tau \rightarrow 0^+$. $\square$

F Experimental Details

This section details the experimental procedures used to empirically validate our theoretical framework. Specifically, we provide comprehensive setups for evaluating: (i) the convergence of finite-batch InfoNCE gradients to their deterministic counterparts; (ii) the Gibbs-shaped concentration and spread behavior of unimodal training under sharp kernels; (iii) the emergence of a persistent, population-level modality gap induced by conditional heterogeneity; and (iv) the real-world implications of cross-modal divergence via probing and fine-tuning CLIP models on MS-COCO.

F.1 Large-Batch Gradient Consistency Across Similarity Functions

This experiment provides a numerical illustration of the large-batch consistency established in Thm. 3.1. We verify that as the number of negatives N increases, the finite-batch InfoNCE gradient aligns increasingly well with a high-fidelity large- N reference gradient. We evaluate this behavior under two different similarity functions and embedding geometries. The corresponding large-batch consistency for the full symmetric multimodal objective is established separately in Thm. 4.1.

Data and pairing. We construct a synthetic unimodal dataset in $\mathcal{X} \subset \mathbb{R}^m$ ($m = 64$) from a mixture of $K = 4$ Gaussians:

$$\mathbf{x} \sim \sum_{k=1}^K \pi_k \mathcal{N}(\boldsymbol{\mu}_k, \sigma^2 \mathbf{I}).$$

The component means $\{\boldsymbol{\mu}_k\}$ are random directions scaled by a separation factor of 4, with $\sigma = 1$. Positive pairs are generated via additive noise $\tilde{\mathbf{x}} = \mathbf{x} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_{\text{aug}}^2 \mathbf{I})$. Negative samples $\{\mathbf{x}_j^-\}_{j=1}^N$ are i.i.d. draws from the same data distribution.

Encoder and regimes. We employ a linear encoder $f_{\mathbf{W}}(x)$ with parameter $\mathbf{W} \in \mathbb{R}^{n \times m}$ ($n = 128$) under two distinct geometric regimes:

$$z = f_{\mathbf{W}}(x) = \begin{cases} \text{normalize}(\mathbf{W}x) \in \mathbb{S}^{n-1}, & (\text{Spherical regime}), \\ \tanh(\mathbf{W}x) \in [-1, 1]^n, & (\text{Compact Euclidean regime}). \end{cases}$$

The spherical case corresponds to the standard feature-normalized setting, while the compact Euclidean case provides a compact non-normalized embedding domain suitable for evaluating distance-based similarities. The latter is intended to test the broader partition-field mechanism numerically rather than to instantiate the constant kernel-volume assumption exactly.

We evaluate the exponential kernel defined in § 2 with two canonical similarity functions: the cosine similarity

$$\text{sim}_{\cos}(z, w) = \langle z, w \rangle$$

and the RBF similarity

$$\text{sim}_{\text{rbf}}(z, w) = -\frac{1}{n} \|z - w\|^2,$$

with temperatures $\tau_{\cos} = 0.1$ and $\tau_{\text{rbf}} = 1.0$, respectively. The RBF similarity includes the factor $1/n$ to maintain $O(1)$ logits. We set augmentation noise $\sigma_{\text{aug}} = 0.05$ for the cosine regime and $\sigma_{\text{aug}} = 0.2$ for the RBF regime.

Loss and gradients. For each seed, we sample a fixed batch of $B = 64$ anchors/positives and a large “reference” pool of $N_{\text{ref}} = 4096$ negatives. For a given N , we compute the loss $\hat{\ell}_{B,N}(\mathbf{W})$ using the first N negatives from this pool:

$$\hat{\ell}_{B,N}(\mathbf{W}) = \frac{1}{B} \sum_{i=1}^B \left[-\frac{1}{\tau} \text{sim}(z_i, \tilde{z}_i) + \log \left(\exp \left(\frac{\text{sim}(z_i, \tilde{z}_i)}{\tau} \right) + \sum_{j=1}^N \exp \left(\frac{\text{sim}(z_i, w_j^-)}{\tau} \right) \right) \right].$$

Let

$$\mathbf{g}_N := \nabla_{\mathbf{W}} \hat{\ell}_{B,N}(\mathbf{W})$$

denote the gradient estimate using N negatives, and let

$$\mathbf{g}_{\text{ref}} := \mathbf{g}_{N_{\text{ref}}}$$

denote the large- N reference gradient. Fixing the negative pool ensures that variations in $\mathbf{g}_N$ across N arise from the number of included negatives rather than from resampling the negative pool.

Metrics and results. We sweep $N \in \{4, 8, 16, 32, 64, 128, 256, 512, 1024\}$ and report two metrics: (i) *gradient alignment* $\cos(\mathbf{g}_N, \mathbf{g}_{\text{ref}})$, measuring directional agreement; and (ii) *relative error* $\|\mathbf{g}_N - \mathbf{g}_{\text{ref}}\| / \|\mathbf{g}_{\text{ref}}\|$, measuring magnitude deviation. Results are averaged over 20 random seeds; error bars indicate ± 1 standard deviation across seeds. As in Figs. 3c and 3d, across both the spherical and RBF regimes, the gradient alignment increases with N while the relative error decreases. These results are consistent with the large-batch gradient-consistency theory: as the negative partition becomes better approximated by a large sample, the finite- N gradient approaches its large- N reference. The rapid convergence (alignment > 0.95 at $N = 256$) further suggests that the large-batch gradient approximation can already be accurate at moderate numbers of negatives in these controlled settings.

F.2 Unimodal Gibbs Equilibrium and Low-Temperature Concentration

This experiment provides a numerical illustration of the intrinsic unimodal geometry characterized in § 3. Specifically, we examine two predictions of the variational perspective: (i) for fixed $\tau > 0$, the functional $\mathcal{F}_{\tau,U}$ admits the unique Gibbs equilibrium characterized in Prop. 3.1; and (ii) as $\tau \rightarrow 0^+$, this equilibrium concentrates around the low-potential regions of U , as predicted by the ground-state concentration result in Prop. 3.2.

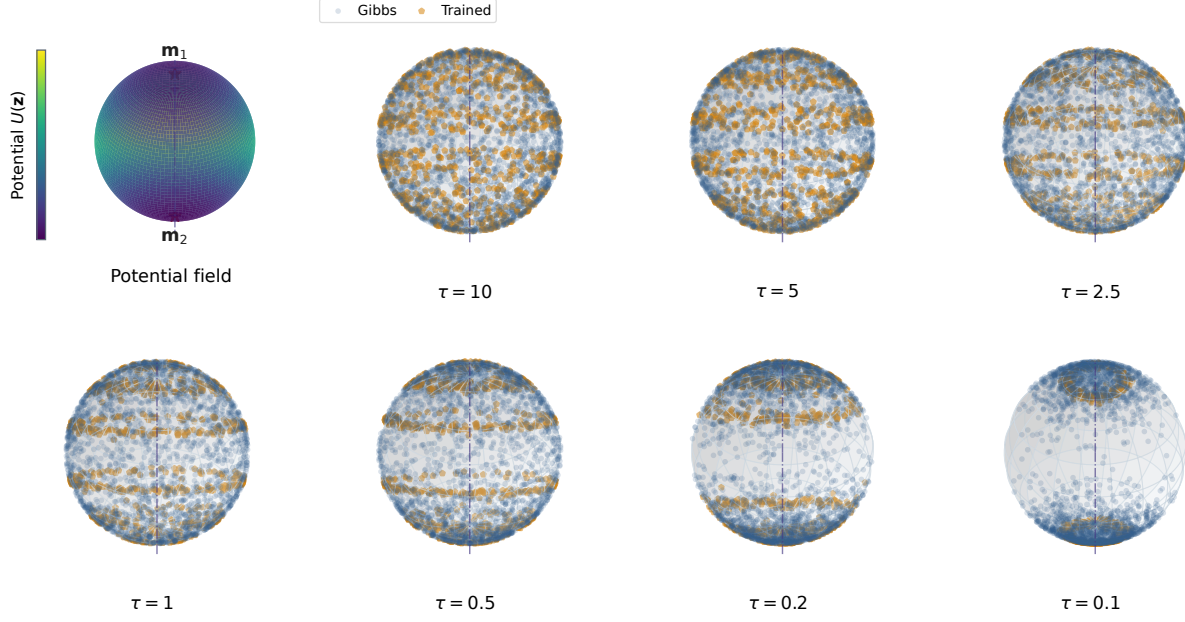


Figure 7: Unimodal potential landscape on $\mathbb{S}^2$ and distributions across temperature τ . Left top: the two-well potential U in Eq. (76) (colored by value), with basin centers $\mathbf{m}_1, \mathbf{m}_2$. Others: Gibbs samples (blue; importance-resampled) and trained particles (orange; minimizing Eq. (78)) across temperatures. As τ decreases, both distributions concentrate around the low-energy basins.

Representation space and potential field. We define the domain as the unit sphere $\mathcal{Z} = \mathbb{S}^2 \subset \mathbb{R}^3$ equipped with surface measure μ and geodesic distance $d_{\mathbb{S}^2}$. We construct a smooth, non-isotropic potential $U(z)$ using a log-sum-exp mixture of two von-Mises–Fisher-like modes:

$$U(z) := -\frac{1}{\gamma} \log(w \exp(\kappa \langle z, \mathbf{m}_1 \rangle) + (1 - w) \exp(\kappa \langle z, \mathbf{m}_2 \rangle)), \quad z \in \mathbb{S}^2, \quad (76)$$

with $\gamma = 12$, concentration $\kappa = 12$, and mixture weight $w = 0.5$. The modes are centered at $\mathbf{m}_1 = (0, 0, 1)$ and the separated direction $\mathbf{m}_2 = (0.85, 0.15, -0.50) / \|(0.85, 0.15, -0.50)\|$, creating two distinct low-energy basins (Fig. 7, left top).

Particle training objective. To emulate minimization of

$$\mathcal{F}_{\tau, U}(\rho) = \frac{1}{\tau} \int_{\mathcal{Z}} U(z) \rho(z) \mu(dz) - H(\rho)$$

(Def. 3.3) without committing to a parametric encoder, we represent ρ by M particles $\{\mathbf{z}_i\}_{i=1}^M \subset \mathbb{S}^2$ and optimize a standard particle/KDE surrogate. We parameterize unconstrained variables $\{\mathbf{v}_i\}_{i=1}^M \subset \mathbb{R}^3$ and project them to the sphere via $\mathbf{z}_i = \mathbf{v}_i / \|\mathbf{v}_i\|$ at each iteration. We approximate the density at particle locations using a geodesic Gaussian kernel on the sphere,

$$\hat{\rho}_h(\mathbf{z}_i) := \frac{1}{M} \sum_{j=1}^M \exp\left(-\frac{d_{\mathbb{S}^2}(\mathbf{z}_i, \mathbf{z}_j)^2}{2h^2}\right), \quad (77)$$

with bandwidth $h = 0.35$. The optimized objective is the particle approximation

$$\hat{\mathcal{F}}_{\tau, U}(\{\mathbf{z}_i\}) := \frac{1}{\tau} \cdot \frac{1}{M} \sum_{i=1}^M U(\mathbf{z}_i) + \frac{1}{M} \sum_{i=1}^M \log \hat{\rho}_h(\mathbf{z}_i), \quad (78)$$

where the second term is a KDE-based surrogate for $-H(\rho)$ up to an additive constant. We minimize Eq. (78) using Adam (Kingma and Ba, 2015) for 5000 steps with learning rate 5×10^{-2} , while injecting Gaussian noise

from $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ with $\sigma = 0.06$ at each iteration. This Langevin-inspired perturbation encourages continued exploration of the landscape and reduces sensitivity to shallow local traps, while the KDE entropy term promotes dispersion among the particles.

We sweep $\tau \in \{10, 5.0, 2.5, 1.0, 0.5, 0.2, 0.1\}$ and run 20 seeds. As a reference, we estimate the Gibbs equilibrium

$$\rho_\tau^*(z) \propto \exp\left(-\frac{U(z)}{\tau}\right)$$

(**Prop. 3.1**) by importance sampling: we draw $n_{\text{mc}} = 120,000$ uniform points on $\mathbb{S}^2$, weight them by $\exp(-U/\tau)$, and compute weighted expectations. For visualization, we use a separate pool of 24,000 uniform points and perform weighted sampling without replacement to draw 2400 Gibbs points per τ .

Evaluation metric. To quantify low- τ concentration, we report the *cap mass* within geodesic radius $\varepsilon = 0.50$ around either potential well:

$$\text{CapMass}_\varepsilon(\rho) := (\rho\mu)(\mathcal{C}_\varepsilon(\mathbf{m}_1) \cup \mathcal{C}_\varepsilon(\mathbf{m}_2)),$$

where

$$\mathcal{C}_\varepsilon(\mathbf{m}) := \{z \in \mathbb{S}^2 : d_{\mathbb{S}^2}(z, \mathbf{m}) \leq \varepsilon\}.$$

For the trained particles, this is simply the fraction of particles whose distance to $\{\mathbf{m}_1, \mathbf{m}_2\}$ is at most ε . For the Gibbs baseline, we compute the same quantity as a weighted Monte Carlo estimate. We report the mean $\pm$ standard error over the 20 seeds.

Results. **Fig. 7** visualizes the potential landscape and the particle distributions across τ together with the corresponding Gibbs reference. At high temperature (e.g., $\tau = 10$), both the Gibbs samples and the trained particles spread broadly over the sphere: the entropy term in $\mathcal{F}_{\tau,U}$ is comparatively strong, producing a diffuse distribution consistent with entropic dispersion. As τ decreases (e.g., $\tau = 0.2, \tau = 0.1$), the potential term $\frac{1}{\tau} \int_{\mathcal{Z}} U(z) \rho(z) \mu(dz)$ becomes increasingly influential, and both distributions concentrate near the two low-energy basins. Notably, the trained particles exhibit semi-regular lattice-like bands, in contrast to the random scatter of the Gibbs samples. This micro-structure is consistent with the repulsion induced by the KDE entropy surrogate, which encourages the finite particle set to spread within the occupied regions. Such finite-particle organization does not alter the macroscopic prediction: at $\tau = 0.1$, the global mass is visibly localized around the low-energy basins, in qualitative agreement with ground-state concentration (**Prop. 3.2**).

This trend is quantified in **Fig. 3a**: the cap mass $\text{CapMass}_\varepsilon(\rho)$ increases as τ decreases for both the Gibbs baseline and the trained particles. The trained curve closely tracks the same trend, with small deviations attributable to the finite particle count M , finite optimization time, and fixed KDE bandwidth h , which imposes a smoothing scale on the entropy surrogate. Overall, the experiment supports the unimodal geometric picture: the Gibbs reference exhibits systematic low-temperature concentration near low-potential regions of U , while optimization of the particle free-energy surrogate qualitatively tracks this behavior across temperatures.

F.3 Multimodal Structural Gap under Controlled Misalignment

This experiment provides a controlled empirical illustration of the *multimodal* mechanism developed in § 4, motivated by the well-documented *modality gap* phenomenon observed in CLIP-style training (Liang et al., 2022). Our theory shows that although symmetric multimodal InfoNCE admits a stable large-batch deterministic limit, its intrinsic landscape is *cross-coupled*, with each coordinate governed by an effective potential that depends on the peer marginal. When the two modalities exhibit heterogeneous conditional laws—arising from noise, distinct semantic density, or mismatched semantic structure—these directional forces need not be simultaneously compatible with coincident marginals and can sustain a persistent separation. We illustrate this mechanism in a controlled setting by progressively introducing structured cross-modal misalignment and measuring the resulting marginal gap.

Latent variables and paired observations. We generate paired observations $(\mathbf{x}_1, \mathbf{x}_2)$ representing two modalities that stem from a shared latent angle $\alpha \in (-\pi, \pi]$. To rule out the trivial solution where marginals are uniform (and thus trivially match regardless of noise), we sample the shared latent α from a *non-uniform*, multi-peaked mixture of von Mises distributions:

$$\alpha \sim w \text{VM}(\mu_1, \kappa) + (1 - w) \text{VM}(\mu_2, \kappa), \quad \text{with } w = 0.7, \quad \mu_1 = 0, \quad \mu_2 = \pi, \quad \kappa = 6.$$

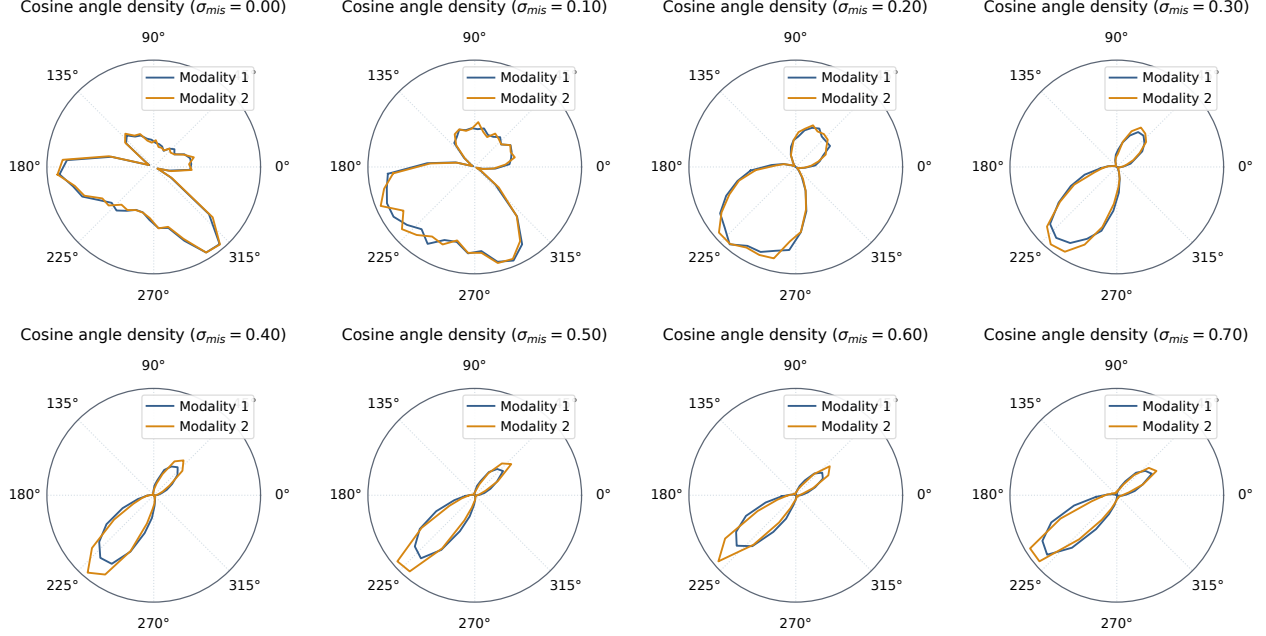


Figure 8: Polar marginals across misalignment. Estimated angle marginals of the two modalities on $\mathbb{S}^1$. The mismatch becomes increasingly apparent as σ_{mis} grows.

Both modalities observe this angle as a projection onto the 2D unit circle, corrupted by independent Gaussian observation noise $\xi_1, \xi_2 \sim \mathcal{N}(\mathbf{0}, \sigma_{\text{obs}}^2 \mathbf{I})$, where $\sigma_{\text{obs}} = 0.02$. Modality 1 is generated by mapping the shared latent directly to the observation space:

$$\mathbf{x}_1 := [\cos \alpha, \sin \alpha]^\top + \xi_1.$$

Modality 2 observes a *misaligned* version of the latent. We introduce an additive angular shift $\eta \sim \mathcal{N}(0, \sigma_{\text{mis}}^2)$ before mapping it to the observation space:

$$\mathbf{x}_2 := [\cos(\alpha + \eta), \sin(\alpha + \eta)]^\top + \xi_2.$$

Thus, increasing the variance of the misalignment σ_{mis} increases the *cross-modal conditional mismatch* between $\mathbf{x}_1$ and $\mathbf{x}_2$, while keeping the marginal complexity of both modalities comparable.

Encoders, learning objective, and protocol. To isolate the directional alignment geometry and prevent magnitude from confounding the representation geometry, we apply bias-free linear encoders $f, g: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ to the observed modalities. The unnormalized embeddings $\tilde{\mathbf{z}}_1 = f(\mathbf{x}_1)$ and $\tilde{\mathbf{z}}_2 = g(\mathbf{x}_2)$ are then projected onto the unit circle:

$$\mathbf{z}_1 = \frac{\tilde{\mathbf{z}}_1}{\|\tilde{\mathbf{z}}_1\|}, \quad \mathbf{z}_2 = \frac{\tilde{\mathbf{z}}_2}{\|\tilde{\mathbf{z}}_2\|} \in \mathbb{S}^1.$$

We adopt the cosine similarity $\text{sim}(z, w) = \langle z, w \rangle$ and the corresponding exponential kernel $\kappa_\tau(z, w) = \exp(\text{sim}(z, w)/\tau)$ with temperature $\tau = 0.07$. This kernel directly instantiates the InfoNCE objective central to our multimodal analysis. Specifically, for a batch of N paired observations, training minimizes the symmetric CLIP loss:

$$\mathcal{L}_{\text{CLIP}}(f, g) = \frac{1}{2} (\mathcal{L}_{\text{CE}}(\mathbf{S}/\tau, \mathbf{y}_{\text{diag}}) + \mathcal{L}_{\text{CE}}(\mathbf{S}^\top/\tau, \mathbf{y}_{\text{diag}})),$$

where $\mathbf{S} \in \mathbb{R}^{N \times N}$ is the pairwise similarity matrix with entries $\mathbf{S}_{ij} := \langle \mathbf{z}_{1,i}, \mathbf{z}_{2,j} \rangle$, $\mathcal{L}_{\text{CE}}$ denotes the standard cross-entropy loss, and the target $\mathbf{y}_{\text{diag}}$ indicates that the matched positive pairs lie on the diagonal ($i = j$).

We use Adam with learning rate 5×10^{-3} for 2000 steps and batch size 256. We sweep eight misalignment scales $\sigma_{\text{mis}} \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$; larger scales induce progressively more severe cross-modal misalignment, yet all settings retain meaningful statistical dependence between positive pairs. For each σ_{mis} ,

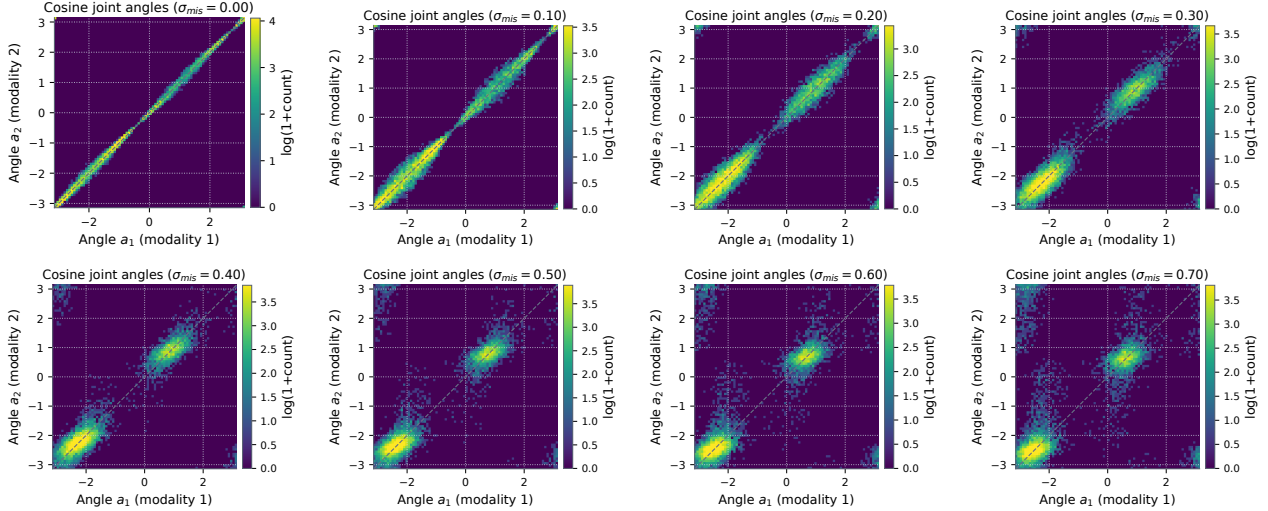


Figure 9: Joint-angle coupling across misalignment. 2D histograms of (a_1, a_2) from $\sigma_{\text{mis}} = 0.0$ to $\sigma_{\text{mis}} = 0.7$. Diagonal concentration at small σ_{mis} indicates near-deterministic alignment, while increasing off-diagonal spread at larger σ_{mis} indicates increasingly noisy cross-modal coupling.

we repeat training over 20 random seeds (seeds 0–19). The large batch size makes the finite-batch objective and gradients better approximate the large-batch regime characterized in [Thm. 4.1](#).

Evaluation metrics. After training, we evaluate the learned representations on $n_{\text{eval}} = 8000$ fresh paired samples. For the unit-normalized embeddings $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{S}^1$, we compute the embedding angles $a_1 = \text{atan2}(\mathbf{z}_{1,y}, \mathbf{z}_{1,x})$ and $a_2 = \text{atan2}(\mathbf{z}_{2,y}, \mathbf{z}_{2,x})$, where the subscripts x and y denote the respective coordinate axes. The following visual diagnostics are considered:

(i) *Marginal gap.* We discretize a_1 and a_2 into 60 bins on $(-\pi, \pi]$ and estimate the two empirical angle densities $\hat{\rho}_\theta$ and $\hat{\rho}_\phi$. As a scalar summary of marginal mismatch, we report the symmetric KL divergence $D_{\text{KL}}^{\text{Sym}}(\hat{\rho}_\theta, \hat{\rho}_\phi)$ averaged over 20 random seeds (seeds 0–19), with error bars or shaded areas showing $\pm$ standard error.

(ii) *Polar marginals.* To visualize *where* the mismatch occurs on the circle, we plot the two estimated angle marginals as polar density curves for each misalignment scale σ_{mis} .

(iii) *Joint-angle heatmaps.* To visualize the learned *cross-modal coupling* beyond marginals, we plot a log-scaled 2D histogram, i.e., a $\log(1 + \text{count})$ heatmap, of the joint distribution of (a_1, a_2) . Diagonal concentration indicates near-deterministic alignment $a_2 \approx a_1$; off-diagonal spread indicates increasing cross-modal mismatch.

(iv) *Angle-shift density.* Finally, we compute the wrapped angular difference $\Delta a := \text{wrap}(a_2 - a_1) \in (-\pi, \pi]$ and plot its empirical density. This directly measures the distribution of residual misalignment in embedding space.

For the scalar marginal-gap curve, we average over 20 random seeds and report the mean with a $\pm$ SEM band. For the qualitative diagnostics (polar marginals, joint-angle heatmaps, and angle-shift densities), we display one fixed representative run (seed 0) and reuse the same held-out latent sample across all misalignment scales, matching the animation setup.

Results. [Fig. 3b](#) shows that the marginal discrepancy $D_{\text{KL}}^{\text{Sym}}(\hat{\rho}_\theta, \hat{\rho}_\phi)$ increases as the latent misalignment scale σ_{mis} grows. Since the encoder architecture, similarity function, and optimization protocol remain fixed across all runs, the controlled perturbation σ_{mis} isolates the role of conditional mismatch in this setting. The result is consistent with the theoretical mechanism: as the two modalities become less conditionally compatible, maintaining both pairwise alignment and coincident marginals becomes increasingly difficult under the symmetric objective, and a persistent marginal separation emerges.

The polar marginals in [Fig. 8](#) provide a structural view of *how* this separation manifests geometrically. First, across all settings ($\sigma_{\text{mis}} \geq 0$), the learned angular marginals are clearly *multi-peaked* rather than uniform. The density concentrates in specific angular regions, reflecting the underlying non-uniform latent law and

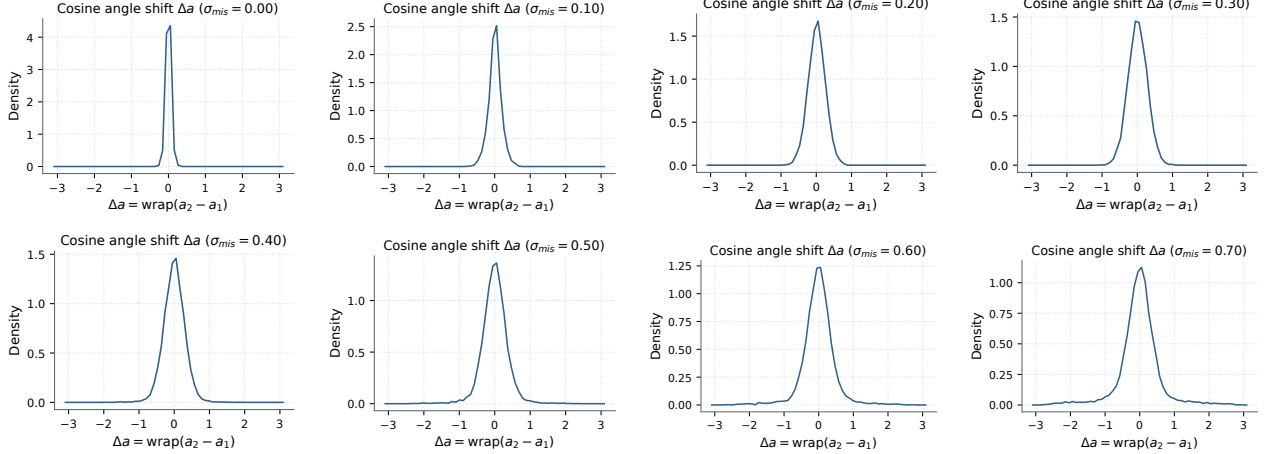


Figure 10: Residual embedding-space misalignment. Densities of $\Delta a = \text{wrap}(a_2 - a_1)$ from $\sigma_{\text{mis}} = 0.0$ to $\sigma_{\text{mis}} = 0.7$. The distribution broadens as σ_{mis} increases, indicating larger and more variable residual mismatch in embedding space.

showing that the learned representations retain its multi-modal structure. Second, the evolution of these peaks reveals the gap mechanism. At $\sigma_{\text{mis}} = 0$, the polar curves for the two modalities nearly overlap, consistent with a compatible pairing law where symmetric InfoNCE can simultaneously achieve alignment and marginal coincidence. As σ_{mis} increases from zero, the curves begin to diverge within the same modal regions: the peaks shift relative to each other, and their probability masses differ. This mismatch becomes visually more pronounced as σ_{mis} increases, tracking the quantitative rise in the gap curve.

The joint-angle heatmaps in Fig. 9 elucidate the mechanism behind this separation. At $\sigma_{\text{mis}} = 0$, probability mass concentrates sharply along the diagonal $a_2 \approx a_1$, indicating a near-deterministic coupling. Because the latent distribution is bimodal, this diagonal concentration is not uniform but forms distinct bright spots (high-density clusters) corresponding to the latent modes. As σ_{mis} increases, these clusters persist—but they deform into thickened, elongated regions. The diagonal band broadens significantly, reflecting increasing uncertainty in the conditional coupling between a_1 and a_2 . The accompanying changes in marginal occupancy observed in Fig. 8 show that, in this controlled setting, broader conditional mismatch is accompanied by increasing marginal separation.

Finally, the angle-shift densities in Fig. 10 quantify the residual embedding misalignment $\Delta a = \text{wrap}(a_2 - a_1)$. As σ_{mis} increases, the distribution of Δa transitions from a sharp peak near zero to a progressively wider distribution. The residual shift becomes both larger in magnitude and more variable, matching the qualitative signature expected from the injection of additive latent noise $\eta \sim \mathcal{N}(0, \sigma_{\text{mis}}^2)$.

In summary, these empirical patterns provide a controlled illustration of the multimodal mechanism characterized in § 4. Our theoretical analysis shows that the symmetric objective induces two directional alignment fields coupled through their marginals, so heterogeneous conditional structure can make marginal coincidence unstable and sustain a separation-inducing contribution. In this experiment, the perturbation $\alpha + \eta$ progressively increases the mismatch between the two directional conditional laws. Empirically, this produces two related effects: (i) the broadening diagonal band in Fig. 9 reflects increasing uncertainty in the cross-modal coupling; and (ii) the growing marginal discrepancy in Fig. 3b and Fig. 8 shows that this conditional mismatch is accompanied by increasing separation between the learned modality marginals. Together, these observations are consistent with the intrinsic coordinate geometry developed in § 4: once strict cross-modal compatibility is weakened, the symmetric objective can sustain a persistent modality gap rather than enforcing identical marginal representations.

F.4 MS-COCO Experiments: Pretrained Gap and Controlled Pair Corruption

We complement the synthetic experiments with two studies on MS-COCO. The first is an observational study of pretrained CLIP-like models, designed to assess whether these models exhibit a modality gap and whether improved retrieval performance necessarily corresponds to a smaller cross-modal distributional discrepancy. The second is a controlled intervention study designed to test how weakening image–text compatibility during

Table 2: Retrieval performance and cross-modal discrepancy for pretrained OpenCLIP checkpoints evaluated on the first 5,000 images of MS-COCO **val2017**. Retrieval is reported as image-to-text and text-to-image Recall@1. Cross-modal discrepancy is quantified by the energy distance, RBF-MMD² at the median bandwidth, the centroid gap $\|\mu_I - \mu_T\|$, and centroid cosine similarity. Lower energy distance, MMD², and centroid gap indicate smaller cross-modal discrepancy.

Model	I→T R@1	T→I R@1	$\mathcal{E}(p_I, p_T)$	$\text{MMD}_{\sigma=\text{med}}^2$	$\ \mu_I - \mu_T\ $	$\cos(\mu_I, \mu_T)$
PE-Core-L-14	0.741	0.559	0.599	0.255	0.854	0.097
PE-Core-B-16	0.703	0.499	0.545	0.237	0.808	0.189
ViT-bigG-14	0.670	0.506	0.288	0.129	0.598	0.384
ViT-H-14	0.662	0.485	0.375	0.159	0.697	0.048
ViT-L-14	0.569	0.357	0.697	0.303	0.898	0.184
ResNet-50x16	0.559	0.354	0.502	0.230	0.761	0.349
RN101	0.498	0.299	0.443	0.228	0.670	0.604
ViT-B-16	0.492	0.302	0.688	0.313	0.871	0.309
ResNet-50	0.480	0.289	0.627	0.276	0.849	0.237

training affects the modality gap. Together, these studies relate the intrinsic geometric analysis to realistic image-text representations.

F.4.1 Probing Pretrained Models on MS-COCO

Setup and evaluations. We evaluate a collection of pretrained OpenCLIP checkpoints [Ilharco et al. \(2021\)](#) on MS-COCO **val2017**. For each checkpoint, we use the first 5000 validation images and up to 5 captions per image, yielding one normalized image embedding ($\mathbf{z}_I$) per image and one normalized text embedding ($\mathbf{z}_T$) per caption. Retrieval is evaluated in both directions under the many-captions-to-one-image matching protocol. Image-to-text Recall@ K assesses whether any of the top- K retrieved captions is associated with the query image, whereas text-to-image Recall@ K assesses whether the correct image appears among the top- K retrieved images for a given caption. We report retrieval performance using Recall@1 (R@1), and define AvgR@1 as the mean of the R@1 scores in the two retrieval directions.

To quantify cross-modal discrepancy, we compute the centroid gap $\|\mu_I - \mu_T\|$ with $\mu_I = \mathbb{E}[\mathbf{z}_I]$ and $\mu_T = \mathbb{E}[\mathbf{z}_T]$, and the energy distance

$$\mathcal{E}(p_I, p_T) = 2\mathbb{E}[d(\mathbf{z}_I, \mathbf{z}_T)] - \mathbb{E}[d(\mathbf{z}_I, \mathbf{z}'_I)] - \mathbb{E}[d(\mathbf{z}_T, \mathbf{z}'_T)],$$

where p_I and p_T denote the image and text embedding marginals, $\mathbf{z}_I, \mathbf{z}'_I \sim p_I$ and $\mathbf{z}_T, \mathbf{z}'_T \sim p_T$ are independent draws, and $d(\cdot, \cdot)$ is the Euclidean distance between normalized embeddings. Whereas the centroid gap captures only the separation between the modality means, the energy distance compares cross-modal distances with within-modal distances and thus provides a broader measure of distributional mismatch. Unlike retrieval, which evaluates alignment between matched image-caption pairs, the energy distance operates at the distribution level: it is small when the image and text embedding marginals are distributionally similar and large when cross-modal distances remain systematically greater than within-modal distances. We additionally report an RBF-MMD diagnostic using a bandwidth selected by the median-distance heuristic.

Results. As shown in [Tab. 2](#), the results indicate that retrieval and cross-modal discrepancy are only partially aligned. As already seen in the AvgR@1 plot in [Fig. 6](#) (left), better retrieval often corresponds to a smaller gap, but the trend is clearly non-monotone. For instance, PE-Core-L-14 achieves the best retrieval yet still has a large energy distance and centroid gap, whereas ViT-bigG-14 attains slightly lower retrieval but the smallest gap values in the table. The separate I→T and T→I versus gap plots ([Fig. 11](#)) show that this is not caused by averaging: models with nearly identical directional retrieval can still differ substantially in energy and centroid discrepancy, e.g., ViT-bigG-14 versus PE-Core-B-16 in T→I retrieval, or ViT-L-14 versus ResNet-50x16 in I→T retrieval. The RBF-MMD values show the same broad pattern, suggesting that the observation is not specific to a single discrepancy metric. Overall, these results show that a modality gap persists across the evaluated pretrained CLIP variants and is not reducible to retrieval performance alone; instead, it reflects an additional geometric property of the learned image and text marginals.

F.4.2 Controlled Same-Category Pair Corruption on MS-COCO

Setup and evaluations. To test more directly whether weakening image-text compatibility enlarges the modality gap, we perform a controlled corruption experiment on MS-COCO **train2017**. We first build an

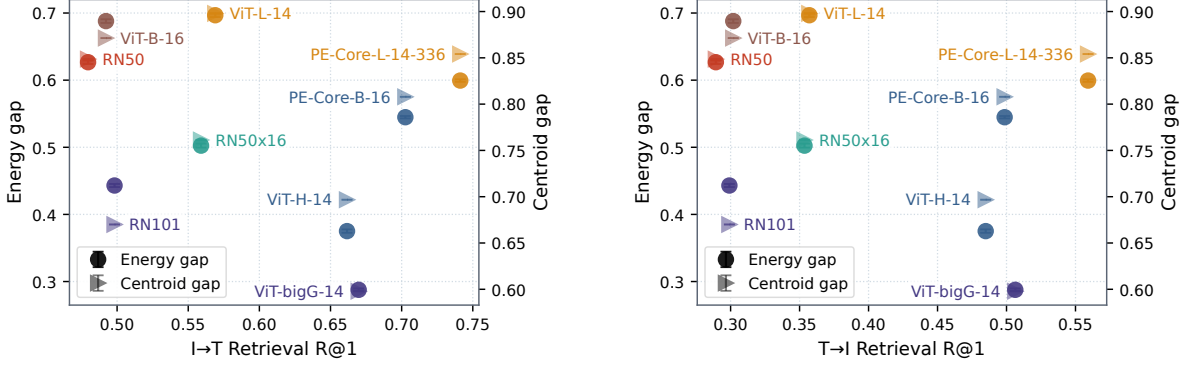


Figure 11: Directional retrieval versus cross-modal gap on MS-COCO for pretrained OpenCLIP checkpoints. Left: image-to-text retrieval ($I \rightarrow T$ R@1) versus gap. Right: text-to-image retrieval ($T \rightarrow I$ R@1) versus gap. In both panels, the left y-axis shows energy distance, and the right y-axis shows centroid gap. The relation is clearly non-monotone: models with similar directional retrieval can exhibit substantially different cross-modal discrepancy. This illustrates that retrieval performance and modality gap are related but not equivalent, and that strong pretrained models may still maintain a substantial structural separation between image and text marginals.

index that records, for each image, its caption set, its COCO category labels, and the pool of other images sharing each category. During training, the image is always kept fixed. With probability $1 - p$, we pair it with one of its own captions; with probability p , we instead replace the caption by a randomly chosen caption from a *different* image that shares at least one category with the current image. Thus, the corruption preserves coarse semantic relatedness while breaking exact instance-level correspondence. We sweep the corruption probability over $p \in \{0, 0.25, 0.50, 0.75, 1.00\}$.

We fine-tune pretrained OpenCLIP backbones (ResNet-50 and ViT-B-16) using the standard symmetric CLIP loss. To isolate the effect of pair corruption rather than full-model adaptation, we freeze the backbone towers and update only the projection heads together with the logit scale. Training uses AdamW with learning rate 10^{-5} , weight decay 0.05, batch size 256, and 5000 optimization steps. Evaluation follows the same protocol as in the previous experiment: we encode the first 5000 images of MS-COCO *val2017* together with up to 5 captions per image, report image-to-text and text-to-image Recall@1, and define AvgR@1 as their mean. As the geometric diagnostic, we report the centroid gap $\|\mu_I - \mu_T\|$.

Results. As shown in Fig. 6 (right), the effect of corruption is clean for both backbones. As the caption corruption probability p increases, retrieval performance degrades while the centroid gap grows. For ResNet-50, AvgR@1 drops from 0.398 at $p = 0$ to 0.226 at $p = 1$, while the centroid gap increases from 0.845 to 1.081. For ViT-B-16, AvgR@1 drops from 0.484 to 0.297, while the centroid gap rises from 0.851 to 1.148. The same pattern holds in both retrieval directions, as shown in Fig. 12: for ResNet-50, $I \rightarrow T$ R@1 decreases from 0.489 to 0.272 and $T \rightarrow I$ R@1 from 0.307 to 0.181; for ViT-B-16, $I \rightarrow T$ R@1 decreases from 0.564 to 0.366 and $T \rightarrow I$ R@1 from 0.405 to 0.229.

Several aspects of this trend are important. First, the corruption is not arbitrary: the replacement caption is sampled from a same-category image, so the supervisory signal remains semantically plausible at a coarse level. Nevertheless, this weaker form of mispairing is already associated with a substantial increase in geometric separation between the image and text marginals. Second, the centroid gap reacts very early: even moving from $p = 0$ to $p = 0.25$ causes a sharp increase in gap for both ResNet-50 and ViT-B-16, while the retrieval drop is initially much milder. This suggests that cross-modal distributional separation is highly sensitive to degradation in pairwise compatibility, and may emerge before retrieval collapses. Third, the monotone behavior is consistent across both architectures, indicating that the phenomenon is not specific to either of the two backbone families evaluated here.

Overall, this experiment provides a controlled real-data analogue of the mechanism developed in the main text. Exact image-text compatibility is progressively weakened, and the learned representations respond in two coupled ways: matched-pair retrieval worsens, and the image/text marginals drift farther apart. This behavior is consistent with the structural-gap view of symmetric contrastive learning: weakening pairwise compatibility

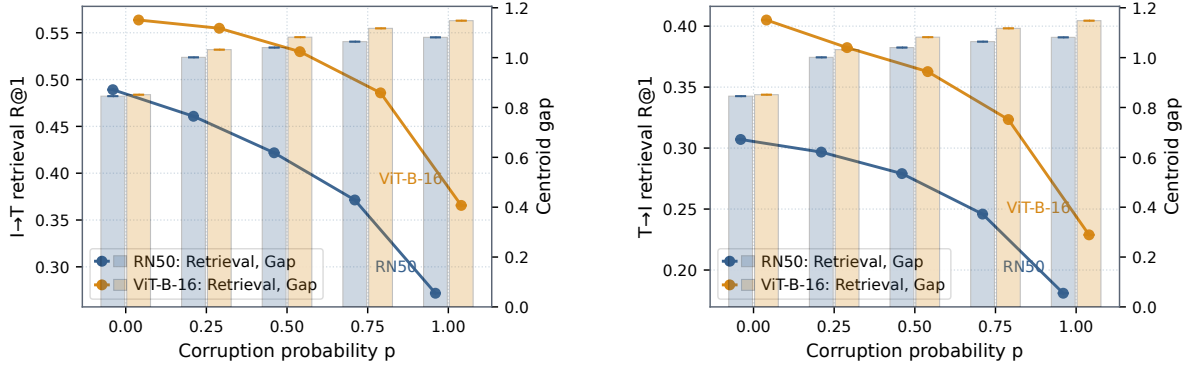


Figure 12: Directional retrieval under same-category caption corruption on MS-COCO. Left: image-to-text retrieval (I→T R@1) versus corruption probability p . Right: text-to-image retrieval (T→I R@1) versus p . Bars show the centroid gap on the right axis. For both RN50 and ViT-B-16, increasing same-category mispairing degrades retrieval in both directions while systematically enlarging the modality gap.

can make cross-modal coincidence harder to sustain, and in this controlled setting the modality gap widens systematically.

G Discussion

G.1 Limitations and Future Directions

We analyze symmetric InfoNCE in a tractable large-batch, low-temperature geometric regime. These assumptions isolate the core mechanism, but they also mark the present boundary of the theory and point to several natural extensions.

Finite-temperature and finite-batch effects. Our analysis combines a large-batch regime ($N \rightarrow \infty$ at fixed τ) with low-temperature asymptotics ($\tau \rightarrow 0^+$) for the intrinsic geometry. Practical training instead uses finite batches and moderate temperatures, introducing stochasticity and entropic smoothing that may soften the barrier-like geometry identified in the limiting analysis. Characterizing finite-temperature and finite-batch corrections, especially how stochastic gradient noise interacts with such barriers, is a natural next step toward sharper practical predictions.

Singularities and boundary behavior. The multimodal analysis imposes a strict density floor to regularize the logarithmic barrier near the boundary of the density space. Removing this technical constraint would require analyzing the singular geometry of the KL divergence near this boundary, and is likely essential for understanding singular measures, including distributions that collapse onto lower-dimensional manifolds or discrete sets.

Geometric homogeneity. We assume a fixed compact embedding manifold with constant kernel volume. Extending the framework to heterogeneous geometries, where the local kernel volume varies across the space, would broaden the present measure-theoretic analysis toward heterogeneous embedding geometries, including settings relevant to hyperbolic contrastive learning and other objectives with spatially varying normalization.

Objective geometry versus optimization dynamics. Our theory describes the geometry of the objective rather than the trajectory of a specific optimization algorithm. Understanding how stochastic optimization explores the resulting barrier-shaped landscape, and under what compatibility conditions it can reach configurations with exact or near-exact marginal matching, remains an important open problem at the interface of geometric analysis and learning theory.

G.2 Broader Implications

Our analysis suggests that contrastive learning should be understood not only through pointwise alignment, but through the population geometry induced on the representation space. In the large-batch regime, InfoNCE gives rise to deterministic fields and intrinsic energies over $\mathcal{Z}$, clarifying why strong retrieval can coexist with persistent modality-level discrepancies: matching positive pairs does not by itself control the induced marginal

geometry. In particular, the multimodal theory shows that improving pairwise alignment alone need not eliminate the modality gap, because pairwise alignment does not by itself determine the induced marginal geometry. This points to a practical lesson for both evaluation and objective design: beyond retrieval, one should monitor distributional diagnostics of the learned marginals, and interventions aimed at closing the gap may need to act on the induced population geometry, for example through explicit discrepancy regularization, shared reference fields, or other mechanisms that weaken the barrier-like coupling between modalities.

More broadly, this perspective suggests that intrinsic population-level analyses can be useful for isolating principled properties of representation-learning objectives that are not obvious from parameter-space dynamics alone. In this view, objective geometry, optimization dynamics, parametrization-specific inductive biases, and identifiability guarantees address distinct but complementary questions: the intrinsic geometry clarifies which population-level configurations the loss favors in the analyzed regime, optimization theory explains how algorithms explore that landscape, and identifiability or generative analyses address when the resulting representations correspond to meaningful latent structure. Our results, therefore, point to a broader methodological lesson: understanding modern representation learning may require combining intrinsic objective-level analysis with complementary theories of optimization, parametrization, and recoverability.

H Changelog

This is a revised presentation of the ICML 2026 version of this paper. The scientific claims, main theoretical results, and empirical conclusions are unchanged. In particular, this revision does not introduce a new theorem-level claim or alter the substantive conclusions of any theorem, proposition, corollary, or experiment. The changes are primarily intended to improve notation, rigor of exposition, scope clarity, and readability.

Changes relative to the ICML 2026 version.

- **Notation and formal setup.** We standardized notation throughout the paper, including distinguishing probability measures from densities, writing $\mathcal{D}_\mu(\mathcal{Z})$ for μ -densities, using `sim` consistently for the similarity function, making the finite- N dependence of InfoNCE objectives explicit, and harmonizing measure, Jacobian, and random-variable notation. Several standing regularity conditions and density-domain conventions are also stated more explicitly.
- **Proof presentation and technical details.** We reorganized the proofs to make intermediate arguments more explicit, including the passage from finite-batch quantities to population limits, differentiation under expectations, kernel-smoothing arguments, Gibbs optimality, the multimodal extremal construction, and the parametric-inheritance results. These revisions clarify the existing arguments and their assumptions; they do not change the claims proved in the ICML version.
- **Scope and interpretation.** We refined the surrounding exposition to distinguish more clearly between pointwise large-batch objective/gradient consistency and optimization trajectories, and between multimodal coordinate geometry and existence of an attained “best response”. Correspondingly, some informal language has been replaced by more precise terminology such as *coordinate infimum*, *infimizing sequence*, and *barrier-like coupling*. The underlying mathematical conclusions are unchanged.
- **Organization and exposition.** The paper has been reformatted for the preprint version, figures and captions have been polished, and the supplementary material has been reorganized for easier navigation. In particular, lemmas, proofs of the unimodal and multimodal results, extended parametric-inheritance statements, experimental details, and discussion are now separated into dedicated appendices.
- **Experimental and discussion text.** The experimental setups and reported results are retained, while their descriptions have been revised to align more precisely with the scope of the theory. We also expanded the discussion of limitations and broader implications, especially the distinction between intrinsic objective geometry, optimization dynamics, parametrization, and identifiability.

For the archival version of record and citation, please refer to the ICML 2026 paper.