UNCOVERING THE ALIGNMENT AND REPELLING TRADEOFF IN CLIP: A PATH TOWARDS GAUSSIAN GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

The contrastive learning framework introduced by CLIP learns image and text representations in a shared embedding space, aligning corresponding samples while repelling representations across the dataset. Building on the alignment–uniformity perspective of Wang and Isola, we identify the term responsible for uniformity in the CLIP objective as a repelling term. We show that its optimum is incompatible with perfect cross-modal alignment: improving the repelling term beyond the uniform distribution on the sphere necessarily requires a positive paired gap. To address this, we propose a prototype objective that shifts repulsion from individual embeddings to representations of image-text pairs. We prove that, under a mild condition on the inverse temperature, its unique population minimizer is the uniform distribution on the sphere with perfect cross-modal alignment. More generally, when perfect alignment is not possible, the optimal distribution is uniform on a sphere whose radius is determined by the HGR maximal correlation. We further show that fixed-dimensional projections of the optimal representations converge to an isotropic Gaussian as the embedding dimension increases. Experiments on MS-COCO and CC3M show that fine-tuning CLIP with the prototype objective substantially reduces the modality gap and improves spherical uniformity and coordinate Gaussianity, while maintaining comparable probe performance.

1 INTRODUCTION

Contrastive learning is a central tool for multimodal representation learning, and CLIP (Radford et al., 2021) remains its most widely used instantiation. Yet the objective is typically analyzed through its empirical performance rather than through the geometry it induces in the joint latent space. This geometry is particularly relevant because recent theory suggests that isotropic-Gaussian representations are especially favorable for linear probes, k-NN, and kernel regression (Balestriero & LeCun, 2025). A basic question therefore remains open: what geometry does the multimodal contrastive objective itself prefer?

Prior work has characterized alignment and uniformity in unimodal contrastive learning (Wang & Isola, 2020), as well as several geometric properties of CLIP representations (Chen et al., 2024; Levi & Gilboa, 2025). Our focus is a more specific structural issue: the population-level tension between aligning matched image–text pairs

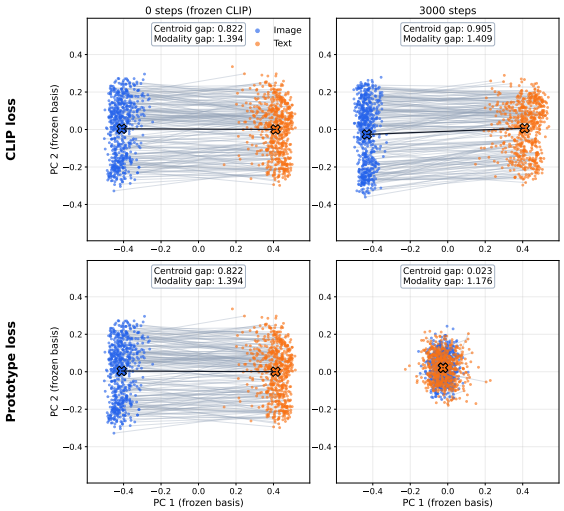


Figure 1: Fine-tuning CLIP ViT-B/32 on MS-COCO (see Section 4.3). All plots show the projection onto the first two principal components of the frozen CLIP features. The prototype loss reduces the centroid gap almost to zero.

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

and repelling the overall embedding distributions, whose respective optima are incompatible. Moreover, we derive a necessary condition for strictly improving upon the uniform distribution: any configuration whose repelling term falls below its value at the uniform distribution on the sphere must exhibit a positive paired gap.

This diagnosis motivates an objective without this conflict, which leads to our second main analytical contribution. We analyze a *prototype* objective, adapting the construction of Dufumier et al. (2023) to paired multimodal data, that shifts repulsion from individual embeddings to pair *midpoints*. Under a mild condition on the inverse temperature, its unique population minimizer is the uniform measure on the sphere with perfect alignment. When perfect alignment is not possible, we further establish that the optimum is a uniform distribution on a sphere of radius strictly less than 1, relating this radius to Hirschfeld–Gebelein–Rényi (HGR) maximal correlation.

At the optimal distribution of the prototype loss, scaled fixed-dimensional projections converge in distribution to a Gaussian by the Poincaré–Maxwell lemma, which was also recently used by Betser et al. (2026) in the unimodal InfoNCE setting. This characterization connects our objective to the motivation of Balestriero & LeCun (2025): a tractable embedding geometry may also provide advantages for downstream prediction.

Experimentally, we fine-tune CLIP ViT-B/32 on MS-COCO and a subset of CC3M, starting from the same pretrained initialization. We find that optimization with the prototype loss produces a stronger reduction in the modality gap defined by Liang et al. (2022) and achieves favorable uniformity and alignment metrics.

Contributions.

- We decompose the population CLIP loss and prove that the minimizers of its alignment and repelling terms are incompatible, without characterizing the global minimizers of their sum. We show that the uniform spherical solution is an ε_d-minimizer, where ε_d is a closed-form constant, and derive necessary conditions on the paired gap and repelling term for strictly improving upon this reference (Section 3.3).
- We prove that the multimodal prototype objective has, as its optimal distribution, the uniform distribution on a sphere of radius R, where R is related to HGR maximal correlation and the modality gap. When perfect alignment is possible, the optimal prototype distribution is the uniform distribution on the unit sphere, and the embeddings are perfectly aligned. We also show that scaled fixed-dimensional projections converge in distribution to a Gaussian (Section 3.4).
- We verify that retraining CLIP with the prototype objective substantially reduces the centroid gap of Liang et al. (2022), improves spherical-uniformity and one-dimensional coordinate-Gaussianity metrics, while achieving probe performance comparable to that of CLIP (Section 4).

Complete proofs are provided in the appendices.

2 Related Work

Contrastive learning and representation geometry. Contrastive objectives such as InfoNCE (Oord et al., 2018), SimCLR (Chen et al., 2020), and supervised contrastive learning (Khosla et al., 2020) are often studied through the geometry they induce. Wang & Isola (2020) formalized the alignment–uniformity perspective in the unimodal setting, while Betser et al. (2026) showed that high-dimensional InfoNCE optima have Gaussian coordinate marginals. We adopt this geometric perspective for the multimodal objective, whose repulsive term couples two modality marginals rather than acting within a single distribution.

Theory and geometry of CLIP, and the modality gap. Chen et al. (2024) derived a population formulation of the CLIP objective, and Levi & Gilboa (2025) documented nontrivial geometric structure in trained CLIP embeddings. Several explanations have been proposed for the modality gap (Liang et al., 2022): Yi et al. (2025) attribute it to dimensional collapse, while Mager et al. (2026) link it to a low-temperature failure mode of InfoNCE. Closest to our setting, Cai et al. (2026) develop a measure-theoretic account in which a persistent divergence term in the multimodal energy

landscape allows strong pairwise alignment to coexist with a gap. We reach a compatible conclusion through a different argument and provide three additional results: a closed-form upper bound ε_d on the optimality gap of the uniform distribution; a maximal-correlation bound for centered encoders that does not depend on the training objective; and an objective whose unconstrained population optimum is both uniform and perfectly aligned.

Prototype-based and energy-based objectives. We build on the prototype loss of Dufumier et al. (2023), adapting it from the unimodal setting to paired multimodal data. Koromilas et al. (2026) likewise recast supervised classification as prototype contrast on the hypersphere. Our analysis shows that, at or below a sharp inverse-temperature threshold $\alpha_c(d)$, the unique population-optimal prototype distribution is uniform on the sphere and admits perfect alignment. We also characterize the constrained optimum at every feasible level of alignment. Our proofs draw on the potential-theoretic framework for energy minimization on spheres and balls summarized by Borodachov et al. (2019). This framework connects the population objectives to their geometric optima and has also informed analyses of contrastive learning, including that of Wang & Isola (2020).

3 Population Analysis of Multimodal Contrastive Geometry

3.1 Preliminaries

We write $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$ and $\overline{\mathbb{B}}^d = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$, $\mathcal{P}(\mathcal{X})$ for the set of probability measures supported in $\mathcal{X}$, σ_d for the uniform measure on $\mathbb{S}^{d-1}$, σ_R for the uniform measure on the sphere of radius R, and id_d for the identity map on $\mathbb{R}^d$. A *coupling* of $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ is a measure π on $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν; we write $\Gamma(\mu, \nu)$ for the set of such couplings. The coupling is the population counterpart of the positive pairs in the contrastive objective. Throughout, a *triplet* refers to two marginals together with a coupling between them.

3.2 The CLIP objective and its two terms

CLIP (Radford et al., 2021) embeds images and text through two encoders whose outputs are ℓ^2-normalized onto $\mathbb{S}^{d-1}$ and compared via inner products; training minimizes the InfoNCE loss $\mathcal{L}_{\mathrm{CLIP},N}$ over batches of N pairs (Appendix A.2). Let $\mu_I, \mu_T \in \mathcal{P}(\mathbb{S}^{d-1})$ denote the laws of the normalized image and text embeddings, and let $\pi \in \Gamma(\mu_I, \mu_T)$ be the positive-pair coupling. Following Chen et al. (2024), at temperature $\tau > 0$ and up to the additive, batch-size-dependent constant $\log N$, the loss decomposes as $\mathcal{L}_{\mathrm{CLIP}} = \mathcal{L}^{\mathrm{align}}_{\mathrm{CLIP}} + \mathcal{L}^{\mathrm{rep}}_{\mathrm{CLIP}}$, where

$$\mathcal{L}^{\mathrm{align}}_{\mathrm{CLIP}}(\mu_I, \mu_T, \pi) = -\frac{1}{\tau}\,\mathbb{E}_{(\hat{z}_I, \hat{z}_T)\sim\pi}\big[\langle \hat{z}_I, \hat{z}_T\rangle\big],$$

$$\mathcal{L}^{\mathrm{rep}}_{\mathrm{CLIP}}(\mu_I, \mu_T) = \frac{1}{2}\,\mathbb{E}_{\hat{z}_I\sim\mu_I}\Big[\log \mathbb{E}_{\hat{z}'_T\sim\mu_T} e^{\langle \hat{z}_I, \hat{z}'_T\rangle/\tau}\Big] + \frac{1}{2}\,\mathbb{E}_{\hat{z}_T\sim\mu_T}\Big[\log \mathbb{E}_{\hat{z}'_I\sim\mu_I} e^{\langle \hat{z}'_I, \hat{z}_T\rangle/\tau}\Big].$$

The population problem is to minimize $\mathcal{L}_{\mathrm{CLIP}}(\mu_I, \mu_T, \pi)$ over $\mu_I, \mu_T \in \mathcal{P}(\mathbb{S}^{d-1})$ and $\pi \in \Gamma(\mu_I, \mu_T)$. We call the second term *repelling* rather than *uniformity* because, unlike its unimodal counterpart (Wang & Isola, 2020), it pushes the two modality distributions away from each other.

3.3 Alignment–repelling tradeoff in CLIP

The repelling term depends only on the marginals, not on the coupling. We first characterize the minimizers of each term separately.

Theorem 1 (Minimizers of the CLIP alignment loss)**.** *The minimizers of $\mathcal{L}^{\mathrm{align}}_{\mathrm{CLIP}}$ are exactly the triplets $\big\{(\mu, \mu, (\mathrm{id}_d, \mathrm{id}_d)_{\#}\mu) : \mu \in \mathcal{P}(\mathbb{S}^{d-1})\big\}$. For every such triplet, $\langle \hat{z}_I, \hat{z}_T\rangle = 1$ a.s., and the optimal value is $-1/\tau$.*

Theorem 2 (Minimizers of the CLIP repelling loss)**.** *The minimizers of $\mathcal{L}^{\mathrm{rep}}_{\mathrm{CLIP}}$ are exactly the antipodal point masses $\mu_I = \delta_z$, $\mu_T = \delta_{-z}$ with $z \in \mathbb{S}^{d-1}$. In this case, $\langle \hat{z}_I, \hat{z}_T\rangle = -1$ a.s., and the optimal value is $-1/\tau$.*

The alignment term is minimized by *any* pair of identical marginals with the diagonal coupling, whereas the repelling term collapses the two modalities onto antipodal points. The two separate optima

are therefore opposed, align $\to \langle \hat{z}_I, \hat{z}_T \rangle = 1$ versus repel $\to \langle \hat{z}_I, \hat{z}_T \rangle = -1$. This distinguishes the multimodal setting from the unimodal one of Wang & Isola (2020), where the uniform distribution minimizes both terms simultaneously and no such tension arises. Consequently, a tradeoff between alignment and uniformity is intrinsic to the minimization of the CLIP objective.

Nevertheless, the uniform distribution on the sphere with the diagonal coupling is nearly optimal; more precisely, it is an ε_d-minimizer, as the following theorem shows.

Theorem 3. *The triplet $\mu_I = \mu_T = \sigma_d$, $\pi = (\mathrm{id}_d, \mathrm{id}_d)_{\#}\sigma_d$ is an ε_d-minimizer of the population CLIP loss, where*

$$\varepsilon_d := \log\Big((2\tau)^{(d-2)/2}\, I_{\frac{d-2}{2}}\big(\tfrac{1}{\tau}\big)\, \Gamma\big(\tfrac{d}{2}\big) \Big),$$

Γ is the Gamma function, and I_ν is the modified Bessel function of the first kind. Moreover, $\varepsilon_d \to 0^+$ as $d \to \infty$. In other words, ε_d is a closed-form upper bound on the suboptimality of the uniform distribution.

The modality gap. The literature usually measures the gap as the distance between the modality centroids (Liang et al., 2022). We instead use the *paired* gap $\Delta_{\text{gap}} := \mathbb{E}_\pi[\|\hat{z}_I - \hat{z}_T\|^2]$, which appears directly in the population objective and it dominates the centroid gap, as prove in the following Proposition, and is therefore a stricter measure of paired cross-modal discrepancy.

Proposition 1 (The paired gap dominates the centroid gap)**.** *Let $\Delta_{\textit{centroid}} := \|\hat{m}_I - \hat{m}_T\|^2$, where $\hat{m}_I = \mathbb{E}_{\mu_I}[\hat{z}_I]$ and $\hat{m}_T = \mathbb{E}_{\mu_T}[\hat{z}_T]$. Then $\Delta_{\textit{gap}} = \Delta_{\textit{centroid}} + \mathrm{tr}\big(\mathrm{Cov}_\pi(\hat{z}_I - \hat{z}_T)\big) \;\geq\; \Delta_{\textit{centroid}}$, so minimizing the paired gap is a stronger requirement than minimizing the centroid gap.*

Relation to maximal correlation. To bound cross-modal dependence independently of any loss, we use the Hirschfeld–Gebelein–Rényi (HGR) maximal correlation (Rényi, 1959) between the raw modalities X_I and X_T, defined over $f \in L^2(\mu_X^I)$ and $g \in L^2(\mu_X^T)$ as

$$\rho_{\mathrm{HGR}}(X_I, X_T) := \sup\big\{ \mathbb{E}[f(X_I)\, g(X_T)] : f,\, g \in L^2,\; \mathbb{E}[f] = \mathbb{E}[g] = 0,\; \mathbb{E}[f^2] = \mathbb{E}[g^2] = 1 \big\}.$$

It captures arbitrary nonlinear dependence between the two modalities and is a property of the data alone.

Proposition 2 (HGR lower bound on the modality gap)**.** *Assume the embeddings are produced by encoders such that $\mathbb{E}_{\mu_I}[\hat{z}_I] = \mathbb{E}_{\mu_T}[\hat{z}_T] = 0$. Then*

$$\mathbb{E}_\pi\left[\langle \hat{z}_I, \hat{z}_T \rangle\right] \;\leq\; \rho_{\mathrm{HGR}}(X_I, X_T), \qquad \textit{and hence} \qquad 2\big(1 - \rho_{\mathrm{HGR}}(X_I, X_T)\big) \;\leq\; \Delta_{\textit{gap}}.$$

In particular, if $\rho_{\mathrm{HGR}}(X_I, X_T) < 1$, perfect alignment is impossible and the modality gap is bounded away from zero.

This lower bound holds for any pair of centered encoders, regardless of the training objective (Proposition 6 gives a version without centering, with explicit mean terms). The next theorem gives necessary conditions for a CLIP configuration to strictly improve on the uniform distribution.

Theorem 4 (Necessary conditions for improving on the uniform distribution)**.** *Let (μ_I', μ_T', π') satisfy $\mathcal{L}_{\textit{CLIP}}(\mu_I', \mu_T', \pi') < \mathcal{L}_{\textit{CLIP}}(\sigma_d, \sigma_d, (\mathrm{id}_d, \mathrm{id}_d)_{\#}\sigma_d)$. Then:*

i) the modalities are not perfectly aligned, i.e. $\langle \hat{z}_I, \hat{z}_T \rangle = 1$ a.s. fails;

ii) $0 < \Delta_{\textit{gap}} < 2 + 2\tau\varepsilon_d$;

iii) the repelling term is controlled by the gap: $\mathcal{L}^{\mathrm{rep}}_{\mathrm{CLIP}}(\mu_I', \mu_T') < \varepsilon_d - \Delta_{\textit{gap}}/(2\tau)$.

Item (iii) is an algebraic reformulation of the strict-improvement hypothesis: a triplet with paired gap Δ_{gap} improves on the reference if and only if its repelling term lies more than $\Delta_{\text{gap}}/(2\tau)$ below ε_d. Combining this inequality with the lower bound $-1/\tau$ on the repelling term yields item (ii).

In summary, the two components of the CLIP objective, alignment and repelling, are in tension, and any configuration that improves on the uniform distribution must have nonzero modality gap.

3.4 THE PROTOTYPE OBJECTIVE AND ITS POPULATION GEOMETRY

To alleviate the conflict between alignment and repelling, we adapt the prototype construction of Dufumier et al. (2023) to paired multimodal data. The objective is

$$m(\hat{z}_I, \hat{z}_T) = \frac{\hat{z}_I + \hat{z}_T}{2} \in \overline{\mathbb{B}}^d,\ \mathcal{L}_{\text{prot}}(\mu_I, \mu_T, \pi) = \log \mathbb{E}_{\pi\otimes\pi}\left[\exp\left(-\frac{1}{\tau}\left\|m(\hat{z}_I, \hat{z}_T) - m(\hat{z}'_I, \hat{z}'_T)\right\|^2\right)\right],$$

that is, the energy functional of the pushforward $\nu_\pi = m_\#\pi$. The main motivation is that this loss can enforce perfect alignment and uniformity on the sphere simultaneously. Indeed, we show that when perfect alignment is attainable, the optimal prototype law is the uniform distribution, which in turn yields uniform marginals and perfect alignment. When perfect alignment is not attainable, we still identify the optimal prototype law: for a fixed modality gap, it is the uniform distribution on a sphere of radius $R < 1$, where R is determined by the gap and bounded in terms of the maximal correlation. This establishes an explicit relation between the modality gap, the maximal correlation of the data, and the optimal distribution, thereby incorporating into the problem the maximal alignment permitted by the data.

For $d \geq 2$, we define the following quantities: let $\alpha_c(d)$ be the unique solution α of $\Phi(2\alpha) = e^\alpha$, where $\Phi(t) = \mathbb{E}_{\omega,\eta\sim\sigma_d}[e^{t\,\omega\cdot\eta}]$ is the angular moment-generating function on $\mathbb{S}^{d-1}$. And define $\beta_c(d) = (d + 2 - \sqrt{2(d+2)})/2$.

Theorem 5 (Prototype geometry)**.** *We have the following.*

1. ***Unconstrained case.*** *If $\tau \geq 1/\alpha_c(d)$, the unique global minimizer of the population prototype loss is the uniform distribution on the unitary sphere, with diagonal coupling. Thus, paired embeddings coincide and are uniformly distributed on the unit sphere. This temperature threshold is sharp.*

2. ***Geometry at a fixed gap.*** *Fix a paired modality gap Δ_{gap} and let $R^2 = 1 - \Delta_{\text{gap}}/4$. If $\tau \geq R^2/\beta_c(d)$, the prototype loss over all triplets with this gap is minimized precisely when the prototype law is uniform on the sphere of radius R. The radius also satisfies*

$$R \leq \sqrt{\frac{1 + \rho_{\text{HGR}}(X_I, X_T)}{2}}.$$

The temperature conditions are mild at standard CLIP scales. For $d = 512$, as in CLIP ViT-B/32, we have $\alpha_c(512) \approx 422.64$ and $\beta_c(512) \approx 240.97$. Thus, the unrestricted result requires only $\tau \gtrsim 0.00237$, while the fixed-gap condition holds for every $R \leq 1$ whenever $\tau \gtrsim 0.00415$. Both conditions therefore hold throughout $\tau \in [0.01, 0.07]$, a range that covers typical CLIP temperature values.

This highlights the advantage of the prototype loss: whereas the CLIP loss exhibits a tradeoff between alignment and repelling, the prototype loss, a single energy term, admits an optimum that is both perfectly aligned and uniform, with no tension between the two. Moreover, prototype objective admits a linear-time, linear-memory estimator, which we briefly discuss in Appendix C.10.

The two sets of results also differ in scope. For CLIP, Theorems 1 and 2 characterize the minimizers of the alignment and repelling terms separately and establish their incompatibility. For the prototype objective, Theorem 5 (item 1) characterizes the *unique* global population minimizer over unrestricted triplets when $\alpha \leq \alpha_c(d)$, and this minimizer is simultaneously perfectly aligned and uniform.

Finally, uniformity on the sphere of radius R implies asymptotic Gaussianity of fixed-dimensional projections of the prototype law, by the Poincaré–Maxwell lemma (Diaconis & Freedman, 1984), as used by Betser et al. (2026) in the unimodal case. This connects our results to the claim of Balestriero & LeCun (2025) that a Gaussian latent distribution is optimal for downstream tasks, a topic we discuss in the next section.

Theorem 6 (Gaussian convergence of fixed-dimensional prototype projections)**.** *Fix $k \in \mathbb{N}$, let $\alpha = 1/\tau$, and let $R \in (0, 1]$ satisfy $\alpha R^2 \leq \beta_c(d)$. Let ν_{π^*} be the optimal prototype law with modality gap $\Delta_{\text{gap}} = 4(1 - R^2)$, and let $P^d = (P_1^d, \ldots, P_d^d) \sim \nu_{\pi^*}$. Then*

$$\sqrt{d}\,(P_1^d, \ldots, P_k^d) \Longrightarrow \mathcal{N}(0_k, R^2 I_k), \qquad d \to \infty.$$

When $R = 1$, the same holds for the marginals: if $\hat{z}^{M,d} = (\hat{z}_1^{M,d}, \ldots, \hat{z}_d^{M,d}) \sim \mu_M^*$ for $M \in \{I, T\}$, then

$$\sqrt{d}\,(\hat{z}_1^{M,d}, \ldots, \hat{z}_k^{M,d}) \Longrightarrow \mathcal{N}(0_k, I_k), \qquad d \to \infty.$$

4 Experiments

The theoretical analysis yields three predictions: the two terms of the CLIP loss have incompatible minimizers; for inverse temperature $\alpha = 1/\tau \leq \alpha_c(d)$, the unrestricted population minimizer of the prototype objective is uniform and aligned, including at the critical threshold; and, as $d \to \infty$, its fixed-dimensional projections scaled by $\sqrt{d}$ converge in distribution to an isotropic Gaussian. We examine these geometric predictions by fine-tuning a pretrained CLIP checkpoint, except in the experiment of Section 4.2. In each comparison, we change only the loss: architecture, initialization, batches, and evaluation are shared between the CLIP and prototype runs.

4.1 Setup

We test on MS-COCO (Lin et al., 2014), for which we use train2017 for training and val2017 for validation. Each batch contains distinct images, with one caption per image. For CC3M (Sharma et al., 2018), we select 100,000 pairs for training and 5,000 for validation from the first 20 .tar files of the dataset available on Hugging Face. We use early stopping based on validation loss in all experiments.

Metrics. We measure alignment by $\mathbb{E}_\pi \langle u, v \rangle$. The modality gap $\mathbb{E}_\pi \|u - v\|^2$ is an affine function of alignment. It differs from the centroid gap, which we report as $\|\bar{u} - \bar{v}\|$, following Liang et al. (2022). The quantity Δ_{centroid} in Section 3.3 is the square of this gap. Uniformity is measured by the Wang–Isola potential, $\log \mathbb{E}_{i \neq j} \exp(-2\|x_i - x_j\|^2)$. Uniformity excess is the difference between this value and that obtained from an i.i.d. sample of the same size from σ_d. To test marginal Gaussianity, we average the Kolmogorov–Smirnov (KS) distances between the coordinate marginals, scaled by $\sqrt{d}$, and $\mathcal{N}(0, 1)$. Mean recall averages image-to-text and text-to-image recall at ranks 1, 5, and 10.

4.2 Two-dimensional visualization with linear heads on frozen features

For visualization, we train separate $512 \to 2$ bias-free linear heads on the frozen image and text features produced by CLIP ViT-B/32 on MS-COCO. Figure 2 shows these results. With the CLIP loss, the modalities align quickly, but the points initially concentrate near two locations on the circle before spreading out. With the prototype loss, the points spread around the circle first, while paired image and text embeddings align later.

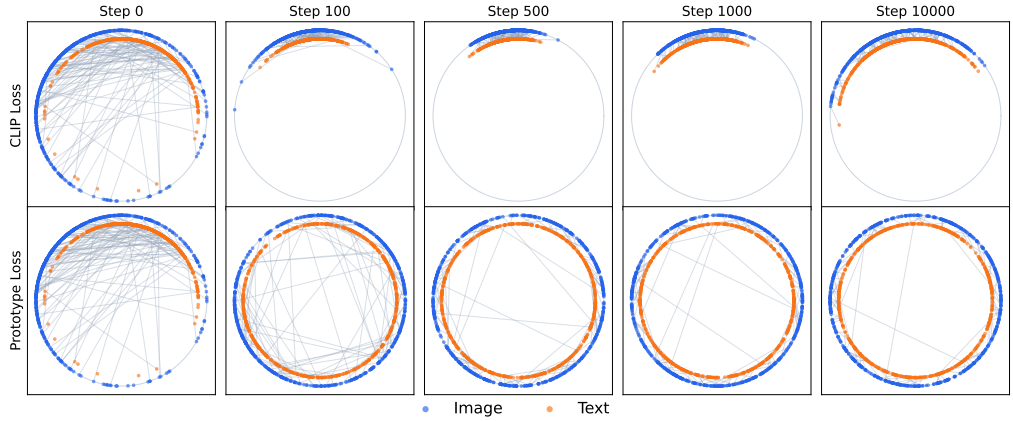


Figure 2: The two objectives applied to the same frozen CLIP features through $512 \to 2$ heads. Columns show steps 0, 100, 500, 1,000, and 10,000; rows correspond to the two objectives. Image embeddings lie on the unit circle. Caption embeddings are drawn at radius 0.9 solely to make paired points visually distinguishable.

4.3 GEOMETRY ON EMBEDDINGS AFTER FINETUNING

Table 1 reports the geometric metrics on MS-COCO. Fine-tuning with the prototype loss improves every reported metric relative to both frozen CLIP and fine-tuning with the CLIP loss. The largest changes are in uniformity excess, the KS distance of the coordinate marginals, and the centroid gap of Liang et al. (2022). The corresponding CC3M results are reported in Appendix D.

The two-dimensional PCA projections in Figure 1 also show a much smaller separation between the image and text embeddings after fine-tuning with the prototype loss. This agrees with the measured centroid gap, which falls from 0.8219 for frozen CLIP to 0.0232. The paired gap decreases as alignment increases, consistent with the relation between these two quantities and with the geometry described in Theorem 5.

Figure 3 tracks alignment and Wang–Isola uniformity (Wang & Isola, 2020) during training, together with their trajectory in the alignment–uniformity plane. Fine-tuning with the CLIP loss changes both metrics little and ends with slightly worse values than frozen CLIP. With the prototype loss, uniformity improves rapidly while alignment initially falls; alignment then rises as training continues. This is the same sequence observed in the two-dimensional visualization: the embeddings spread out first and align later.

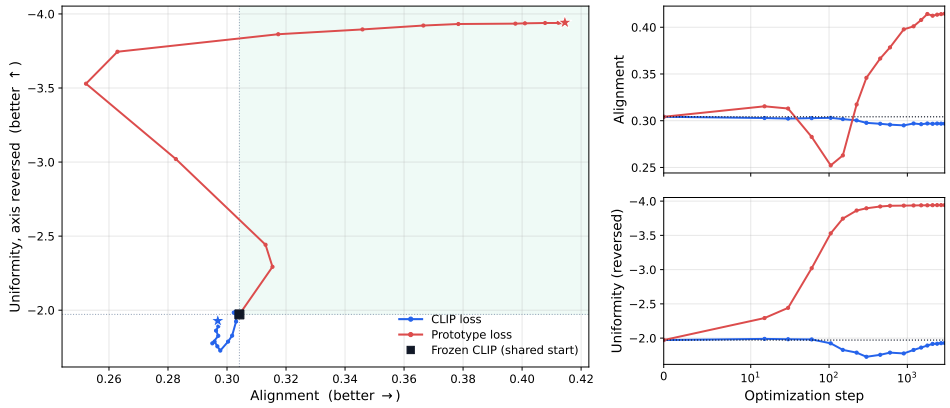


Figure 3: Alignment and uniformity during fine-tuning of CLIP ViT-B/32 on MS-COCO train2017, evaluated on val2017 (3,000 steps and batch size 256).

Metric	Frozen CLIP	CLIP loss	Prototype loss
Alignment	0.3042	0.2969	**0.4145**
Sphere uniformity excess	2.0128	2.0559	**0.0421**
Centroid gap $\|\bar{u} - \bar{v}\|$	0.8219	0.9051	**0.0232**
Modality gap $\mathrm{E}\|u - v\|^2$	1.3916	1.4062	**1.1710**
Marginal Gaussianity (KS)	0.1923	0.1995	**0.0302**
kNN (CIFAR-10)	92.88	94.01	**94.13**
NCM (CIFAR-10)	89.31	91.01	**91.95**
kNN (CIFAR-100)	71.82	76.35	**76.42**
NCM (CIFAR-100)	66.05	70.38	**72.76**

Table 1: Embedding geometry after 3,000 fine-tuning steps on MS-COCO train2017, evaluated on val2017 (higher alignment is better, lower is better for the other geometry metrics), together with downstream top-1 accuracy (kNN and NCM) on CIFAR-10 and CIFAR-100.

4.4 GAUSSIANITY OF THE COORDINATE MARGINALS

We test the Gaussianity of the coordinate marginals predicted by Theorem 6. Figure 4 shows the densities of all $\sqrt{d}$-scaled coordinate marginals of the MS-COCO embeddings against the standard Gaussian density, together with the mean KS distance to $\mathcal{N}(0, 1)$, averaged over coordinates. The

marginals obtained with the prototype loss closely match the standard Gaussian, and their mean KS distance approaches the finite-sample floor. In contrast, the CLIP marginals, although bell-shaped, have non-unit variance and therefore remain far from the standard Gaussian. Overall, the prototype loss yields markedly stronger Gaussianity than CLIP.

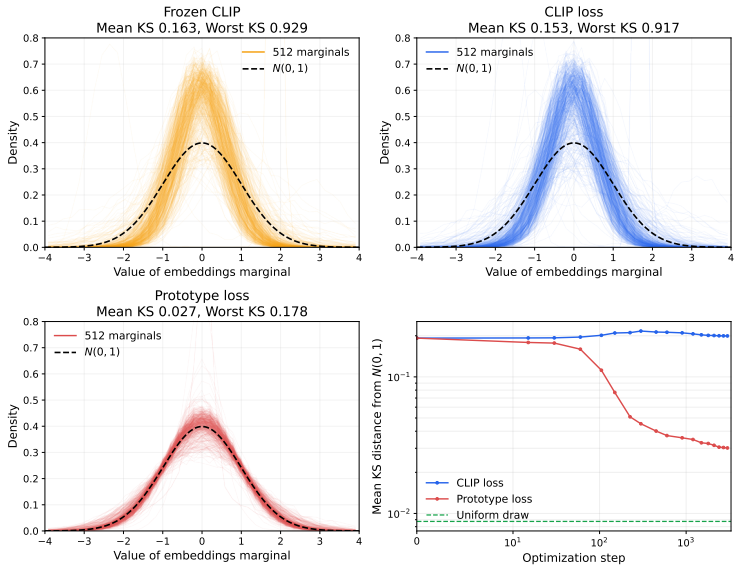


Figure 4: Coordinate-marginal diagnostics for the same unfrozen end-to-end fine-tuning run (one seed, MS-COCO val2017, step 3,000), pooling the embeddings of both towers. Thin curves show the $\sqrt{d}$-scaled marginals against $\mathcal{N}(0, 1)$. The rightmost panel tracks the mean KS distance over training, compared with the finite-sample floor given by an equally sized sample drawn uniformly from the sphere.

4.5 Does the geometry improve downstream task performance?

Finally, we ask whether this geometry translates into better downstream performance. Under certain assumptions, Balestriero & LeCun (2025) identify isotropic Gaussian embeddings as optimal for downstream prediction with linear, kNN, and kernel probes. Our characterization of the prototype optimum, together with the Gaussian limit of its fixed-dimensional projections observed in Section 4.4, motivates testing whether the resulting embeddings perform well on downstream tasks compared with those obtained using the CLIP loss.

We evaluate linear, RBF-kernel, and kNN probes, Nearest Class Mean (NCM) classification, and zero-shot classification on CIFAR-10 and CIFAR-100. For zero-shot classification, we use the prompt template `a photo of a ..` Table 1 reports top-1 accuracy for kNN and NCM, which are the most closely tied to geometry, as they classify based solely on the representation's geometric structure, without any task-specific supervision. Other tasks are in the Supplementary Material. The prototype loss yields higher kNN and NCM accuracy on both datasets, although the kNN differences are small.

Dataset	Model	I→T R@1	T→I R@1	Mean recall
CC3M validation	Frozen CLIP	47.08	44.84	64.77
	CLIP loss	**61.12**	**59.72**	**77.93**
	Prototype loss	55.96	55.36	73.93
MS-COCO val2017	Frozen CLIP	**50.28**	30.28	59.89
	CLIP loss	50.00	**34.52**	**62.62**
	Prototype loss	44.46	32.22	58.39

Table 2: Retrieval after fine-tuning on a 100,000-pair subset of CC3M, evaluated on CC3M validation and MS-COCO val2017.

We also evaluate image–text retrieval using checkpoints fine-tuned on a 100,000-pair subset of CC3M. Table 2 reports results on both CC3M validation and MS-COCO val2017 to assess transfer between datasets. On CC3M validation, fine-tuning with either loss improves on frozen CLIP, with the CLIP loss performing best on all three retrieval metrics. The prototype loss improves text-to-image R@1 over frozen CLIP but yields lower image-to-text R@1 and mean recall.

Taken together, the pattern splits by how each task uses the representation: kNN and NCM, which classify directly from the embedding's geometry with no task-specific fitting, favor the prototype loss, while CLIP keeps the advantage on linear and kernel probes, zero-shot classification, and retrieval, the tasks it was explicitly trained to optimize or that rely on additional fitted structure. We discuss the implications of this pattern for the role of Gaussianity in downstream performance in Section 5.

5 Conclusion

This work develops a population-level account of the geometry induced by multimodal contrastive learning and analyzes a prototype objective built from the midpoints of paired image and text embeddings. Our analysis establishes that the alignment and repelling terms of CLIP have incompatible minimizers, motivating an alternative whose geometric optimum can be characterized directly. When data permits perfect alignment, under the stated temperature conditions, the prototype objective admits a unique population optimum that combines perfect alignment with uniformity. For a fixed positive paired gap, we characterize the optimal prototype law as uniform on a sphere of smaller radius, connecting this radius to the gap and with HGR maximal correlation.

A further consequence is the convergence of appropriately scaled, fixed-dimensional projections of the optimal prototype distribution to an isotropic Gaussian; under perfect alignment, this also applies to the modality distributions themselves. These results characterize how a single prototype energy can promote alignment, uniformity, and Gaussianity, as confirmed by our experiments. The objective reshapes the geometry of pretrained multimodal representations within a modest training budget, and it does so for the reason our theory predicts.

These geometric gains do not translate into the downstream performance we expected. Balestriero and LeCun report that pushing embeddings toward Gaussianity improves downstream results, and since the prototype objective achieves stronger Gaussianity, we expected similar gains to follow. Instead, each objective wins mainly on the tasks it was built for: prototype training improves the metrics that read the latent geometry directly, while CLIP keeps the advantage on zero-shot classification and retrieval, the tasks its loss explicitly targets. This suggest that, even if Gaussianity in the latent space is desirable, or even necessary, is not sufficient on its own to improve downstream performance. Balestriero and LeCun do not evaluate retrieval, so their results say nothing about whether Gaussianizing embeddings preserves the structure retrieval depends on; our results suggest Gaussianizing the space and preserving CLIP's retrieval-favorable geometry are, at least in part, competing goals. What appears to be missing is a loss that fits the latent space toward a Gaussian-type geometry while keeping the properties responsible for CLIP's retrieval and zero-shot performance.

This is primarily a theoretical contribution, with experiments constrained by the available computational resources. The empirical evaluation centers on fine-tuning one pretrained architecture on limited datasets and training budgets, and does not establish the behavior of the prototype objective under large-scale pretraining. Designing and testing a hybrid Gaussianity-retrieval objective, together with experiments on larger and more diverse datasets, multiple architectures, and repeated runs, is left to future work.

AI Use Statement

In this work, we used generative AI tools to assist with software changes and tests, and to revise the manuscript's writing and grammar. We have not used generative AI tools to generate research ideas or experimental results, and synthetic data generation is not applicable to this work. We have reviewed all AI-assisted work. AI-assisted code changes and tests were reviewed and run by the authors before use, and all reported numbers were checked against the original experiment logs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Linear heads are trained on an NVIDIA RTX 4050 GPU with 6 GB of memory; full fine-tuning uses an NVIDIA Tesla V100 with 16 GB. All code needed to reproduce our results is available in an anonymized repository: `https://anonymous.4open.science/r/align_repel_tradeoff_towards_gaussianity`. The repository has one script per experiment, and the paper's results run as seven steps in a fixed order: (i) the two-dimensional experiment on frozen CLIP features of MS-COCO; (ii) fine-tuning all of CLIP ViT-B/32 on MS-COCO; (iii) downstream evaluation on CIFAR-10/100 (zero-shot and few-shot nearest class mean); (iv) building the CC3M subset; (v) fine-tuning on CC3M; (vi) downstream evaluation of the CC3M models; and (vii) retrieval evaluation of the CC3M models on MS-COCO.

All datasets and models are public. We use MS-COCO 2017, CIFAR-10/100 and the official OpenAI CLIP ViT-B/32 checkpoint. For CC3M, our script builds the 100k-pair training subset and the 5k-pair validation subset deterministically from a pinned revision of the public `pixparse/cc3m-wds` WebDataset. Model checkpoints, feature caches and raw data are not included because of their size, and the scripts regenerate them.

REFERENCES

M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. Applied mathematics series. Dover Publications, 1965. ISBN 9780486612720. URL https://books.google.cl/books?id=MtU8uP7XMvoC.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and Scalable Self-Supervised Learning Without the Heuristics, 2025. URL https://arxiv.org/abs/2511.08544.

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=BlSH7gNQSq.

Sergiy V. Borodachov, Douglas P. Hardin, and Edward B. Saff. *Discrete Energy on Rectifiable Sets*. Springer Monographs in Mathematics. Springer New York, New York, NY, 2019. ISBN 978-0-387-84807-5 978-0-387-84808-2. doi: 10.1007/978-0-387-84808-2. URL http://link.springer.com/10.1007/978-0-387-84808-2.

Yichao Cai, Zhen Zhang, Yuhang Liu, and Javen Qinfeng Shi. The geometric mechanics of contrastive representation learning: Alignment potentials, entropic dispersion, and cross-modal divergence. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=X9VtjdBlQw.

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

Zixiang Chen, Yihe Deng, Yuanzhi Li, and Quanquan Gu. Understanding Transferable Representation Learning and Zero-shot Transfer in CLIP. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/f41b6e5af73421e46ceed9cb036e72e7-Paper-Conference.pdf.

Persi Diaconis and David Freedman. Asymptotics of graphical projection pursuit. *The Annals of Statistics*, 12(3):793–815, 1984. doi: 10.1214/aos/1176346703. URL https://doi.org/10.1214/aos/1176346703.

Benoit Dufumier, Carlo Alberto Barbano, Robin Louiset, Edouard Duchesnay, and Pietro Gori. Integrating prior knowledge in contrastive learning with kernel. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8851–8878. PMLR, 2023. URL https://proceedings.mlr.press/v202/dufumier23a.html.

Gerald B. Folland. *A Course in Abstract Harmonic Analysis*. Textbooks in Mathematics. Chapman and Hall/CRC, 2 edition, 2015. ISBN 9781498727136. doi: 10.1201/b19172. URL https://doi.org/10.1201/b19172.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

Panagiotis Koromilas, Theodoros Giannakopoulos, Mihalis Nicolaou, and Yannis Panagakis. Neural collapse by design: Learning class prototypes on the hypersphere, 2026. URL https://openreview.net/forum?id=avFvCd1lPj.

Meir Yossef Levi and Guy Gilboa. The double-ellipsoid geometry of CLIP. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 33999–34019. PMLR, 2025. URL https://proceedings.mlr.press/v267/levi25b.html.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/702f4db7543a7432431df588d57bc7c9-Abstract-Conference.html.

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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Fabian Mager, Hiba Nassar, and Lars Kai Hansen. On the modality gap and the contrastive loss in multi-modal representation learning, 2026. URL https://arxiv.org/abs/2607.10698.

Elizabeth S. Meckes. *The Random Matrix Theory of the Classical Compact Groups*, volume 218 of *Cambridge Tracts in Mathematics*. Cambridge University Press, 2019. doi: 10.1017/9781108303453.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding, 2018. URL https://arxiv.org/abs/1807.03748.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Alfréd Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungaricae*, 10:441–451, 1959. doi: 10.1007/BF02024507. URL https://doi.org/10.1007/BF02024507.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL https://aclanthology.org/P18-1238/.

Tongzhou Wang and Phillip Isola. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 2020. URL https://proceedings.mlr.press/v119/wang20k.html.

Zhen-Hang Yang and Jing-Feng Tian. Convexity of ratios of the modified Bessel functions of the first kind with applications. *Revista Matemática Complutense*, 36:799–825, 2023. doi: 10.1007/s13163-022-00439-w. URL https://doi.org/10.1007/s13163-022-00439-w.

Lingjie Yi, Raphael Douady, and Chao Chen. Decipher the Modality Gap in Multimodal Contrastive Learning: From Convergent Representations to Pairwise Alignment, 2025. URL https://arxiv.org/abs/2510.03268.

A THEORETICAL SETTING

This appendix fixes notation and groups the objects and the analytic facts that Appendices B and C use repeatedly. Sections A.1–A.3 restate the two objectives in the population form in which they are analysed; §A.4 makes explicit the encoder model that underlies every statement involving maximal correlation; and §A.5 records the analytic tools used on both sides of the paper, of which the angular moment generating function Φ is by far the most important: it appears in the value of the CLIP repelling term at the uniform distribution with diagonal coupling (Appendix B) and in the radial analysis of the prototype loss (Appendix C).

A.1 NOTATION

Throughout the paper, unless stated otherwise, the space dimension satisfies $d \geq 2$, in which case the setting for the sphere $\mathbb{S}^{d-1}$ is well defined.

Sets and measures. We write $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$ for the unit sphere and $\overline{\mathbb{B}}^d = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$ for the closed unit ball, both in $\mathbb{R}^d$ with the Euclidean inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$. For a metric space $\mathcal{X}$, $\mathcal{P}(\mathcal{X})$ denotes the Borel probability measures on $\mathcal{X}$. We write σ_d for the uniform (normalized surface) measure on $\mathbb{S}^{d-1}$ and, for $r \in (0, 1]$, σ_r for the uniform measure on the sphere $r \cdot \mathbb{S}^{d-1}$ of radius r; thus $\sigma_1 = \sigma_d$, and we set $\sigma_0 = \delta_0$. The identity map on $\mathbb{R}^d$ is id_d, and $T_{\#}\mu$ is the pushforward of μ by a Borel map T. Finally δ_x is the Dirac mass at x, $O(d)$ is the orthogonal group and H its Haar measure.

Couplings and triplets. A *coupling* of $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ is a measure π on $\mathcal{X} \times \mathcal{Y}$ whose marginals are μ and ν; we write $\Gamma(\mu, \nu)$ for the set of such couplings. The coupling is the population object corresponding to the positive pairs of a contrastive objective. A *triplet* is a pair of marginals together with a coupling of them,

$$(\mu_I, \mu_T, \pi), \qquad \mu_I, \mu_T \in \mathcal{P}(\mathbb{S}^{d-1}), \qquad \pi \in \Gamma(\mu_I, \mu_T),$$

and it is the variable over which both population problems in this paper are posed. We call $\big(\sigma_d, \sigma_d, (\mathrm{id}_d, \mathrm{id}_d)_{\#}\sigma_d\big)$ the *uniform diagonal triplet*: both modalities are uniform on the sphere and paired embeddings coincide almost surely.

Temperature. Throughout, $\tau > 0$ is the temperature and $\alpha := 1/\tau$ the inverse temperature. The CLIP analysis reads more naturally in τ and the prototype analysis in α; we keep both and always state which is in force.

Special functions. Γ is the Gamma function and I_ν the modified Bessel function of the first kind of order ν. Two distinct objects in this paper carry the letter I, and we separate them by their brackets: $I_\nu(t)$, with parentheses and a subscript order, is always the Bessel function, while $I_\alpha[\mu]$, with square brackets and a measure argument, is always the Gaussian energy functional of §A.3. The same square-bracket notation indexed by a kernel, $I_K[\mu]$, denotes the energy of μ under that kernel; it appears in Appendix B with $K = K_f$, the kernel $K_f(x, y) = f(\|x - y\|^2)$ induced by a radial profile f.

Norm and dot product identities on the sphere. For $u, v \in \mathbb{S}^{d-1}$,

$$\|u - v\|^2 = 2 - 2\langle u, v \rangle, \qquad \text{equivalently} \qquad \langle u, v \rangle = 1 - \tfrac{1}{2}\|u - v\|^2. \tag{1}$$

It is used below without further comment to pass between inner products and squared distances.

A.2 THE CLIP OBJECTIVE

The batch-level loss. CLIP consists of two parametrized encoders, an image encoder $f_\theta^I : \mathcal{X}_I \to \mathbb{R}^{d_I}$ and a text encoder $f_\theta^T : \mathcal{X}_T \to \mathbb{R}^{d_T}$, whose representations are projected into a shared embedding space by linear maps $W_I \in \mathbb{R}^{d \times d_I}$ and $W_T \in \mathbb{R}^{d \times d_T}$ and normalized,

$$\hat{z}_I = \frac{W_I f_\theta^I(x_I)}{\|W_I f_\theta^I(x_I)\|}, \qquad \hat{z}_T = \frac{W_T f_\theta^T(x_T)}{\|W_T f_\theta^T(x_T)\|},$$

where we assume throughout that $W_I f_\theta^I(x_I) \neq 0$ and $W_T f_\theta^T(x_T) \neq 0$ almost surely, so that $\hat{z}_I, \hat{z}_T$ are well defined; this holds generically and we do not revisit it. Given a batch of N image–text pairs, the logits matrix is $\mathbf{L}_{ij} = \langle \hat{z}_{I,i}, \hat{z}_{T,j} \rangle / \tau$, and with labels $[0, 1, \ldots, N-1]$ the CLIP loss is the symmetrized cross-entropy, which is also a symmetrized InfoNCE loss:

$$\mathcal{L}_{\mathrm{CLIP},N} = -\frac{1}{2}\left(\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp(\mathbf{L}_{ii})}{\sum_{j=1}^{N}\exp(\mathbf{L}_{ij})}\right) + \frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp(\mathbf{L}_{ii})}{\sum_{j=1}^{N}\exp(\mathbf{L}_{ji})}\right)\right).$$

From batch to population loss. The population loss defined below is *not* the limit of $\mathcal{L}_{\mathrm{CLIP},N}$ itself: each denominator $\sum_{j=1}^{N}\exp(\mathbf{L}_{ij})$ sums N terms of order $O(1)$, so $\mathcal{L}_{\mathrm{CLIP},N} \to \infty$ as $N \to \infty$. If the pairs $(x_{I,i}, x_{T,i})_{i=1}^{N}$ are i.i.d. draws from the joint law of (X_I, X_T) inducing μ_I, μ_T, π, the law of large numbers applied to each in-batch sum of negatives gives

$$\mathcal{L}_{\mathrm{CLIP},N} - \log N \xrightarrow{\text{prob.}} \mathcal{L}_{\mathrm{CLIP}}(\mu_I, \mu_T, \pi) \qquad (N \to \infty),$$

which is analogous to the unimodal result of Wang & Isola (2020, Theorem 1), who prove exactly this $-\log M$ convergence, at rate $O(M^{-1/2})$, for a single encoder and negatives drawn from one data distribution; we do not reprove the two-modality analogue here. Every population statement in this paper is about $\mathcal{L}_{\mathrm{CLIP}}$, i.e. the batch loss up to the additive, batch-size-dependent constant $\log N$, and not about $\mathcal{L}_{\mathrm{CLIP},N}$ itself.

The population loss and its two terms. Let $\mu_I, \mu_T \in \mathcal{P}(\mathbb{S}^{d-1})$ be the laws of the normalized image and text embeddings and $\pi \in \Gamma(\mu_I, \mu_T)$ the positive-pair coupling. Up to the additive $\log N$ term made precise above, the loss splits, following Chen et al. (2024), as $\mathcal{L}_{\mathrm{CLIP}} = \mathcal{L}_{\mathrm{CLIP}}^{\mathrm{align}} + \mathcal{L}_{\mathrm{CLIP}}^{\mathrm{rep}}$, with

$$\mathcal{L}_{\mathrm{CLIP}}^{\mathrm{align}}(\mu_I, \mu_T, \pi) = -\frac{1}{\tau}\,\mathbb{E}_{(\hat{z}_I, \hat{z}_T)\sim\pi}\big[\langle \hat{z}_I, \hat{z}_T \rangle\big],$$

$$\mathcal{L}_{\mathrm{CLIP}}^{\mathrm{rep}}(\mu_I, \mu_T) = \frac{1}{2}\,\mathbb{E}_{\hat{z}_I\sim\mu_I}\Big[\log \mathbb{E}_{\hat{z}'_T\sim\mu_T} e^{\langle \hat{z}_I, \hat{z}'_T\rangle/\tau}\Big] + \frac{1}{2}\,\mathbb{E}_{\hat{z}_T\sim\mu_T}\Big[\log \mathbb{E}_{\hat{z}'_I\sim\mu_I} e^{\langle \hat{z}'_I, \hat{z}_T\rangle/\tau}\Big].$$

The population problem is to minimize $\mathcal{L}_{\mathrm{CLIP}}(\mu_I, \mu_T, \pi)$ over triplets. The first term depends on the coupling and the second only on the marginals; §B.1 characterizes the minimizers of each of them on its own.

The two modality gaps. We use throughout the *paired* gap and the *centroid* gap,

$$\Delta_{\mathrm{gap}} := \mathbb{E}_\pi\big[\|\hat{z}_I - \hat{z}_T\|^2\big], \qquad \Delta_{\mathrm{centroid}} := \|\hat{m}_I - \hat{m}_T\|^2, \qquad \hat{m}_I := \mathbb{E}_{\mu_I}[\hat{z}_I], \quad \hat{m}_T := \mathbb{E}_{\mu_T}[\hat{z}_T],$$

the latter being the quantity usually reported in the literature (Liang et al., 2022). Substituting (1) into the alignment term gives the identity that drives the whole of Appendix B: the alignment term *is* the paired gap, up to an affine change of variable,

$$\mathcal{L}_{\mathrm{CLIP}}^{\mathrm{align}}(\mu_I, \mu_T, \pi) = \frac{1}{2\tau}\,\mathbb{E}_\pi\big[\|\hat{z}_I - \hat{z}_T\|^2\big] - \frac{1}{\tau} = \frac{\Delta_{\mathrm{gap}}}{2\tau} - \frac{1}{\tau}. \tag{2}$$

Since $\Delta_{\mathrm{gap}} \in [0, 4]$, the alignment term ranges over $[-1/\tau, 1/\tau]$.

A.3 The prototype objective

Midpoints and the prototype law. The prototype construction replaces each positive pair by its midpoint. Write

$$m(u, v) = \frac{u+v}{2}, \qquad p = m(\hat{z}_I, \hat{z}_T) = \frac{\hat{z}_I + \hat{z}_T}{2}, \qquad \nu_\pi := m_\#\pi.$$

By (1), for $\hat{z}_I, \hat{z}_T \in \mathbb{S}^{d-1}$,

$$\|p\|^2 = \frac{1}{4}\|\hat{z}_I + \hat{z}_T\|^2 = \frac{1 + \langle \hat{z}_I, \hat{z}_T \rangle}{2} = \frac{1 + \cos\theta}{2}, \tag{3}$$

where θ is the angle between the two vectors. As θ ranges over $[0, \pi]$ the norm $\|p\|$ ranges over $[0, 1]$, so $p \in \overline{\mathbb{B}}^d$ and $\nu_\pi \in \mathcal{P}(\overline{\mathbb{B}}^d)$. A pair sits on the boundary sphere exactly when it is perfectly aligned (Lemma 10).

The loss as a Gaussian energy. For $\alpha > 0$ define the *Gaussian energy functional* on $\mathcal{P}(\overline{\mathbb{B}}^d)$,

$$I_\alpha[\mu] \coloneqq \iint k_\alpha(x,y)\, d\mu(x)\, d\mu(y), \qquad k_\alpha(x,y) \coloneqq e^{-\alpha\|x-y\|^2},$$

and the product between two measures $\mu,\ \nu$ is

$$\langle \mu, \nu \rangle = \iint k_\alpha(x,y)\, d\mu(x)\, d\nu(y).$$

The prototype loss of Section 3.4 is exactly the logarithm of this energy evaluated at the prototype law:

$$\mathcal{L}_{\text{prot}}(\mu_I, \mu_T, \pi) = \log \mathbb{E}_{p,p'\sim\nu_\pi}\left[\exp\left(-\frac{\|p-p'\|^2}{\tau}\right)\right] = \log I_{1/\tau}[\nu_\pi].$$

The population loss is not the expectation of the minibatch estimator. Training minimizes the logarithm U-statistic $\widehat{\mathcal{L}}_N \coloneqq \log\big[\frac{1}{N(N-1)}\sum_{i\neq j} e^{-\alpha\|p_i-p_j\|^2}\big]$ over a batch of N prototypes, with $\alpha = 1/\tau$. Since the logarithm does not commute with expectation, $\mathbb{E}[\widehat{\mathcal{L}}_N] \neq \log I_\alpha[\nu_\pi]$ in general. At $N = 2$ the two off-diagonal terms coincide, so $\widehat{\mathcal{L}}_2 = -\alpha\|p_1 - p_2\|^2$ exactly and $\mathbb{E}[\widehat{\mathcal{L}}_2] = -2\alpha\big(\mathbb{E}\|p\|^2 - \|\mathbb{E}p\|^2\big)$: every centered law supported on $\mathbb{S}^{d-1}$ attains the same value, including non-uniform examples such as $\frac{1}{2}\delta_{e_1} + \frac{1}{2}\delta_{-e_1}$. The population theorem of Section 3.4 remains valid as a statement about $\log I_\alpha[\nu_\pi]$; it does not by itself characterize the minimizers of the finite-N training objective $\mathbb{E}[\widehat{\mathcal{L}}_N]$, and we do not study here how the two are related as N grows.

The modality gap and mean-square radius relation. Taking expectations in (3) and using $\Delta_{\text{gap}} = 2 - 2\,\mathbb{E}_\pi[\langle \hat{z}_I, \hat{z}_T\rangle]$ shows that fixing the paired gap is the same as fixing the second moment of the prototype law. Then,

$$R^2 \coloneqq \mathbb{E}_{\nu_\pi}\big[\|p\|^2\big] = \frac{1 + \mathbb{E}_\pi[\langle \hat{z}_I, \hat{z}_T\rangle]}{2} = 1 - \frac{\Delta_{\text{gap}}}{4}, \qquad R \in [0,1]. \tag{4}$$

We call R the root of the *mean-square radius* of the prototype law; perfect alignment is $R = 1$. Fixing this second moment does not impose $\|p\| = R$ almost surely or spherical support. It is the variable in which the prototype objective is best understood: Appendix C shows that, under the sufficient condition $\alpha R^2 \leq \beta_c(d)$ on inverse temperature, including equality, the optimal prototype law at fixed second moment R^2 is the uniform measure σ_R on the sphere of that radius, and that the optimal value decreases strictly in R.

A.4 Encoder-induced embeddings and maximal correlation

Every statement in this paper involving the Hirschfeld–Gebelein–Rényi maximal correlation makes explicit an assumption that is implicit in the CLIP formulation of Section 3.3: that the embeddings are produced by encoders acting on each modality separately. We record it once here.

Assumption 1 (Encoder-induced triplet)**.** *There exist measurable maps*

$$F_I : \mathcal{X}_I \to \mathbb{S}^{d-1}, \qquad F_T : \mathcal{X}_T \to \mathbb{S}^{d-1}, \qquad \hat{z}_I = F_I(X_I), \quad \hat{z}_T = F_T(X_T),$$

where (X_I, X_T) is the raw image–text pair with marginals μ_X^I, μ_X^T, so that $\mu_I = (F_I)_\# \mu_X^I$, $\mu_T = (F_T)_\# \mu_X^T$, and the coupling $\pi = (F_I, F_T)_\#\,\mathrm{law}(X_I, X_T)$ is the one induced by the positive-pair distribution. We write $F_I = (F_{I,1}, \ldots, F_{I,d})$ and $F_T = (F_{T,1}, \ldots, F_{T,d})$ for the coordinate functions.

Assumption 1 fixes a triplet, not a pair of encoders: whenever (μ_I, μ_T, π) is a minimizer of one of the two population objectives, any two encoder pairs (F_I, F_T) and (F_I', F_T') satisfying the assumption for that same triplet are equally optimal. Uniqueness statements in this paper are uniqueness of the optimal triplet, not of the encoder parameters realizing it; e.g. post-composing F_I, F_T with a common rotation, or with any measure-preserving relabeling of $\mathcal{X}_I, \mathcal{X}_T$ that leaves the induced law unchanged, gives another encoder pair inducing the same triplet.

Definition 1 (Maximal correlation). *The Hirschfeld–Gebelein–Rényi maximal correlation (Rényi, 1959) of the pair (X_I, X_T) is*

$$\rho_{\mathrm{HGR}}(X_I, X_T) \coloneqq \sup\big\{\, \mathbb{E}[f(X_I)\, g(X_T)] : f,\, g \in L^2,\ \mathbb{E}[f] = \mathbb{E}[g] = 0,\ \mathbb{E}[f^2] = \mathbb{E}[g^2] = 1 \,\big\}.$$

If X_I or X_T is almost surely constant, no pair (f, g) satisfies the unit-variance constraints and the supremum is over the empty set; by convention we set $\rho_{\mathrm{HGR}}(X_I, X_T) \coloneqq 0$ in that degenerate case, consistent with the trivial covariance bound of Lemma 3.

A.5 Analytic tools

A.5.1 The Wasserstein distance

Definition 2 (Wasserstein distance). *For measures $\mu,\ \nu \in \mathcal{P}(\mathcal{X})$ with finite second moments, the 2-Wasserstein distance is*

$$\mathcal{W}_2(\mu, \nu) = \inf_{\pi \in \Gamma(\mu,\nu)} \sqrt{\mathbb{E}_{(x,y)\sim\pi}[\|x - y\|^2]}.$$

It vanishes if and only if $\mu = \nu$, and by (2) the minimization of the CLIP alignment term over triplets is exactly the quadratic optimal transport problem that it solves.

A.5.2 The angular moment generating function

The following function is the single analytic object to which both halves of the theory reduce.

Lemma 1 (Closed form and power series of Φ). *Let ω, η be independent random points uniformly distributed on $\mathbb{S}^{d-1}$, $d \geq 2$, and define*

$$\Phi(t) \coloneqq \mathbb{E}\big[e^{\,t\,\omega\cdot\eta}\big], \qquad t \in \mathbb{R}.$$

Then

$$\Phi(t) \;=\; \Gamma\!\left(\frac{d}{2}\right)\left(\frac{t}{2}\right)^{1-d/2} I_{d/2-1}(t) \;=\; \sum_{k=0}^{\infty} c_k\, t^{2k},$$

with

$$c_0 = 1, \quad c_k = \frac{1}{2^k\, k!\, d(d+2)\cdots(d+2k-2)} \;\; (k \geq 1).$$

In particular Φ is even and positive, and strictly increasing on $[0, \infty)$, extended by continuity to $t = 0$ and by evennes to $\mathbb{R}$. Moreover, using the recurrence $I_\nu'(t) = I_{\nu+1}(t) + \frac{\nu}{t} I_\nu(t)$ (Abramowitz & Stegun, 1965),

$$\Phi'(t) = R_{d/2-1}(t)\, \Phi(t), \qquad R_\nu(t) \coloneqq \frac{I_{\nu+1}(t)}{I_\nu(t)}. \tag{5}$$

Proof. By rotational invariance the law of $s = \omega \cdot \eta$ has density $c_d\,(1 - s^2)^{(d-3)/2}$ on $(-1, 1)$, with $c_d = \Gamma(d/2)/\big(\sqrt{\pi}\,\Gamma((d-1)/2)\big)$. Poisson's integral representation of the modified Bessel function (Abramowitz & Stegun, 1965), valid for $\nu > -1/2$,

$$I_\nu(t) = \frac{(t/2)^\nu}{\sqrt{\pi}\,\Gamma(\nu + \frac{1}{2})} \int_{-1}^{1} e^{ts}\,(1 - s^2)^{\nu - 1/2}\, ds,$$

applied with $\nu = d/2 - 1$ (so that $\nu - \frac{1}{2} = \frac{d-3}{2}$) gives

$$\int_{-1}^{1} e^{ts}(1 - s^2)^{(d-3)/2}\, ds = \sqrt{\pi}\,\Gamma\big(\tfrac{d-1}{2}\big)\,\big(\tfrac{t}{2}\big)^{1-d/2}\, I_{d/2-1}(t),$$

and multiplying by c_d yields the closed form. For the series, insert

$$I_\nu(t) = \sum_{k\geq 0} (t/2)^{2k+\nu} / \big(k!\,\Gamma(k + \nu + 1)\big)$$

with $\nu = d/2 - 1$:

$$\Phi(t) = \Gamma\big(\tfrac{d}{2}\big) \sum_{k\geq 0} \frac{(t/2)^{2k}}{k!\,\Gamma(k + d/2)} = \sum_{k\geq 0} \frac{t^{2k}}{2^{2k}\,k!} \cdot \frac{\Gamma(d/2)}{\Gamma(k + d/2)},$$

and

$$\Gamma(k + d/2)/\Gamma(d/2) = \tfrac{d}{2} \cdot \tfrac{d+2}{2} \cdots \tfrac{d+2k-2}{2} = 2^{-k}\, d(d+2)\cdots(d+2k-2)$$

gives the stated c_k. All $c_k > 0$, whence positivity, evenness, and strict monotonicity on $[0, \infty)$. For (5), differentiate the closed form:

$$\frac{\Phi'(t)}{\Phi(t)} = \frac{1 - d/2}{t} + \frac{I'_{d/2-1}(t)}{I_{d/2-1}(t)} = \frac{1 - d/2}{t} + \frac{I_{d/2}(t)}{I_{d/2-1}(t)} + \frac{d/2 - 1}{t} = R_{d/2-1}(t). \qquad \square$$

We use the following standard facts about the Bessel ratio $R_\nu(t) = I_{\nu+1}(t)/I_\nu(t)$, restricted here to $\nu \geq 0$, which covers every application $\nu = d/2 - 1$ with $d \geq 2$. At $t = 0$, the ratio is defined by continuous extension. For the series and asymptotics see Abramowitz & Stegun (1965); for monotonicity and concavity see Yang & Tian (2023, Section 1), which records Watson's result for $\nu \geq -1/2$:

(R1) $R_\nu(0) = 0$ and $R_\nu(t) < 1$ for all $t \geq 0$;

(R2) R_ν is strictly increasing and strictly concave on $[0, \infty)$;

(R3) $R_\nu(t) = \dfrac{t}{2(\nu + 1)} + O(t^3)$ as $t \to 0^+$, and $R_\nu(t) \to 1$ as $t \to \infty$.

Combining (R2) and (R3) with $R_\nu(0) = 0$ gives the bound

$$R_\nu(t) < \frac{t}{2(\nu + 1)} \qquad \text{for every } t > 0. \tag{6}$$

Where the angular moment generating function appears. This function appears twice in the proofs of this work. On the CLIP side, Proposition 3 shows that its value at the uniform reference is

$$\varepsilon_d = \mathcal{L}^{\mathrm{rep}}_{\mathrm{CLIP}}(\sigma_d, \sigma_d) = \log \Phi(1/\tau).$$

On the prototype side, the same computation shows that the energy between two uniform spheres of radii r and s is

$$g(r, s) = \iint e^{-\alpha \|r\omega - s\eta\|^2}\, d\sigma_d(\omega)\, d\sigma_d(\eta) = e^{-\alpha(r^2+s^2)}\, \Phi(2\alpha r s),$$

which is precisely the kernel of the one-dimensional problem solved in Appendix C. Remarkably, this single object, which appears somewhat involved at first sight, governs the proofs for both objectives studied in this work.

B Theoretical Results for the CLIP Loss

This appendix proves the results of Section 3.3. All notation, the population form of the objective, and the analytic tools are those of Appendix A; in particular we use the alignment–gap identity (2) and the angular moment generating function Φ of Lemma 1 throughout.

The argument proceeds in five steps. §B.1 minimizes the two terms of the loss separately and shows that their optima are opposed. §B.2 evaluates the repelling term at the uniform distribution with diagonal coupling, obtaining the constant $\varepsilon_d = \log \Phi(1/\tau)$ together with a non-asymptotic two-sided bound on it. §B.3 proves the global lower bound $-1/\tau$ for CLIP loss. §B.4 combines both to prove the ε_d-minimality of the uniform distribution. §B.5 derives the necessary conditions for improving the uniform reference in Theorem 4, and §B.6 derives HGR based bounds for modality gap.

B.1 The two CLIP loss terms

The alignment term depends on the coupling and the repelling term only on the marginals. We minimize each term on its own, over all triplets, and find that their minimizer sets are disjoint: the first forces $\langle \hat{z}_I, \hat{z}_T \rangle = 1$ almost surely, the second $\langle \hat{z}_I, \hat{z}_T \rangle = -1$.

B.1.1 PROOF OF THEOREM 1

Theorem 1 states that the minimizers of $\mathcal{L}_{\text{CLIP}}^{\text{align}}$ are exactly the diagonal triplets $\big(\mu, \mu, (\mathrm{id}_d, \mathrm{id}_d)_{\#}\mu\big)$, with optimal value $-1/\tau$. By (2) the alignment term is the expected squared distance between paired embeddings, up to an affine change of variable. Minimizing an expected squared distance over all couplings between two probability measures is precisely the *quadratic optimal transport* problem, whose minimum is the squared Wasserstein-2 distance $\mathcal{W}_2^2(\mu_I, \mu_T)$ of Definition 2. That distance is zero if and only if $\mu_I = \mu_T$, so perfect alignment forces the two marginals to be equal.

Proof. By (2),

$$\mathcal{L}_{\text{CLIP}}^{\text{align}}(\mu_I, \mu_T, \pi) = \frac{1}{2\tau}\mathbb{E}_{(\hat{z}_I, \hat{z}_T)\sim\pi}\big[\|\hat{z}_I - \hat{z}_T\|^2\big] - \frac{1}{\tau}.$$

Since the $-1/\tau$ term is a constant, minimizing $\mathcal{L}_{\text{CLIP}}^{\text{align}}$ over all (μ_I, μ_T, π) is equivalent to minimizing the expected squared distance. At fixed marginals the inner minimization is the quadratic optimal transport problem,

$$\inf_{\pi\in\Gamma(\mu_I,\mu_T)} \mathbb{E}_{\pi}\big[\|\hat{z}_I - \hat{z}_T\|^2\big] = \mathcal{W}_2^2(\mu_I, \mu_T),$$

so minimizing over triplets amounts to minimizing that distance over the marginals,

$$\inf_{\mu_I,\mu_T\in\mathcal{P}(\mathbb{S}^{d-1})} \mathcal{W}_2^2(\mu_I, \mu_T) = 0.$$

The value 0 is attained, by $\mu_I = \mu_T$ with the identity coupling $T = \mathrm{id}_d$, and it is the global minimum. Therefore, every triplet with $\mu_I \equiv \mu_T$ and $\pi = (\mathrm{id}_d, \mathrm{id}_d)_{\#}\mu_I$ belongs to the optimal set of $\mathcal{L}_{\text{CLIP}}^{\text{align}}$. Conversely, any global minimizer must make the expected squared distance equal to zero, so $\hat{z}_I = \hat{z}_T$ π-a.s.; hence the marginals are equal and the coupling is supported on the diagonal. Thus the optimizer is not a single measure, but the family of diagonal couplings over arbitrary $\mu \in \mathcal{P}(\mathbb{S}^{d-1})$. □

B.1.2 PROOF OF THEOREM 2

Theorem 2 states that the repelling term is minimized exactly when the two embedding distributions are concentrated at antipodal points z and $-z$, achieving the value $-1/\tau$. The argument is a direct application of the bound $\langle \hat{z}_I, \hat{z}'_T\rangle \geq -1$ (which follows from the unit-sphere constraint) to show that $-1/\tau$ is a lower bound, and then a verification that antipodal point masses achieve it.

Proof. Since $\hat{z}_I, \hat{z}'_T \in \mathbb{S}^{d-1}$, we have $\langle \hat{z}_I, \hat{z}'_T\rangle \geq -1$, so

$$\exp\big(\langle \hat{z}_I, \hat{z}'_T\rangle/\tau\big) \geq \exp(-1/\tau).$$

Taking expectations and then a logarithm (which preserves the inequality since $\log$ is increasing):

$$\log\left(\mathbb{E}_{\hat{z}'_I\sim\mu_I} \exp\left(\frac{\langle \hat{z}'_I, \hat{z}_T\rangle}{\tau}\right)\right) \geq \log\exp(-1/\tau) = -\frac{1}{\tau}.$$

Averaging over $\hat{z}_T \sim \mu_T$ and symmetrising across both terms of the repelling loss gives $\mathcal{L}_{\text{CLIP}}^{\text{rep}}(\mu_I, \mu_T) \geq -1/\tau$.

Equality requires $\langle \hat{z}_I, \hat{z}'_T\rangle = -1$ a.s., i.e., every image embedding is opposite every text embedding. This condition forces both supports to be singletons: if two distinct image points both had to be opposite every text point, they would have to be equal. Hence all equality cases are antipodal point masses, $\mu_I = \delta_z$ and $\mu_T = \delta_{-z}$:

$$\mathcal{L}_{\text{CLIP}}^{\text{rep}}(\delta_z, \delta_{-z}) = \log\exp\left(\frac{\langle z, -z\rangle}{\tau}\right) = -\frac{1}{\tau},$$

confirming that antipodal point masses achieve the minimum. □

B.2 THE REPELLING TERM AT THE UNIFORM DISTRIBUTION

The two theorems above characterize the terms separately, not the minimizer of their sum. To compare candidate triplets for the complete CLIP loss we need a reference value, and the natural candidate is the uniform diagonal triplet. Its alignment term is $-1/\tau$ by Theorem 1; this subsection computes its repelling term, which is the constant ε_d appearing throughout the paper.

Proposition 3. *We have that*

$$\mathcal{L}^{\mathrm{rep}}_{CLIP}(\sigma_d, \sigma_d) = \log \Phi(1/\tau) = \varepsilon_d \coloneqq \log\left((2\tau)^{(d-2)/2} I_{\frac{d-2}{2}}\left(\frac{1}{\tau}\right)\Gamma\left(\frac{d}{2}\right)\right),$$

where Γ is the Gamma function and I_ν is the modified Bessel function of first kind of order ν. Furthermore $\varepsilon_d > 0$ for every $d \geq 2$, and $\varepsilon_d \to 0^+$ as $d \to \infty$.

Proof. Fix $\hat{z}_T \in \mathbb{S}^{d-1}$. Because σ_d is invariant under every rotation of $\mathbb{S}^{d-1}$, and the integrand depends on $\hat{z}'_I$ only through $\langle \hat{z}'_I, \hat{z}_T\rangle$, the inner expectation does not depend on $\hat{z}_T$:

$$\mathbb{E}_{\hat{z}'_I \sim \sigma_d}\left[e^{\langle \hat{z}'_I, \hat{z}_T\rangle/\tau}\right] = \mathbb{E}_{\omega,\eta\sim\sigma_d}\left[e^{\omega\cdot\eta/\tau}\right] = \Phi(1/\tau),$$

with Φ the angular moment generating function of Lemma 1. The outer expectation is therefore an expectation of a constant, and the same holds for the symmetric term, so

$$\mathcal{L}^{\mathrm{rep}}_{\mathrm{CLIP}}(\sigma_d, \sigma_d) = \log \Phi(1/\tau).$$

The closed form follows from Lemma 1 evaluated at $t = 1/\tau$, where $(t/2)^{1-d/2} = (2\tau)^{d/2-1}$ and $d/2 - 1 = (d-2)/2$:

$$\log \Phi(1/\tau) = \log\Big(\Gamma\big(\tfrac{d}{2}\big)\,(2\tau)^{(d-2)/2} I_{(d-2)/2}\big(\tfrac{1}{\tau}\big)\Big) = \varepsilon_d.$$

Both the strict positivity of ε_d and the limit $\varepsilon_d \to 0$ follow from the two-sided non-asymptotic bound of Proposition 4 below, whose proof is independent of this one. □

Lemma 2 (Proposition A.11.2, Borodachov et al. (2019))**.** *Let $f : (0, 4] \to \mathbb{R}$ be continuous, write $K_f(x, y) \coloneqq f(\|x - y\|^2)$ for the kernel it induces, and suppose the energy*

$$I_{K_f}[\sigma_d] = \int_{\mathbb{S}^{d-1}} \int_{\mathbb{S}^{d-1}} K_f(x, y)\, d\sigma_d(x) d\sigma_d(y)$$

is finite. Then

$$I_{K_f}[\sigma_d] = \frac{2^{d-2}\,\Gamma\big(\frac{d}{2}\big)}{\sqrt{\pi}\,\Gamma\big(\frac{d-1}{2}\big)} \int_0^1 f(4u)\, u^{\frac{d-3}{2}} (1-u)^{\frac{d-3}{2}}\, du.$$

Proof of Proposition 3. Set $\alpha = 1/\tau$ and $\nu = d/2 - 1$. By rotational invariance of σ_d, the quantity

$$A_\alpha(v) := \int_{\mathbb{S}^{d-1}} e^{\alpha\langle u, v\rangle}\, d\sigma_d(u)$$

is independent of $v \in \mathbb{S}^{d-1}$. Hence $\mathcal{L}^{\mathrm{rep}}_{\mathrm{CLIP}}(\sigma_d, \sigma_d) = \log A_\alpha(v)$.

For $u \sim \sigma_d$, the random variable $t = \langle u, v\rangle$ has density

$$\frac{\Gamma(d/2)}{\sqrt{\pi}\,\Gamma((d-1)/2)}(1-t^2)^{(d-3)/2}, \quad -1 < t < 1.$$

Therefore, using the integral representation of I_ν Abramowitz & Stegun (1965),

$$\begin{aligned} A_\alpha(v) &= \frac{\Gamma(d/2)}{\sqrt{\pi}\,\Gamma((d-1)/2)} \int_{-1}^1 e^{\alpha t}(1-t^2)^{\nu-1/2}\, dt \\ &= \Gamma(d/2)\left(\frac{2}{\alpha}\right)^\nu I_\nu(\alpha). \end{aligned}$$

Substituting $\alpha = 1/\tau$ gives

$$\mathcal{L}^{\mathrm{rep}}_{\mathrm{CLIP}}(\sigma_d, \sigma_d) = \log\left[\Gamma\left(\frac{d}{2}\right)(2\tau)^{(d-2)/2} I_{(d-2)/2}\left(\frac{1}{\tau}\right)\right] = \varepsilon_d.$$

□

B.2.1 A NON-ASYMPTOTIC BOUND ON ε_d

The closed form of Proposition 3 is exact but hard to read. The following elementary bound identifies a regime in which the optimality-error certificate is small: the inverse temperature $\alpha = 1/\tau$ is small compared with $\sqrt{d}$. A small certificate guarantees that the uniform distribution on the sphere is close to optimal in loss value. This is a sufficient, but not necessary, condition for near-optimality.

Proposition 4 (Non-asymptotic bounds for ε_d)**.** *For every $d \geq 2$ and every $\tau > 0$,*

$$\log\left(1 + \frac{1}{2d\tau^2}\right) \;\leq\; \varepsilon_d \;\leq\; \frac{1}{2d\tau^2}.$$

In particular $\varepsilon_d > 0$, and $\varepsilon_d \to 0$ as $d \to \infty$ at fixed τ.

Proof. By the proof of Proposition 3, $\varepsilon_d = \log \Phi(1/\tau)$, and by Lemma 1,

$$\Phi(t) = \sum_{k\geq 0} c_k\, t^{2k}, \qquad c_0 = 1, \quad c_k = \frac{1}{2^k\, k!\, d(d+2)\cdots(d+2k-2)} \quad (k \geq 1).$$

Every factor of the product in the denominator is at least d, so $d(d+2)\cdots(d+2k-2) \geq d^k$ and hence $c_k \leq 1/(2^k k!\, d^k)$ for all $k \geq 1$. Comparing termwise with the exponential series,

$$\Phi(t) \;\leq\; \sum_{k\geq 0} \frac{1}{k!}\left(\frac{t^2}{2d}\right)^k = \exp\left(\frac{t^2}{2d}\right),$$

which gives $\varepsilon_d = \log \Phi(1/\tau) \leq 1/(2d\tau^2)$. For the lower bound, all the coefficients c_k are positive, so truncating the series after $k = 1$ gives $\Phi(t) \geq 1 + c_1 t^2 = 1 + t^2/(2d)$, and taking logarithms at $t = 1/\tau$ yields the claim. Strict positivity follows since $\log(1 + x) > 0$ for $x > 0$. □

B.3 A GLOBAL LOWER BOUND ON THE CLIP LOSS

We now bound the full loss from below. Both separate minima computed in §B.1 equal $-1/\tau$, so one might expect $-2/\tau$ for the sum, but the following proposition shows that the correct bound is $-1/\tau$, and the whole cost of the alignment–repelling tradeoff is visible in that discrepancy.

Proposition 5. *For any $\mu_I, \mu_T \in \mathcal{P}(\mathbb{S}^{d-1})$ and $\pi \in \Gamma(\mu_I, \mu_T)$, we have*

$$\mathcal{L}_{\mathit{CLIP}}(\mu_I, \mu_T, \pi) \geq -\frac{1}{\tau}.$$

The key computation identifies the lower bound with the variance of centered embeddings, which is at most 1 since embeddings lie on the unit sphere.

Proof of Proposition 5. Write $\alpha = 1/\tau$, and let $m_I = \int x\, d\mu_I(x)$ and $m_T = \int y\, d\mu_T(y)$. Applying Jensen's inequality to each direction of the repelling term gives

$$\mathcal{L}^{\mathrm{rep}}_{\mathrm{CLIP}}(\mu_I, \mu_T) \geq \alpha \langle m_I, m_T \rangle.$$

Consequently,

$$\begin{aligned}\mathcal{L}_{\mathrm{CLIP}}(\mu_I, \mu_T, \pi) &\geq -\alpha\left(\int \langle x, y\rangle\, d\pi(x,y) - \langle m_I, m_T\rangle\right) \\ &= -\alpha \int \langle x - m_I, y - m_T\rangle\, d\pi(x,y).\end{aligned}$$

Cauchy–Schwarz in $L^2(\pi; \mathbb{R}^d)$ now yields

$$\int \langle x - m_I, y - m_T\rangle\, d\pi \leq \left(\int \|x - m_I\|^2\, d\mu_I\right)^{1/2} \left(\int \|y - m_T\|^2\, d\mu_T\right)^{1/2}.$$

Since both marginals are supported on the unit sphere, the two integrals equal $1 - \|m_I\|^2$ and $1 - \|m_T\|^2$, respectively. Thus, more precisely,

$$\mathcal{L}_{\mathrm{CLIP}}(\mu_I, \mu_T, \pi) \geq -\frac{1}{\tau}\sqrt{(1 - \|m_I\|^2)(1 - \|m_T\|^2)} \geq -\frac{1}{\tau}.$$

□

B.4 ε_d-MINIMALITY OF THE UNIFORM DISTRIBUTION ON THE SPHERE

Theorem 3 now follows by putting the previous two subsections together: the uniform diagonal triplet attains the alignment optimum exactly and pays ε_d on the repelling term, so it sits at most ε_d above the global lower bound $-1/\tau$ of Proposition 5.

Proof of Theorem 3. By Theorem 1, under the coupling $\pi = (\mathrm{id}_d, \mathrm{id}_d)_{\#}\sigma_d$ we have $\hat{z}_I = \hat{z}_T$ almost surely, so

$$\mathcal{L}^{\text{align}}_{\text{CLIP}}(\sigma_d, \sigma_d, \pi) = -\frac{1}{\tau}\,\mathbb{E}_\pi[\langle \hat{z}_I, \hat{z}_T\rangle] = -\frac{1}{\tau}\,\mathbb{E}[\|\hat{z}\|^2] = -\frac{1}{\tau},$$

using $\|\hat{z}\|^2 = 1$ on $\mathbb{S}^{d-1}$.

As shown in Proposition 3, the repelling loss evaluated at $\mu_I = \mu_T = \sigma_d$ is

$$\varepsilon_d = \log\left((2\tau)^{(d-2)/2}\, I_{(d-2)/2}\left(\frac{1}{\tau}\right)\,\Gamma\left(\frac{d}{2}\right)\right),$$

where I_ν is the modified Bessel function of the first kind.

Combining both terms:

$$\mathcal{L}_{\text{CLIP}}(\sigma_d, \sigma_d, \pi) = -\frac{1}{\tau} + \varepsilon_d.$$

By the lower bound established in Proposition 5, every admissible triplet satisfies $\mathcal{L}_{\text{CLIP}} \geq -1/\tau$. Therefore the triplet $(\sigma_d, \sigma_d, \pi)$ achieves this lower bound up to the additive gap ε_d.

As shown in Proposition 3, $\varepsilon_d \to 0^+$ as $d \to \infty$, so as $d \to \infty$ the triplet $(\sigma_d, \sigma_d, \pi)$ achieves the global lower bound $-1/\tau$ and is therefore asymptotically optimal. □

B.5 NECESSARY CONDITIONS FOR IMPROVING THE UNIFORM DISTRIBUTION ON THE SPHERE FOR CLIP LOSS

B.5.1 PROOF OF PROPOSITION 1

The paired gap and the centroid gap of Liang et al. (2022) differ by exactly the dispersion of the difference vector around its mean, which is the usual bias–variance decomposition applied to $\hat{z}_I - \hat{z}_T$.

Proof. Write $W \coloneqq \hat{z}_I - \hat{z}_T$, so that $\Delta_{\text{gap}} = \mathbb{E}_\pi[\|W\|^2]$. Since the marginals of π are μ_I and μ_T, we have $\mathbb{E}_\pi[W] = \hat{m}_I - \hat{m}_T$, and therefore

$$\mathbb{E}_\pi\big[\|W\|^2\big] = \big\|\mathbb{E}_\pi[W]\big\|^2 + \mathbb{E}_\pi\big[\|W - \mathbb{E}_\pi[W]\|^2\big] = \Delta_{\text{centroid}} + \sum_{k=1}^{d} \mathrm{Var}_\pi(W_k),$$

where the first equality is the coordinatewise identity $\mathbb{E}[W_k^2] = (\mathbb{E}[W_k])^2 + \mathrm{Var}(W_k)$ summed over k, and the cross terms vanish by construction. Since $\sum_k \mathrm{Var}_\pi(W_k) = \mathrm{tr}\big(\mathrm{Cov}_\pi(W)\big)$ and a covariance matrix is positive semidefinite, its trace is non-negative, giving $\Delta_{\text{gap}} \geq \Delta_{\text{centroid}}$. Equality holds if and only if $\mathrm{Cov}_\pi(W) = 0$, that is, if and only if $\hat{z}_I - \hat{z}_T$ is π-a.s. constant. □

B.5.2 PROOF OF THEOREM 4

This theorem gives necessary conditions for strictly improving the uniform distribution on the sphere with diagonal coupling; it does not establish that such an improvement exists. The key identity is again (2): the alignment term is exactly the modality gap, up to the constant shift $-1/\tau$. Once it is substituted into the CLIP objective, the upper bound on Δ_{gap} follows from the global lower bound on the repelling term, while the strict positivity of the gap follows by ruling out the perfectly aligned case.

Proof. For any coupling π', identity (2) gives

$$\mathcal{L}^{\text{align}}_{\text{CLIP}}(\mu'_I, \mu'_T, \pi') = \frac{1}{2\tau}\,\mathbb{E}_{\pi'}\big[\|\hat{z}_I - \hat{z}_T\|^2\big] - \frac{1}{\tau} = \frac{\Delta_{\text{gap}}}{2\tau} - \frac{1}{\tau},$$

hence

$$\mathcal{L}_{\mathrm{CLIP}}(\mu'_I, \mu'_T, \pi') = \frac{\Delta_{\mathrm{gap}}}{2\tau} - \frac{1}{\tau} + \mathcal{L}^{\mathrm{rep}}_{\mathrm{CLIP}}(\mu'_I, \mu'_T). \tag{7}$$

For item i), argue by contradiction. Assume that $\langle \hat{z}_I, \hat{z}_T \rangle = 1$ almost surely. Since both embeddings lie on $\mathbb{S}^{d-1}$, this is equivalent to $\|\hat{z}_I - \hat{z}_T\|^2 = 0$ almost surely, so $\Delta_{\mathrm{gap}} = 0$ and the coupling is supported on the diagonal. In particular, $\mu'_I = \mu'_T =: \mu$, and the CLIP loss reduces to

$$\mathcal{L}_{\mathrm{CLIP}}(\mu, \mu, (\mathrm{id}_d, \mathrm{id}_d)_{\#}\mu) = -\frac{1}{\tau} + \mathcal{L}^{\mathrm{rep}}_{\mathrm{CLIP}}(\mu, \mu).$$

Under this diagonal restriction, the repelling term reduces to $\mathbb{E}_{\hat{z}\sim\mu}\big[\log \mathbb{E}_{\hat{z}'\sim\mu} \exp(\langle \hat{z}, \hat{z}' \rangle/\tau)\big]$, which is exactly the second term of the asymptotic contrastive loss in Theorem 1 of Wang & Isola (2020). We note that the outer expectation sits *outside* the logarithm, so this is that asymptotic term and not their uniformity functional $\mathcal{L}_{\mathrm{uniform}} = \log \mathbb{E}\mathbb{E}[\cdot]$; the two coincide precisely when the inner expectation is constant, which is the case at a rotation-invariant μ. Their theorem states that whenever perfectly uniform encoders are realizable they are the *exact* minimizers of that term, and here the minimization ranges over all of $\mathcal{P}(\mathbb{S}^{d-1})$, so σ_d is realizable by construction and the clause applies unconditionally. Therefore

$$\mathcal{L}^{\mathrm{rep}}_{\mathrm{CLIP}}(\mu, \mu) \geq \mathcal{L}^{\mathrm{rep}}_{\mathrm{CLIP}}(\sigma_d, \sigma_d) = \varepsilon_d.$$

Consequently,

$$\mathcal{L}_{\mathrm{CLIP}}(\mu'_I, \mu'_T, \pi') \geq -\frac{1}{\tau} + \varepsilon_d,$$

which contradicts the strict hypothesis of the theorem. Hence perfect alignment cannot hold.

For item ii), Theorem 2 gives $\mathcal{L}^{\mathrm{rep}}_{\mathrm{CLIP}}(\mu'_I, \mu'_T) \geq -1/\tau$. Combining this with (7) yields

$$\mathcal{L}_{\mathrm{CLIP}}(\mu'_I, \mu'_T, \pi') \geq \frac{\Delta_{\mathrm{gap}}}{2\tau} - \frac{2}{\tau}.$$

Using the assumption $\mathcal{L}_{\mathrm{CLIP}}(\mu'_I, \mu'_T, \pi') < -1/\tau + \varepsilon_d$, we obtain

$$\frac{\Delta_{\mathrm{gap}}}{2\tau} - \frac{2}{\tau} < -\frac{1}{\tau} + \varepsilon_d,$$

which rearranges to

$$\Delta_{\mathrm{gap}} < 2 + 2\tau\varepsilon_d.$$

Moreover, by item 1 we do not have $\hat{z}_I = \hat{z}_T$ almost surely, so $\|\hat{z}_I - \hat{z}_T\|^2$ is not almost surely zero. Since it is nonnegative, its expectation is strictly positive, and therefore

$$0 < \Delta_{\mathrm{gap}} < 2 + 2\tau\varepsilon_d.$$

This upper bound is always strictly smaller than the universal bound of four. Indeed, for independent $\omega, \eta \sim \sigma_d$ and finite $d \geq 2$, $\omega \cdot \eta < 1$ almost surely, so

$$\Phi(1/\tau) = \mathbb{E}e^{\omega\cdot\eta/\tau} < e^{1/\tau}, \qquad \tau\varepsilon_d = \tau \log \Phi(1/\tau) < 1$$

for every $\tau > 0$.

For item iii), we have from (7) for the repelling term,

$$\mathcal{L}^{\mathrm{rep}}_{\mathrm{CLIP}}(\mu'_I, \mu'_T) = \mathcal{L}_{\mathrm{CLIP}}(\mu'_I, \mu'_T, \pi') + \frac{1}{\tau} - \frac{\Delta_{\mathrm{gap}}}{2\tau},$$

and insert the hypothesis $\mathcal{L}_{\mathrm{CLIP}}(\mu'_I, \mu'_T, \pi') < -1/\tau + \varepsilon_d$ to obtain

$$\mathcal{L}^{\mathrm{rep}}_{\mathrm{CLIP}}(\mu'_I, \mu'_T) < \varepsilon_d - \frac{\Delta_{\mathrm{gap}}}{2\tau}.$$

This proves the claim. By (7), this inequality is equivalent to the strict-improvement hypothesis. It specifies the reduction in the repelling term required to improve the reference at a given paired gap, without establishing that this reduction is realizable. □

B.6 THE MAXIMAL-CORRELATION BOTTLENECK

Theorem 4 gives necessary conditions for improving a reference value of CLIP. This subsection proves loss independent bounds on alignment: the general version retains explicit mean terms, and the centered version gives the lower bound on the paired gap. Neither bound identifies separate data and loss contributions to the observed gap or guarantees attainability.

The idea is that each coordinate of the normalized embeddings is a square-integrable scalar transformation of its own modality, and hence an *admissible* competitor in the variational problem of Definition 1. The maximal correlation therefore controls every coordinatewise covariance simultaneously, and summing over the d coordinates controls the expected inner product itself. Since we have $\Delta_{\mathrm{gap}} = 2 - 2\,\mathbb{E}_\pi[\langle \hat{z}_I, \hat{z}_T \rangle]$, the bound propagates directly to the gap.

Throughout this subsection we work under Assumption 1, with the notation and conventions of Appendix A.4. We first record the elementary fact that ρ_{HGR} bounds the covariance of *any* pair of square-integrable scalar statistics of the two modalities, not only of standardized ones.

Lemma 3. *Let $\phi \in L^2(\mu_X^I)$ and $\psi \in L^2(\mu_X^T)$ be real valued. Then*

$$\mathrm{Cov}\big(\phi(X_I),\, \psi(X_T)\big) \;\le\; \rho_{\mathrm{HGR}}(X_I, X_T)\, \sqrt{\mathrm{Var}\big(\phi(X_I)\big)\; \mathrm{Var}\big(\psi(X_T)\big)}.$$

Proof. If $\mathrm{Var}(\phi(X_I)) = 0$ or $\mathrm{Var}(\psi(X_T)) = 0$, the corresponding random variable is almost surely constant, so both sides vanish and the inequality is trivial. Assume then that both variances are strictly positive and set

$$\tilde{\phi} := \frac{\phi(X_I) - \mathbb{E}[\phi(X_I)]}{\sqrt{\mathrm{Var}(\phi(X_I))}}, \qquad \tilde{\psi} := \frac{\psi(X_T) - \mathbb{E}[\psi(X_T)]}{\sqrt{\mathrm{Var}(\psi(X_T))}}.$$

By construction $\tilde{\phi}$ is a square-integrable function of X_I alone and $\tilde{\psi}$ of X_T alone, with $\mathbb{E}[\tilde{\phi}] = \mathbb{E}[\tilde{\psi}] = 0$ and $\mathrm{Var}(\tilde{\phi}) = \mathrm{Var}(\tilde{\psi}) = 1$. Hence $(\tilde{\phi}, \tilde{\psi})$ is admissible in the supremum defining $\rho_{\mathrm{HGR}}(X_I, X_T)$, and therefore $\mathbb{E}[\tilde{\phi}\,\tilde{\psi}] \le \rho_{\mathrm{HGR}}(X_I, X_T)$. Since

$$\mathbb{E}[\tilde{\phi}\,\tilde{\psi}] = \frac{\mathrm{Cov}(\phi(X_I), \psi(X_T))}{\sqrt{\mathrm{Var}(\phi(X_I))\; \mathrm{Var}(\psi(X_T))}},$$

multiplying by $\sqrt{\mathrm{Var}(\phi(X_I))\,\mathrm{Var}(\psi(X_T))}$ gives the claim. □

The next proposition is the general form of the bound; Proposition 2 corresponds to the centered case $\hat{m}_I = \hat{m}_T = 0$. The proof technique is similar to the one used by Betser et al. (2026) in their Proposition 1 for a single shared marginal. The extension here is to two encoder-induced marginals $\mu_I \neq \mu_T$ with possibly distinct means $\hat{m}_I \neq \hat{m}_T$, which replaces the single squared-norm term with the cross term $\langle \hat{m}_I, \hat{m}_T \rangle$ and an averaged pair of variance budgets.

Proposition 6 (HGR obstruction to cross-modal alignment)**.** *Under Assumption 1, for the coupling π induced by the encoders,*

$$\mathbb{E}_\pi\big[\langle \hat{z}_I, \hat{z}_T \rangle\big] \;\le\; \langle \hat{m}_I, \hat{m}_T \rangle + \rho_{\mathrm{HGR}}(X_I, X_T)\, \frac{\big(1 - \|\hat{m}_I\|^2\big) + \big(1 - \|\hat{m}_T\|^2\big)}{2}.$$

In particular, if the embedding distributions are centered, $\hat{m}_I = \hat{m}_T = 0$, then

$$\mathbb{E}_\pi\big[\langle \hat{z}_I, \hat{z}_T \rangle\big] \;\le\; \rho_{\mathrm{HGR}}(X_I, X_T).$$

Proof. Write $U = \hat{z}_I = F_I(X_I)$ and $V = \hat{z}_T = F_T(X_T)$. Since $\|U\| = \|V\| = 1$ almost surely, every coordinate satisfies $|U_k| \le 1$ and $|V_k| \le 1$, so $F_{I,k} \in L^2(\mu_X^I)$ and $F_{T,k} \in L^2(\mu_X^T)$ for all $k = 1, \dots, d$, and all the quantities below are finite. Decomposing the inner product into centered and centroid parts,

$$\mathbb{E}_\pi\big[\langle U, V \rangle\big] = \sum_{k=1}^{d} \mathbb{E}_\pi[U_k V_k] = \langle \hat{m}_I, \hat{m}_T \rangle + \sum_{k=1}^{d} \mathrm{Cov}(U_k, V_k), \tag{8}$$

where we used that $\mathbb{E}[U_k] = \hat{m}_{I,k}$ and $\mathbb{E}[V_k] = \hat{m}_{T,k}$ are determined by the marginals μ_I, μ_T.

Applying Lemma 3 with $\phi = F_{I,k}$ and $\psi = F_{T,k}$, and then the arithmetic–geometric mean inequality $\sqrt{ab} \le (a+b)/2$, gives for every k

$$\mathrm{Cov}(U_k, V_k) \le \rho_{\mathrm{HGR}}(X_I, X_T)\sqrt{\mathrm{Var}(U_k)\,\mathrm{Var}(V_k)} \le \rho_{\mathrm{HGR}}(X_I, X_T)\,\frac{\mathrm{Var}(U_k) + \mathrm{Var}(V_k)}{2}.$$

Summing over k and using that, on the unit sphere,

$$\sum_{k=1}^{d} \mathrm{Var}(U_k) = \mathbb{E}\big[\|U - \hat{m}_I\|^2\big] = \mathbb{E}\big[\|U\|^2\big] - \|\hat{m}_I\|^2 = 1 - \|\hat{m}_I\|^2,$$

and likewise $\sum_{k=1}^{d} \mathrm{Var}(V_k) = 1 - \|\hat{m}_T\|^2$, we obtain

$$\sum_{k=1}^{d} \mathrm{Cov}(U_k, V_k) \le \rho_{\mathrm{HGR}}(X_I, X_T)\,\frac{\big(1 - \|\hat{m}_I\|^2\big) + \big(1 - \|\hat{m}_T\|^2\big)}{2}.$$

Substituting into (8) proves the first claim. □

Proof of Proposition 2. By (1),

$$\Delta_{\mathrm{gap}} = \mathbb{E}_\pi\big[\|\hat{z}_I - \hat{z}_T\|^2\big] = 2 - 2\,\mathbb{E}_\pi\big[\langle \hat{z}_I, \hat{z}_T\rangle\big].$$

By the centered case of Proposition 6, $\mathbb{E}_\pi[\langle \hat{z}_I, \hat{z}_T\rangle] \le \rho_{\mathrm{HGR}}(X_I, X_T)$, hence

$$\Delta_{\mathrm{gap}} \;\ge\; 2\big(1 - \rho_{\mathrm{HGR}}(X_I, X_T)\big).$$

For the second claim, suppose perfect alignment held, that is, $\hat{z}_I = \hat{z}_T$ π-almost surely. Then $\langle \hat{z}_I, \hat{z}_T\rangle = \|\hat{z}_I\|^2 = 1$ almost surely, so $\mathbb{E}_\pi[\langle \hat{z}_I, \hat{z}_T\rangle] = 1$ and the bound above forces $\rho_{\mathrm{HGR}}(X_I, X_T) \ge 1$. Contrapositively, whenever $\rho_{\mathrm{HGR}}(X_I, X_T) < 1$ perfect alignment is impossible, and the modality gap is bounded away from zero by $\Delta_{\mathrm{gap}} \ge 2(1 - \rho_{\mathrm{HGR}}(X_I, X_T)) > 0$. □

C Theoretical Results for the Prototype Loss

This appendix proves the results of Section 3.4. The prototype loss, the prototype law $\nu_\pi = m_\#\pi$, the Gaussian energy I_α, the relation between the modality gap and the mean-square radius (4), and the angular moment-generating function Φ are defined in Appendix A and used here without further comment.

C.1 Overview and proof strategy

The population problem is to minimize

$$\mathcal{L}_{\mathrm{prot}}(\mu_I, \mu_T, \pi) = \log I_\alpha[\nu_\pi], \qquad I_\alpha[\mu] = \iint e^{-\alpha\|x-y\|^2}\, d\mu(x)\, d\mu(y), \qquad \alpha = \frac{1}{\tau},$$

over triplets (μ_I, μ_T, π). Since $\log$ is strictly increasing, this is the minimization of the Gaussian energy I_α over the set of *attainable* prototype laws

$$\mathcal{A} \coloneqq \big\{\, m_\#\pi \;:\; \pi \in \mathcal{P}(\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}) \,\big\} \;\subseteq\; \mathcal{P}(\overline{\mathbb{B}}^d).$$

The following lemma says that this set is no restriction at all: the midpoint map is onto the measures on the ball. The population problem is therefore *exactly* the minimization of I_α over $\mathcal{P}(\overline{\mathbb{B}}^d)$, and the only work left on the triplet side is to unwind an optimal prototype law back into a pair of marginals and a coupling.

Lemma 4 (Every law on the ball is a prototype law)**.** *For every $d \ge 2$, $\mathcal{A} = \mathcal{P}(\overline{\mathbb{B}}^d)$.*

Proof. Since the midpoint of two unit vectors belongs to $\overline{\mathbb{B}}^d$, only the reverse inclusion needs proof. Define

$$a(p) = \begin{cases} e_2, & p = \lambda e_1,\ \lambda \in \mathbb{R} \\ e_1, & \text{otherwise.} \end{cases}$$

For $p \neq 0$, let

$$v(p) = a(p) - \frac{\langle a(p), p\rangle}{\|p\|^2} p, \qquad w(p) = \sqrt{1 - \|p\|^2}\, \frac{v(p)}{\|v(p)\|},$$

and set $w(0) = e_1$. By construction $v(p) \neq 0$, so w is Borel, $w(p) \perp p$, and $\|w(p)\|^2 = 1 - \|p\|^2$. Consequently $T(p) = (p + w(p), p - w(p))$ belongs to $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$ and satisfies $m(T(p)) = p$. For any $\nu \in \mathcal{P}(\overline{\mathbb{B}}^d)$, the measure $\pi = T_{\#}\nu$ therefore satisfies $m_{\#}\pi = \nu$. □

The argument runs in six steps, each occupying one subsection.

- **(S1)** *Prove uniqueness.* The Gaussian kernel is strictly positive definite, so I_α is strictly convex and weak-$*$ continuous on the compact set $\mathcal{P}(\overline{\mathbb{B}}^d)$. The problem therefore has exactly one minimizer $\mu^\star$ (§C.2).
- **(S2)** *Rotational invariance.* The kernel is rotation invariant, so averaging $\mu^\star$ over $O(d)$ cannot increase the energy. Uniqueness then forces $\mu^\star$ to be $O(d)$-invariant (§C.3).
- **(S3)** *Reduction to a one-dimensional problem.* Every $O(d)$-invariant measure on the ball is a mixture of uniform spheres, $\int_0^1 \sigma_r \, d\rho(r)$, so the problem reduces to minimizing the scalar energy $E[\rho] = \iint g \, d\rho \, d\rho$ of the radial law over $\rho \in \mathcal{P}([0,1])$, with the explicit kernel $g(r,s) = e^{-\alpha(r^2+s^2)}\Phi(2\alpha rs)$ (§C.4).
- **(S4)** *Turn optimality into a potential condition.* Criterion says that δ_R minimizes E as soon as the potential $U_R(r) = g(r, R)$ is minorized appropriately(§C.5).
- **(S5)** *Analyse the potential.* For the unconstrained problem the condition is $U(r) \geq U(1)$ on $[0,1]$. It holds exactly for $\alpha \leq \alpha_c(d)$, where $\alpha_c(d)$ is the unique solution of $\Phi(2\alpha) = e^\alpha$(§C.6).
- **(S6)** *Conclude.* In the case $\alpha \leq \alpha_c(d)$, the minimizer of I_α is σ_d (as a measure in $\mathcal{P}(\overline{\mathbb{B}}^d)$), using this through the midpoint map recovers $\mu_I = \mu_T = \sigma_d$ and the diagonal coupling (§C.7).

Section C.8 then redo steps (S4)–(S6) under the additional linear constraint $\int r^2 d\rho = R^2$, which by (4) is exactly the constraint of a given modality gap; §C.9 derives the Gaussian limit of fixed-dimensional projections at the optimum, with the appropriate scaling; and §C.10 obtains the linear time and memory estimator for the prototype loss based on random features, using the fact that the energy is constructed as a kernel energy.

Two conditions on the inverse temperature appear. The unrestricted case involves the *sharp* constant $\alpha_c(d)$ of Definition 3: the uniform measure is optimal for $\alpha \leq \alpha_c(d)$ and fails to be optimal for $\alpha > \alpha_c(d)$. The fixed-gap result instead relies on a *sufficient* constant $\beta_c(d)$, which is likely not sharp.

C.2 (S1) Strict convexity and a unique minimizer

Strict convexity follows from the positive-definiteness of the Gaussian kernel: for any non-zero finite signed measure ν on $\overline{\mathbb{B}}^d$ (in particular for $\nu = \mu - \mu'$ with μ, μ' probability measures), the quadratic form $I_\alpha[\nu]$ is strictly positive; boundedness of $\overline{\mathbb{B}}^d$ and continuity of k_α make $I_\alpha[\nu] = \iint k_\alpha \, d\nu \, d\nu$ well defined and finite for every such ν.

Proposition 7 (Strict convexity)**.** *The functional*

$$I_\alpha[\mu] = \iint e^{-\alpha\|x-y\|^2} \, d\mu(x) \, d\mu(y)$$

is strictly convex on $\mathcal{P}(\overline{\mathbb{B}}^d)$.

The proof relies on the following lemma, which is also used in the context of energy theory in Borodachov et al. (2019).

Lemma 5 (Borodachov–Hardin–Saff 4.4.2)**.** $I_\alpha[\nu] \geq 0$ *for every signed measure ν, with equality if and only if $\nu \equiv 0$.*

Proof of Proposition 7. Let $\mu, \nu \in \mathcal{P}(\overline{\mathbb{B}}^d)$ be distinct and $\lambda \in (0,1)$. Expanding the bilinear form,

$$I_\alpha[\lambda\mu + (1-\lambda)\nu] = \lambda^2 I_\alpha[\mu] + 2\lambda(1-\lambda)\langle\mu,\nu\rangle + (1-\lambda)^2 I_\alpha[\nu]$$

we want to show that $I_\alpha(\lambda\mu + (1-\lambda)\nu) < \lambda I_\alpha[\mu] + (1-\lambda)I_\alpha[\nu]$, for this compute

$$\begin{aligned}
&\lambda I_\alpha[\mu] + (1-\lambda)I_\alpha[\nu] - I_\alpha(\lambda\mu + (1-\lambda)\nu)\\
=&\lambda(1-\lambda)I_\alpha[\mu] + \lambda(1-\lambda)I_\alpha[\nu] - 2\lambda(1-\lambda)\langle\mu,\nu\rangle\\
=&\lambda(1-\lambda)(I_\alpha[\mu] + I_\alpha[\nu] - 2\langle\mu,\nu\rangle)\\
=&\lambda(1-\lambda)I_\alpha[\mu-\nu]\\
>&0,
\end{aligned}$$

the last inequality holds because $\lambda \in (0,1)$ and $\mu \neq \nu$, so $\mu - \nu$ is a non-null signed measure; strict convexity then follows from the lemma. □

For the existence of a minimizer: the set $\mathcal{P}(\overline{\mathbb{B}}^d)$ is weak-$*$ compact, since $\overline{\mathbb{B}}^d$ is compact, and the kernel $k_\alpha(x,y) = e^{-\alpha\|x-y\|^2}$ is continuous and bounded on $\overline{\mathbb{B}}^d \times \overline{\mathbb{B}}^d$, so $\mu \mapsto I_\alpha[\mu]$ is weak-$*$ continuous and attains its infimum. Combined with the strict convexity of Proposition 7, the minimizer exists and is unique.

C.3 (S2) Rotational invariance of the minimizer

The Gaussian kernel is rotationally invariant: $k_\alpha(Rx, Ry) = k_\alpha(x,y)$ for all $R \in O(d)$. We now prove that the unique minimizer of the Gaussian energy functional is invariant under all rotations in $O(d)$, which is the key step that reduces the minimization over the ball to a one-dimensional problem over the radial distribution.

Proposition 8 (Rotational invariance)**.** *The unique minimizer of I_α in $\mathcal{P}(\overline{\mathbb{B}}^d)$ is invariant under all rotations in $O(d)$.*

Proof. Let $\mu^\star$ be the unique minimizer, and fix $R \in O(d)$. Since R preserves $\overline{\mathbb{B}}^d$ and Euclidean distances,

$$\begin{aligned}
I_\alpha[R_\#\mu^\star] &= \iint k_\alpha(Rx, Ry)\, d\mu^\star(x)\, d\mu^\star(y)\\
&= \iint k_\alpha(x,y)\, d\mu^\star(x)\, d\mu^\star(y) = I_\alpha[\mu^\star].
\end{aligned}$$

Thus $R_\#\mu^\star$ is also a minimizer. By uniqueness, $R_\#\mu^\star = \mu^\star$. Since $R \in O(d)$ was arbitrary, $\mu^\star$ is invariant under $O(d)$. □

C.4 (S3) Radial decomposition and reduction to one dimension

A rotation-invariant probability measure on $\overline{\mathbb{B}}^d$ must be a mixture of uniform surface measures σ_r on spheres $r \cdot \mathbb{S}^{d-1}$, parameterized by their radius $r \in [0,1]$. This reduces the infinite-dimensional optimization over $\mathcal{P}(\overline{\mathbb{B}}^d)$ to a one-dimensional problem over the radial distribution $\rho \in \mathcal{P}([0,1])$.

Proposition 9 (Radial decomposition)**.** *Let $\mu \in \mathcal{P}(\overline{\mathbb{B}}^d)$ be $O(d)$-invariant. Then there exists a unique $\rho \in \mathcal{P}([0,1])$ such that*

$$\mu = \int_0^1 \sigma_r\, d\rho(r),$$

where σ_r is the normalized surface measure on $r \cdot \mathbb{S}^{d-1}$.

Since $O(d)$ is a compact group, it admits a normalized Haar probability measure H, invariant under left and right multiplication Folland (2015). For any $x \in \overline{\mathbb{B}}^d$, the pushforward of H under the map $R \mapsto Rx$ is the uniform probability measure $\sigma_{\|x\|}$ on the sphere of radius $\|x\|$, with the convention $\sigma_0 = \delta_0$, the Dirac delta on the origin. Uniqueness of the invariant probability measure on a compact homogeneous space (Folland, 2015, Theorem 2.51) therefore identifies it with $\sigma_{\|x\|}$. The same identity holds at $x = 0$, where $\sigma_0 = \delta_0$.

Proof of Proposition 9. Let H be the normalized Haar measure on $O(d)$, and define $\rho = (x \mapsto \|x\|)_{\#}\mu$. For each $R \in O(d)$, the invariance $R_{\#}\mu = \mu$ gives

$$\int f(Rx)\, d\mu(x) = \int f(y)\, d(R_{\#}\mu)(y) = \int f(y)\, d\mu(y).$$

Averaging this identity with respect to H and applying Fubini's theorem, we obtain

$$\int f\, d\mu = \int_{O(d)} \int f(Rx)\, d\mu(x)\, dH(R) = \int \left(\int_{O(d)} f(Rx)\, dH(R) \right) d\mu(x).$$

The pushforward of H under $R \mapsto Rx$ is $\sigma_{\|x\|}$ (Meckes, 2019, Section 1.2). Therefore,

$$\int f\, d\mu = \int \left(\int f\, d\sigma_{\|x\|} \right) d\mu(x) = \int_0^1 \left(\int f\, d\sigma_r \right) d\rho(r).$$

This proves

$$\mu = \int_0^1 \sigma_r\, d\rho(r).$$

Applying the radial projection $x \mapsto \|x\|$ to this identity shows that ρ is uniquely determined by μ. □

The one-dimensional energy. With the radial decomposition in hand, we write the energy of a rotation-invariant measure $\mu = \int_0^1 \sigma_r\, d\rho(r)$ as a scalar functional of ρ:

$$I_\alpha[\mu] = \int_0^1 \int_0^1 g(r,s)\, d\rho(r)\, d\rho(s) =: E[\rho].$$

Let $\omega, \eta \in \mathbb{S}^{d-1}$ and $r, s \in [0,1]$. Using

$$\|r\omega - s\eta\|^2 = r^2 + s^2 - 2rs\, \omega \cdot \eta$$

and the definition of Φ in Lemma 1,

$$\begin{aligned} g(r,s) &= \int_{\mathbb{S}^{d-1}} \int_{\mathbb{S}^{d-1}} e^{-\alpha\|r\omega - s\eta\|^2} d\sigma_d(\omega)\, d\sigma_d(\eta) \\ &= e^{-\alpha(r^2+s^2)} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{S}^{d-1}} e^{2\alpha rs\, \omega\cdot\eta} d\sigma_d(\omega) d\sigma_d(\eta) \\ &= e^{-\alpha(r^2+s^2)} \Phi(2\alpha rs). \end{aligned} \tag{9}$$

If $\rho = \delta_1$, then $\mu = \sigma_1 = \sigma_d$, the uniform surface measure on $\mathbb{S}^{d-1}$. The unrestricted case part of Theorem 5 is therefore equivalent to showing that δ_1 minimizes E over $\mathcal{P}([0,1])$.

C.5 (S4) Criterion for the optimal radial profile

The following lemma reduces the global minimization of E to a pointwise condition on the potential $U(r) = g(r,1)$: if the potential of the candidate minimizer δ_1 is everywhere at least as large as its value at the support point $r = 1$, then δ_1 is optimal.

Lemma 6 (Criterion)**.** *If $U(r) = g(r,1) \geq U(1)$ for every $r \in [0,1]$, then δ_1 is the* unique *minimizer of E on $\mathcal{P}([0,1])$.*

Proof. For $\rho \in \mathcal{P}([0,1])$, write $\nu = \rho - \delta_1$. Then

$$\begin{aligned} E[\rho] &= E[\nu + \delta_1] \\ &= E[\nu] + 2\int_0^1 \int_0^1 g(r,s) d\nu(r) d\delta_1(s) + E[\delta_1] \\ &= E[\delta_1] + 2\int_0^1 U(r)\, d\nu(r) + \int_0^1 \int_0^1 g(r,s)\, d\nu(r)\, d\nu(s) \end{aligned}$$

Since $U(r) \geq U(1)$,

$$\begin{aligned}&\int_0^1 U(r)d\nu(r)\\ =&\int_0^1 U(r)d(\rho-\delta_1)(r)\\ =&\int_0^1 U(r)d\rho(r) - U(1)\\ =&\int_0^1 (U(r)-U(1))d\rho(r) \geq 0\end{aligned}$$

Call

$$\tilde{\nu} = \int_0^1 \sigma_r d\nu(r),$$

then

$$\iint g(r,s)\, d\nu(r)\, d\nu(s) = \iint k_\alpha(x,y)\, d\widetilde{\nu}(x)\, d\widetilde{\nu}(y) \;\geq\; 0$$

by Lemma 5, with equality if and only if $\widetilde{\nu} = 0$. Moreover $\widetilde{\nu} = 0$ forces $\nu = 0$: each σ_r is supported on $r \cdot \mathbb{S}^{d-1}$, so the radial projection $x \mapsto \|x\|$ pushes $\widetilde{\nu}$ forward to ν, and the map $\nu \mapsto \widetilde{\nu}$ is therefore injective. Hence the quadratic term vanishes exactly when $\rho = \delta_1$.

Combining them,

$$E[\rho] \;=\; E[\delta_1] + 2\int_0^1 \big(U(r) - U(1)\big)\, d\rho(r) + \iint g\, d\nu\, d\nu \;\geq\; E[\delta_1],$$

and for $\rho \neq \delta_1$ the last term is strictly positive while the middle one is non-negative, so the inequality is strict. Thus δ_1 is the unique minimizer, under the hypothesis $U(r) \geq U(1)$. □

C.6 (S5) The radial potential and the sharp critical constant

By Lemma 6, δ_1 minimises E when $U(r) \geq U(1)$ for every $r \in [0,1]$, where

$$U(r) = e^{-\alpha(r^2+1)}\Phi(2\alpha r).$$

The critical constant then follows by comparing $U(0)$ and $U(1)$.

Write $R = R_{d/2-1}$. By (5),

$$\frac{d}{dr}\log U(r) = 2\alpha\big(R(2\alpha r) - r\big). \tag{10}$$

We use the facts that R is strictly increasing and strictly concave, $R(0) = 0$, $R'(0) = 1/d$, $R(t) < 1$, and $R(t) \to 1$ as $t \to \infty$; see (R1)–(R3) in §A.5.2.

Lemma 7 (Monotonicity for $\alpha \leq d/2$)**.** *If $0 < \alpha \leq d/2$, then U is strictly decreasing on $[0,\infty)$. In particular, $U(r) > U(1)$ for $0 \leq r < 1$.*

Proof. Strict concavity gives $R(t) < t/d$ for every $t > 0$. Thus, for $r > 0$,

$$R(2\alpha r) - r < \left(\frac{2\alpha}{d} - 1\right) r \leq 0.$$

The claim follows from (10). □

For $\alpha > d/2$, the potential is no longer decreasing throughout $[0,1]$. But we can study the regime for α comparing:

$$U(0) = e^{-\alpha}, \qquad U(1) = e^{-2\alpha}\Phi(2\alpha).$$

In particular, if

$$h(\alpha) := \log\Phi(2\alpha) - \alpha,$$

then $h(\alpha) = \log\big(U(1)/U(0)\big)$.

Definition 3 (Critical inverse-temperature constant)**.** *For $d \geq 2$, let $\alpha_c(d)$ be the unique positive solution of*

$$\Phi(2\alpha) = e^{\alpha}, \qquad \textit{equivalently,} \qquad \Gamma\left(\tfrac{d}{2}\right) \alpha^{1-d/2} I_{d/2-1}(2\alpha) = e^{\alpha}. \tag{11}$$

Lemma 8 (Existence and location of the critical constant)**.** *The constant $\alpha_c(d)$ is well defined and satisfies $\alpha_c(d) > d/2$. Moreover, $h(\alpha) < 0$ for $0 < \alpha < \alpha_c(d)$ and $h(\alpha) > 0$ for $\alpha > \alpha_c(d)$.*

Proof. We have $h(0) = 0$ and

$$h'(\alpha) = 2R(2\alpha) - 1.$$

Since R increases strictly from 0 to 1, h' increases strictly from -1 to 1. Hence h first decreases and then increases to $+\infty$, so it has exactly one positive zero. Lemma 7 gives $U(0) > U(1)$ at $\alpha = d/2$, or $h(d/2) < 0$. Thus the zero lies above $d/2$, and the stated signs follow. □

Lemma 9 (Sharp regime)**.** *If $0 < \alpha \leq \alpha_c(d)$, then $U(r) \geq U(1)$ for every $r \in [0, 1]$. The inequality is strict for $r < 1$ when $\alpha < \alpha_c(d)$. If $\alpha > \alpha_c(d)$, then $U(0) < U(1)$ and δ_1 does not minimise E. Consequently, $\alpha_c(d)$ is the sharp threshold.*

Proof. For $\alpha \leq d/2$, the result follows from Lemma 7. Suppose $\alpha > d/2$, and set $f(r) = R(2\alpha r) - r$. This function is strictly concave, with

$$f(0) = 0, \qquad f'(0) = \frac{2\alpha}{d} - 1 > 0, \qquad f(1) = R(2\alpha) - 1 < 0.$$

It follows that f changes sign exactly once in $(0, 1)$, from positive to negative. By (10), U first increases and then decreases. Therefore its minimum on $[0, 1]$ is attained at 0 or 1, and every interior value is strictly larger than the smaller endpoint value.

By Lemma 8,

$$U(1) \leq U(0) \quad \Longleftrightarrow \quad h(\alpha) \leq 0 \quad \Longleftrightarrow \quad \alpha \leq \alpha_c(d).$$

This proves the claimed inequalities.

Finally, if $\alpha > \alpha_c(d)$, then $U(0) < U(1)$. For $\rho_t = (1 - t)\delta_1 + t\delta_0$, the energy expansion from the proof of Lemma 6 gives

$$E[\rho_t] - E[\delta_1] = 2t\big(U(0) - U(1)\big) + O(t^2) < 0$$

for all sufficiently small $t > 0$. Thus δ_1 is not a minimiser. □

Remark 1 (Numerical values of α_c)**.** *The critical constant $\alpha_c(d)$ can be evaluated by solving* (11) *numerically. Representative values are collected in Table 3.*

d	2	4	8	16	32	64	128	256	512	1024
$\alpha_c(d)$	1.36	2.97	6.25	12.85	26.07	52.51	105.38	211.13	422.64	845.66

Table 3: Numerical values of the sharp critical constant $\alpha_c(d)$.

C.7 (S6) Proof of the unrestricted case in Theorem 5

We now assemble the preceding subsections. The only remaining ingredient is the elementary fact that a midpoint reaches the boundary sphere exactly for a perfectly aligned pair, which is what converts a statement about ν_π back into a statement about the triplet.

Lemma 10 (Midpoint characterisation)**.** *For $x, y \in \mathbb{S}^{d-1}$, $\left\|\frac{x+y}{2}\right\| = 1$ if and only if $x = y$.*

Proof. By (3),

$$\left\|\frac{x+y}{2}\right\|^2 = \frac{1 + \langle x, y\rangle}{2} = \frac{1 + \cos\theta}{2},$$

which equals 1 iff $\cos\theta = 1$ iff $x = y$. □

Proof of the unrestricted case. By Lemma 9, whenever $\alpha = 1/\tau \leq \alpha_c(d)$ the radial potential satisfies $U(r) \geq U(1)$ on $[0,1]$, so the criterion of Lemma 6 gives δ_1 as the unique minimiser of E on $\mathcal{P}([0,1])$. Together with the radial decomposition (Proposition 9) and rotational invariance (Proposition 8), this identifies σ_d as the unique minimiser of I_α over all of $\mathcal{P}(\overline{\mathbb{B}}^d)$.

By Lemma 4 the triplets realise every prototype law on the ball, $\mathcal{A} = \mathcal{P}(\overline{\mathbb{B}}^d)$, so minimising $\mathcal{L}_{\text{prot}}$ over triplets (μ_I, μ_T, π) is exactly minimising I_α over $\mathcal{P}(\overline{\mathbb{B}}^d)$, and uniqueness is inherited. Concretely, the minimiser is realised by the diagonal coupling: $\pi = (\mathrm{id}_d, \mathrm{id}_d)_\# \sigma_d$ gives $m_\# \pi = \sigma_d$, since $m(x,x) = x$. Hence any minimising triplet satisfies

$$\nu_\pi = \sigma_d.$$

Since σ_d is supported on the boundary sphere $\mathbb{S}^{d-1} \subset \overline{\mathbb{B}}^d$, this implies $\nu_\pi(\mathbb{S}^{d-1}) = 1$, so, using Lemma 10 in the last step,

$$\begin{aligned}
1 &= \nu_\pi(\mathbb{S}^{d-1}) \\
&= \pi(m^{-1}(\mathbb{S}^{d-1})) \\
&= \pi\left(\left\{(x,y) \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} : \left\|\frac{x+y}{2}\right\| = 1\right\}\right) \\
&= \pi\left(\left\{(x,y) \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} : x = y\right\}\right)
\end{aligned}$$

With this we prove $\mu_I = \mu_T$. Let A be a Borel set in $\mathbb{S}^{d-1}$, so

$$\begin{aligned}
\mu_I(A) &= \pi(A \times \mathbb{S}^{d-1}) \\
&= \pi\left(\left\{(x,y) \in A \times \mathbb{S}^{d-1} : x = y\right\}\right) \\
&= \pi\left(\left\{(x,y) \in \mathbb{S}^{d-1} \times A : x = y\right\}\right) \\
&= \pi(\mathbb{S}^{d-1} \times A) \\
&= \mu_T(A).
\end{aligned}$$

We conclude $\mu_I = \mu_T$ for all Borel sets of $\mathbb{S}^{d-1}$. Now, we must prove that $\mu_I = \mu_T = \sigma_d$, and this is because

$$\nu_\pi = \sigma_d = m_\# \pi = (\mathrm{id}_d)_\# \mu_I = \mu_I$$

it follows that $\mu_I = \sigma_d = \mu_T$ and the theorem follows. □

C.8 The prototype optimum at a given modality gap

The previous part just proved the unconstrained optimum, which demands perfect alignment. Proposition 2 rules this out for centered encoders whenever $\rho_{\mathrm{HGR}}(X_I, X_T) < 1$, without identifying the minimum attainable gap. It is therefore natural to ask what the objective selects once the achievable level of alignment is fixed. By the identity (4), setting the paired gap Δ_{gap} is exactly setting the second moment of the prototype law,

$$\mathbb{E}_{\nu_\pi}\left[\|p\|^2\right] = 1 - \frac{\Delta_{\mathrm{gap}}}{4} =: R^2, \qquad R \in (0,1],$$

which in the radial coordinates of §C.4 is the linear constraint $\int_0^1 r^2 \, d\rho(r) = R^2$. The argument now repeats steps (S4)–(S6) with this constraint in force: the criterion can be generalized, with the constant $U(1)$ replaced by an affine function of r^2 whose slope is the Lagrange multiplier of the moment constraint.

C.8.1 The constrained criterion

The second-moment constraint suggests comparing the potential U_R with a function of the form $a + br^2$: its integral is the same for every admissible prototype law.

Proposition 10 (Constrained criterion)**.** *Let $R \in (0,1]$ and set $U_R(r) := g(r,R)$. Suppose that, for some $a, b \in \mathbb{R}$,*

$$U_R(r) \geq a + br^2 \quad (0 \leq r \leq 1), \qquad U_R(R) = a + bR^2. \tag{12}$$

Then σ_R is the unique minimizer of I_α among probability measures μ on $\overline{\mathbb{B}}^d$ satisfying $\int \|x\|^2 \, d\mu(x) = R^2$. Consequently, δ_R is the unique minimizer of E among radial laws with second moment R^2.

Equivalently, among triplets with paired modality gap $\Delta_{\mathrm{gap}} = 4(1 - R^2)$, the minimizing prototype law is uniquely $\nu_\pi = \sigma_R$. For $R = 1$, the unconstrained criterion of Lemma 6 is recovered by taking $a = U_1(1)$ and $b = 0$.

Proof. Let $\eta = \mu - \sigma_R$, so that $\mu = \sigma_R + \eta$. Expanding the quadratic energy gives

$$\begin{aligned} I_\alpha[\mu] - I_\alpha[\sigma_R] &= 2 \iint k_\alpha(x, y) \, d\eta(x) \, d\sigma_R(y) + \iint k_\alpha(x, y) \, d\eta(x) \, d\eta(y) \\ &= 2 \int U_R(\|x\|) \, d\eta(x) + \iint k_\alpha(x, y) \, d\eta(x) \, d\eta(y). \end{aligned}$$

The second equality follows because the potential of σ_R at x depends only on $\|x\|$ and equals $U_R(\|x\|)$.

To rewrite the first term, note that both μ and σ_R have mass one and second moment R^2. Hence

$$\int d\eta = 0, \qquad \int \|x\|^2 \, d\eta(x) = 0.$$

We may therefore subtract $a + b\|x\|^2$ inside the integral:

$$\begin{aligned} \int U_R(\|x\|) \, d\eta(x) &= \int \left[U_R(\|x\|) - a - b\|x\|^2\right] d\eta(x) \\ &= \int \left[U_R(\|x\|) - a - b\|x\|^2\right] d\mu(x). \end{aligned}$$

For the last equality, the integral against σ_R is zero: $\|x\| = R$ on its support, and $U_R(R) = a + bR^2$. Combining these identities yields

$$I_\alpha[\mu] - I_\alpha[\sigma_R] = 2 \int \left[U_R(\|x\|) - a - b\|x\|^2\right] d\mu(x) + \iint k_\alpha(x, y) \, d\eta(x) \, d\eta(y).$$

By (12), the first term is non-negative. The second is also non-negative and, by Lemma 5, vanishes only if $\eta = 0$. Thus

$$I_\alpha[\mu] \geq I_\alpha[\sigma_R],$$

with equality only when $\mu = \sigma_R$.

For a radial law ρ, the corresponding measure is $\mu_\rho = \int \sigma_r \, d\rho(r)$, with $I_\alpha[\mu_\rho] = E[\rho]$ and $\int \|x\|^2 \, d\mu_\rho(x) = \int r^2 \, d\rho(r)$. Hence its unique minimizing value is attained at $\rho = \delta_R$. Finally, σ_R is attainable as a prototype law by Lemma 4. □

For $R \in (0, 1)$, equality in (12) also determines a and b. The quadratic lower bound must touch U_R at R.

The constants a and b have a geometric interpretation. In the variable $s = r^2$, the bound $a + br^2$ is a line supporting $F_R(s) \coloneqq U_R(\sqrt{s})$ at $s = R^2$. Since R^2 is an interior point, its slope must equal $F_R'(R^2)$. Thus the two constants are determined by tangency.

Lemma 11 (Tangency of the supporting function)**.** *Let $R \in (0, 1)$. If (12) holds, then a and b are uniquely determined by*

$$b = \frac{U_R'(R)}{2R} = \alpha U_R(R)\big(R_{d/2-1}(2\alpha R^2) - 1\big) < 0, \qquad a = U_R(R) - bR^2.$$

Proof. Set $F_R(s) = U_R(\sqrt{s})$. The function $F_R(s) - (a + bs)$ is non-negative and vanishes at the interior point $s = R^2$. Its derivative must therefore vanish there. By the chain rule,

$$b = F_R'(R^2) = \frac{U_R'(R)}{2R}.$$

The equality $F_R(R^2) = a + bR^2$ then gives $a = U_R(R) - bR^2$.

Finally, differentiating U_R and using (5) gives

$$\frac{U_R'(R)}{U_R(R)} = 2\alpha R\big(R_{d/2-1}(2\alpha R^2) - 1\big).$$

Substitution yields the formula for b. Since $R_{d/2-1}(z) < 1$ for $z > 0$, we have $b < 0$. □

C.8.2 A SUFFICIENT TEMPERATURE CONDITION

To verify (12), it is enough to show that $V(u) \coloneqq U_R(\sqrt{u})$ is convex. Indeed, the tangent line to a convex function lies below its graph, and the tangent at $u = R^2$ is precisely $a + bu$ from Lemma 11.

Lemma 12 (Sufficient temperature condition)**.** *If*

$$\alpha R^2 \le \beta_c(d), \tag{13}$$

then V is convex on $[0, 1]$. Consequently, (12) *holds with the constants from Lemma 11.*

Proof. By Lemma 1, the even power series for Φ becomes a power series in u:

$$V(u) = e^{-\alpha R^2} e^{-\alpha u} F(u), \qquad F(u) = \sum_{k\ge 0} a_k u^k, \qquad a_k = c_k(2\alpha R)^{2k} > 0.$$

Thus $V''(u)$ has the sign of

$$F''(u) - 2\alpha F'(u) + \alpha^2 F(u).$$

We show that every coefficient of this series is non-negative.

Set $q = 2\alpha R^2$ and $D_k = d + 2k$. The coefficient ratio from Lemma 1 gives

$$(k+1)a_{k+1} = \frac{\alpha q}{D_k}\, a_k, \qquad (k+2)(k+1)a_{k+2} = \frac{\alpha^2 q^2}{D_k(D_k+2)}\, a_k.$$

Hence the coefficient of u^k in $F'' - 2\alpha F' + \alpha^2 F$ is

$$\alpha^2 a_k \left(1 - \frac{2q}{D_k} + \frac{q^2}{D_k(D_k+2)}\right).$$

The expression in parentheses is non-negative whenever

$$q \le D_k + 2 - \sqrt{2(D_k+2)}.$$

The right-hand side increases with k, so it suffices to check $k = 0$. This gives

$$q \le d + 2 - \sqrt{2(d+2)} = 2\beta_c(d),$$

which is exactly (13). Therefore $V'' \ge 0$ on $[0, 1]$.

Finally, convexity gives

$$V(u) \ge V(R^2) + V'(R^2)(u - R^2).$$

Taking $u = r^2$, this is (12), with $b = V'(R^2)$ and $a = V(R^2) - bR^2$. □

Remark 2 (β_c versus α_c)**.** *These constants address different questions. The condition in* (13) *guarantees convexity of V, which is sufficient for the supporting-line inequality; it need not be necessary for that inequality. By contrast, $\alpha_c(d)$ is the sharp constant for the unconstrained criterion at $R = 1$.*

Moreover, when $R = 1$, the second-moment constraint itself forces $\rho = \delta_1$, regardless of α. Thus the constrained problem at $R = 1$ does not determine the behavior of the supporting-line condition, or of any sharp fixed-gap threshold, as $R \to 1$.

C.8.3 Monotonicity in the root-mean-square radius, and the HGR bound

The remaining ingredients of Theorem 5 are elementary. The first says that the optimal value decreases as the root of the mean-square radius increases, or equivalently as the modality gap decreases. The second gives a bound for the maximum radius based on HGR maximal correlation.

Lemma 13 (The optimal value decreases in the radius)**.** *The map $R \mapsto g(R,R) = e^{-2\alpha R^2}\Phi(2\alpha R^2)$ is strictly decreasing on $(0,1]$.*

Proof. Put $t = 2\alpha R^2$, so that $\log g(R,R) = -t + \log\Phi(t)$. By (5), $\frac{d}{dt}\big(-t + \log\Phi(t)\big) = R_{d/2-1}(t) - 1 < 0$ for every $t > 0$, by (R1); and t is strictly increasing in R. Composing gives the claim. □

Lemma 14 (HGR bound for the root-mean-square radius)**.** *Suppose the embeddings are given by modality-wise encoders as in Assumption 1, and that the prototype law is centered, $\mathbb{E}_\pi[p] = 0$. Then*

$$\mathbb{E}_\pi\big[\langle \hat{z}_I, \hat{z}_T\rangle\big] \le \rho_{\mathrm{HGR}}(X_I, X_T), \qquad \textit{hence} \qquad R^2 = \mathbb{E}_\pi\big[\|p\|^2\big] \le \frac{1+\rho_{\mathrm{HGR}}(X_I,X_T)}{2}.$$

Proof. Centering of the prototype law means $0 = \mathbb{E}_\pi[p] = (\hat{m}_I + \hat{m}_T)/2$, so $\hat{m}_T = -\hat{m}_I =: -\hat{m}$. Substituting into Proposition 6, whose hypotheses do not require centering,

$$\begin{aligned}
&\mathbb{E}_\pi\big[\langle \hat{z}_I, \hat{z}_T\rangle\big] \\
\le& -\|\hat{m}\|^2 + \rho_{\mathrm{HGR}}\,\frac{(1-\|\hat{m}\|^2) + (1-\|\hat{m}\|^2)}{2} \\
=& \rho_{\mathrm{HGR}} - \|\hat{m}\|^2\big(1+\rho_{\mathrm{HGR}}\big) \\
\le& \rho_{\mathrm{HGR}},
\end{aligned}$$

since $\rho_{\mathrm{HGR}} \ge 0$. The second claim follows from (3) by taking expectations. □

C.9 Proof of Theorem 6

Once the prototype law is characterized as the uniform distribution on a sphere (by Theorem 5), Gaussian convergence of scaled fixed-dimensional projections follows from the Poincaré–Maxwell (Diaconis & Freedman, 1984) argument used in Betser et al. (2026) for the InfoNCE loss in the unimodal case. The sphere has radius $R \le 1$ rather than 1 as soon as the modality gap is positive, and the only change this forces is a division by R: the dilation $x \mapsto x/R$ carries σ_R to σ_d exactly, for every d, so no new limiting argument is needed.

Lemma 15 (Poincaré–Maxwell)**.** *Let $k \in \mathbb{N}$ be a fixed integer and $\mu \in \mathcal{P}(\mathbb{S}^{d-1})$ be the uniform distribution on the unit sphere, and let $U^d \sim \mu$, then if we denote $U^d = (U_1^d, \dots, U_d^d)$ the coordinates of U^d, we have that*

$$\sqrt{d}(U_1^d, \dots, U_k^d) \Rightarrow \mathcal{N}(0, I_k)$$

as $d \to \infty$, where the convergence is in distribution and $\mathcal{N}(0, I_k)$ is the standard Gaussian distribution in $\mathbb{R}^k$.

Proof of Theorem 6. Under $\alpha R^2 \le \beta_c(d)$, part 2 of Theorem 5 gives $\nu_{\pi^*} = \sigma_R$ for the given gap $\Delta = 4(1-R^2)$. The dilation $x \mapsto x/R$ pushes σ_R forward to σ_d, so $U^d := P^d/R$ is uniform on $\mathbb{S}^{d-1}$ for every d, and Lemma 15 applies to it: $\sqrt{d}\,(U_1^d, \dots, U_k^d) \Rightarrow \mathcal{N}(0_k, I_k)$ as $d \to \infty$. Rewriting this in P^d is the first claim.

For the case $R = 1$, also both distributions are the uniform distribution on the sphere, so the claim follows for them too in this case. □

C.10 A computational consequence: linear-time estimation

The structural feature that makes the prototype loss tractable in the analysis above, a single logarithm placed outside one bilinear energy, also has a consequence for how the objective can be estimated. We record it here because it is independent of the rest of the appendix.

Remark 3 (Linear-time estimation of the prototype loss)**.** *Because $\mathcal{L}_{prot} = \log I_\alpha[\nu_\pi]$ places a single logarithm outside one bilinear energy, and $k_\alpha(x, y) = e^{-\alpha\|x-y\|^2}$ is a positive-definite Gaussian kernel, the energy is a squared norm of the kernel mean embedding,*

$$I_\alpha[\nu] = \left\|\mathbb{E}_{p\sim\nu}\varphi(p)\right\|_{\mathcal{H}}^2,$$

where φ is the feature map of k_α. By Bochner's theorem,

$$k_\alpha(x, y) = \mathbb{E}_{\omega\sim\mathcal{N}(0,2\alpha I_d)}\,\mathbb{E}_{b\sim\mathrm{Unif}[0,2\pi]}\left[2\cos(\omega^\top x + b)\cos(\omega^\top y + b)\right]$$

so with M random features $z_j(p) = \sqrt{2/M}\,\cos(\omega_j^\top p + b_j)$, $j = 1, \ldots, M$, we have

$$k_\alpha(x, y) = \mathbb{E}_{\omega,b}[z(x) \cdot z(y)]$$

and

$$\left\|\frac{1}{n}\sum_{i=1}^n z(p_i)\right\|^2 - \frac{1}{n^2}\sum_{i=1}^n \|z(p_i)\|^2 = \frac{1}{n^2}\sum_{i\neq j} z(p_i)\cdot z(p_j),$$

whose expectation over i.i.d. $p_1, \ldots, p_n \sim \nu$ is $\frac{n-1}{n} I_\alpha[\nu]$, not $I_\alpha[\nu]$: the diagonal subtraction removes the n self-pairs, but the remaining $n(n-1)$ off-diagonal terms are still divided by n^2 rather than $n(n-1)$. Rescaling by $n/(n-1)$ gives the unbiased statistic

$$\widehat{I}_\alpha := \frac{1}{n(n-1)}\sum_{i\neq j} z(p_i)\cdot z(p_j) = \frac{n}{n-1}\left[\left\|\frac{1}{n}\sum_{i=1}^n z(p_i)\right\|^2 - \frac{1}{n^2}\sum_{i=1}^n \|z(p_i)\|^2\right],$$

which runs in $O(nMd)$ time, since the feature mean is accumulated in a single streaming pass and the $n \times n$ Gram matrix is never formed, giving linear rather than quadratic dependence on the batch size n. Storing the M frequencies $\omega_j \in \mathbb{R}^d$ costs $O(Md)$ memory in general. Unlike $I_\alpha[\nu] \geq 0$, a given realization of $\widehat{I}_\alpha$ can be negative, since the features z_j are not sign-definite, in which case $\log \widehat{I}_\alpha$ is undefined; using $\widehat{I}_\alpha$ in place of $I_{1/\tau}[\nu_\pi]$ inside $\mathcal{L}_{prot}$ therefore requires a safeguard, such as clipping $\widehat{I}_\alpha$ away from zero, that we do not analyze here.

D Experimental Supplementary Material

Dataset	Loss	Linear	Kernel (RBF)	Zero-shot
CIFAR-10	CLIP	**95.48**	**94.46**	**88.49**
	Prototype	94.77	94.24	87.54
CIFAR-100	CLIP	**81.49**	**75.77**	**65.33**
	Prototype	81.35	75.66	62.44

Table 4: Top-1 accuracy (Linear, Kernel RBF, Zero-shot) on CIFAR-10 and CIFAR-100 using CLIP ViT-B/32 fine-tuned on MS-COCO with either the CLIP or prototype loss.

	Frozen CLIP	CLIP loss	Prototype loss
Image-to-text R@1	50.24	**64.32**	55.64
Image-to-text R@5	74.96	**85.94**	79.74
Image-to-text R@10	83.28	**91.92**	87.88
Text-to-image R@1	30.25	**46.14**	40.84
Text-to-image R@5	54.75	**73.40**	66.70
Text-to-image R@10	66.12	**82.42**	76.00
Mean bidirectional recall	59.93	**74.02**	67.80

Table 5: Retrieval metrics on MS-COCO val2017 of CLIP finetuned on MS-COCO train2017. Mean bidirectional recall is the average of the six columns above it.

	Frozen CLIP	CLIP loss	Prototype loss
Alignment	0.2884	0.2899	**0.4283**
Centroid gap $\|\bar{u} - \bar{v}\|$	0.8198	0.7476	**0.0457**
Paired modality gap $\mathbb{E}\|u - v\|^2$	1.4232	1.4202	**1.1434**
Sphere-uniformity excess	2.0534	1.6122	**0.0329**
Marginal Gaussianity (KS)	0.1796	0.1563	**0.0404**

Table 6: Geometry after fine-tuning of CLIP on 100,000 subset of CC3M for 3,000 steps, metrics were computed on validation set. Higher is better for alignment and lower is better for all the others.

Dataset	Loss	Linear	Kernel (RBF)	k-NN	NCM	Zero-shot
CIFAR-100	CLIP	**80.02**	**73.50**	72.65	68.01	**64.68**
	Prototype	78.25	72.65	**73.84**	**70.69**	63.09
CIFAR-10	CLIP	**94.36**	**93.46**	92.59	89.76	**89.67**
	Prototype	93.35	92.64	**92.86**	**90.74**	86.81

Table 7: Top-1 accuracy (Linear probe, Kernel RBF probe, Zero-shot) on CIFAR-10 and CIFAR-100 using CLIP ViT-B/32 fine-tuned on 100,000 subset of CC3M with either the CLIP or prototype loss.