

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 DECRYPT MODALITY GAP IN MULTIMODAL CONTRASTIVE LEARNING: FROM CONVERGENT REPRESENTATION TO PAIR ALIGNMENT

Anonymous authors

Paper under double-blind review

ABSTRACT

Multimodal contrastive learning (MCL) aims to embed data from different modalities in a shared embedding space. However, empirical evidence shows that representations from different modalities occupy completely separate regions of embedding space, a phenomenon referred to as the modality gap. Moreover, experimental findings on how the size of the modality gap influences downstream performance are inconsistent. These observations raise two key questions: (1) What causes the modality gap? (2) How does it affect downstream tasks? To address these questions, this paper introduces the first theoretical framework for analyzing the convergent optimal representations of MCL and the modality alignment when training is optimized. Specifically, we prove that without any constraint or under the cone constraint, the modality gap converges to zero. Under the subspace constraint (i.e., representations of two modalities fall into two distinct hyperplanes due to dimension collapse), the modality gap converges to the smallest angle between the two hyperplanes. This result identifies *dimension collapse* as the fundamental origin of the modality gap. Furthermore, our theorems demonstrate that paired samples cannot be perfectly aligned under the subspace constraint. The modality gap influences downstream performance by affecting the alignment between sample pairs. We prove that, in this case, perfect alignment between two modalities can still be achieved via two ways: hyperplane rotation and shared space projection.

1 INTRODUCTION

Pre-trained vision–language models (VLMs) (Radford et al., 2021; Mu et al., 2022; Li et al., 2022) have achieved remarkable success across a wide range of tasks, including zero-shot image classification, zero-shot cross-modal retrieval, and visual question answering. These models are typically trained with multimodal contrastive learning on large-scale image–text pairs. Despite their strong empirical performance, our theoretical understanding of how VLMs learn representations and how these representations relate to downstream performance remains limited. In this work, we provide a theoretical study of these issues.

Our understanding of **unimodal** contrastive representation learning (Chen et al., 2020; Khosla et al., 2020) has advanced considerably. From a theoretical standpoint, when training is optimized (i.e., the training loss reaches its minimum), the learned representations converge to an optimal configuration. We refer to this process as *representational convergence* and to its limiting configuration as the *convergent optimal representation* (COR). Prior work has demonstrated that the COR of self-supervised learning (SSL) corresponds to a uniform distribution on the surface of an h -dimensional unit hypersphere (\mathbb{S}^{h-1}) (Wang & Isola, 2020). For supervised contrastive learning (SupCon), the COR forms a regular simplex inscribed in \mathbb{S}^{h-1} (Graf et al., 2021), and a skewed simplex when the data is imbalanced (Yi et al., 2025b). (See additional related work in Sec. A.1.) These prior research on unimodal data demonstrate that examining the geometric and distributional properties of CORs yields critical insights into how pretraining with contrastive learning affects downstream performance.

This motivates us to investigate the COR of **multimodal** contrastive learning (MCL). Intuitively, MCL intends to align representations from different modalities in a shared embedding space. However, this is not supported by empirical evidence. Instead, representations of different modalities cluster

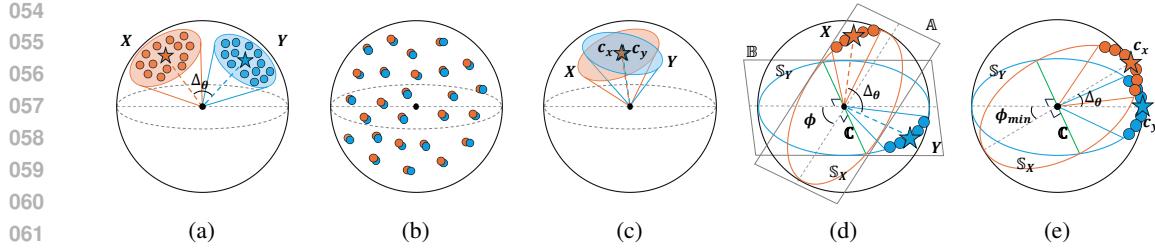


Figure 1: The COR of MCL. Orange and blue dots represent X and Y . Starts are centers of X and Y (i.e., c_x, c_y). Δ_θ denotes the size of modality gap. (a): When a model is initialized, (X, Y) are within two distinct cones. (b): Without any constraint, (X, Y) converge to a paired uniform distribution and $\Delta_\theta \rightarrow 0$. (c): Under the cone constraint, $\Delta_\theta \rightarrow 0$. (d): (X, Y) collapse into two distinct subspaces $\mathbb{S}_X \in \mathbb{A}$ (orange circle) and $\mathbb{S}_Y \in \mathbb{B}$ (blue circle), respectively. ϕ is the angle between \mathbb{A} and \mathbb{B} . Green line represent the shared space \mathbb{C} . See Definition 5 for details. (e): Under the subspace constraint, when training is optimized, $c_x, c_y \perp \mathbb{C}$ and $\Delta_\theta \rightarrow \phi_{\min}$.

into disjoint cones in \mathbb{S}^{h-1} , forming a geometric phenomenon called the *modality gap* (Liang et al., 2022). To explain the origin of this gap, several hypotheses have been proposed, including the cone effect (Liang et al., 2022), the contrastive learning object (Fahim et al., 2024), insufficient training (Shi et al., 2023) and information bias (Schrodi et al., 2025). The impact of the modality gap on downstream performance also remains unclear. Some studies (Liang et al., 2022; Schrodi et al., 2025) show that narrowing the modality gap pos hoc may lead to degraded downstream performance, indicating that such reduction is not always beneficial. (See Sec. A.1 for more details). Prior work have mostly focused on numerical analysis. None of them has offered a satisfactory theoretical explanation of what causes the modality gap and how it affects downstream performance.

In this paper, we in turn focus on the theoretical explanation of the modality gap. We establish the first theoretical framework to systematically analyze the COR of MCL. In particular, we prove (Theorem 1) that, without any distributional constraints, representations of two modalities converge to a paired uniform distribution on \mathbb{S}^{h-1} (Fig. 1b). As a result, the modality gap converges to zero. Meanwhile, the dispersion degree (i.e., how wild a distribution is spread) of the learned representation becomes infinite (Corollary 1). This shows that **the contrastive learning objective tends to close the modality gap**. However, we observe that dispersion degrees of the learned representation always remain finite in practice. Therefore, representations of each modality fall into a cone in \mathbb{S}^{h-1} (Fig. 1a), a phenomenon known as *cone effect*. We prove (Theorem 2) that even under this cone constraint, the modality gap still converges to zero, regardless of the initial locations or sizes of the cones (Fig. 1c). This elucidates that **the cone effect is not the cause of the modality gap**.

The preceding analysis prompts us to ask whether there are any other geometric or distributional constraints on representations that ultimately give rise to the modality gap. Jing et al. (2022) show that the SSL learned representations collapse into a lower-dimensional subspace rather than spanning the entire embedding space, a phenomenon referred to as *dimension collapse*. Inspired by this insight, we observe that dimension collapse also arises in the MCL learned representations. We then prove (Theorem 3) that if representations of two modalities collapse into distinct hyperplanes (Fig. 1d), the modality gap converges to the smallest angle between these hyperplanes (Fig. 1e). This finding demonstrates that **the true origin of the modality gap is dimension collapse**.

That how modality gap influences downstream tasks still confuses researchers. We argue that downstream performance is determined by the alignment between all paired samples, i.e., *modality alignment*. First, we prove (Theorem 4 and Corollary 2) that when representations converge, **the mutual information between two modalities in the shared space is maximized** and in this case **paired samples cannot be perfectly aligned**. Next, we demonstrate that changes in the size of the modality gap alter the representation distribution, which in turn affects modality alignment. Then, we show that existing translation approaches, e.g., shifting image embeddings toward language embeddings by the average distance between image–language pairs, modify the representation distribution in arbitrary ways. This explains the worsen downstream performance observed when such methods are applied. Lastly, we prove derive two methods, hyperplane rotation (Corollary 3) and shared subspace projection (Corollary 4), that achieve perfect alignment and modality gap reduction without harming downstream performance. The major contributions of our work are listed below:

- 108 • We theoretically show that the contrastive learning objective tends to close the modality gap
109 regardless of the existence of the cone effect.
- 110 • We reveal that the origin of the modality gap is dimension collapse. And under the subspace
111 constraint, the modality gap converges to the smallest angle between two hyperplanes.
- 112 • We prove that paired samples cannot be perfectly aligned under the subspace constraint.
- 113 • We derive that perfect alignment can be achieved via hyperplane rotation or shared subspace
114 projection.

116 2 PRELIMINARY

119 Suppose we have a dataset $D = \{(I_n, T_n)\}_{n=1}^N$ of N image-text pairs, where $I = (i_1, \dots, i_N) \in$
120 $(\mathcal{I})^N$ and $T = (t_1, \dots, t_N) \in (\mathcal{T})^N$. The unit hypersphere in \mathbb{R}^h is defined as $\mathbb{S}^{h-1} =$
121 $\{z \in \mathbb{R}^h : \|z\| = 1\}$. An image encoder $f_I(\cdot) : \mathcal{I} \rightarrow \mathbb{R}^h$ and a text encoder $f_T(\cdot) : \mathcal{T} \rightarrow \mathbb{R}^h$ map
122 image and text data, respectively, into a shared embedding space. The resulting representations are de-
123 noted as $X = (f_I(i_1), \dots, f_I(i_N)) = (x_1, \dots, x_N) \in (\mathbb{S}^{h-1})^N$ and $Y = (f_T(t_1), \dots, f_T(t_N)) =$
124 $(y_1, \dots, y_N) \in (\mathbb{S}^{h-1})^N$.

125 **Multimodal Contrastive Learning (MCL).** MCL aims to embed data from different modalities into
126 a shared embedding space. This is achieved by minimizing the MCL loss, defined as:

127 **Definition 1** (Multimodal Contrastive Loss (MCL Loss)). *Let (X, Y) be an N -pair configuration,
128 where $X = (x_1, \dots, x_N) \in (\mathbb{S}^{h-1})^N$ and $Y = (y_1, \dots, y_N) \in (\mathbb{S}^{h-1})^N$. $\forall \tau > 0$, the multimodal
129 contrastive loss $\mathcal{L}_{\text{MCL}}(\cdot, \cdot) : (\mathbb{S}^{h-1})^N \times (\mathbb{S}^{h-1})^N \rightarrow \mathbb{R}$ is defined as:*

$$130 \quad \mathcal{L}_{\text{MCL}} = \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{\text{MCL}}^i, \text{ where } \mathcal{L}_{\text{MCL}}^i = \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) + \mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}(y_i; X). \quad (1)$$

133 Here, $\mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}$ is the \mathcal{X} -to- \mathcal{Y} alignment and $\mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}$ is the \mathcal{Y} -to- \mathcal{X} alignment, defined respectively as:

$$135 \quad \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) = -\log \frac{\exp(x_i \cdot y_i / \tau)}{\sum_{j=1}^N \exp(x_i \cdot y_j / \tau)}, \quad \mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}(y_i; X) = -\log \frac{\exp(x_i \cdot y_i / \tau)}{\sum_{j=1}^N \exp(x_j \cdot y_i / \tau)}. \quad (2)$$

138 In practice, contrastive learning is performed in a batch-wise manner due to memory limitations. For
139 analytical simplicity, we assume unlimited memory to train on all samples in a single batch.

140 **Modality Gap.** Define $\mu_x = \frac{1}{N} \sum_{i=1}^N x_i$, $c_x = \frac{\mu_x}{\|\mu_x\|}$ as the mean and the center representation of X ,
141 $\mu_y = \frac{1}{N} \sum_{i=1}^N y_i$, $c_y = \frac{\mu_y}{\|\mu_y\|}$ as the mean and the center representation of Y .

143 **Definition 2** (Modality Gap). *Let (X, Y) be an N -pair configuration, where $X = (x_1, \dots, x_N) \in$
144 $(\mathbb{S}^{h-1})^N$ and $Y = (y_1, \dots, y_N) \in (\mathbb{S}^{h-1})^N$. The modality gap between X and Y can be defined as
145 the difference between their mean representations:*

$$146 \quad \Delta_\mu = \|\mu_x - \mu_y\|_2, \quad (3)$$

148 or as the angle between their center representations:

$$149 \quad \Delta_\theta = \cos^{-1}(c_x \cdot c_y). \quad (4)$$

151 In this study, we use Eq. (4) to define the *modality gap*.

152 3 REPRESENTATIONAL CONVERGENCE AND MODALITY GAP

155 In this section, we study the relationship between MCL and the modality gap. To understand this,
156 we establish a theoretical framework for analyzing the convergent optimal representations (COR) of
157 (X, Y) . We prove that, with or without the cone constraint, as the MCL loss approaches its minimum,
158 the modality gap converges to **zero**.

159 Both cases implicitly assume that X and Y are embedded in the same space \mathbb{S}^{h-1} . Empirical evidence,
160 however, shows that X and Y tend to collapse into different subspaces. We further demonstrate that
161 if X and Y lie in two distinct hyperplanes, then when the MCL loss is minimized, the modality gap
converges to the **smallest angle between the two hyperplanes**.

162 3.1 VON MISES–FISHER (vMF) DISTRIBUTIONS
163

164 As shown in (Liang et al., 2022), when a model is initialized, the representations of each modality
165 reside within a hypercone (Fig. 1a). During training, the representation distribution evolves as the
166 size and shape of the hypercone change. The von Mises-Fisher (vMF) distribution (Mardia & Jupp,
167 2009), a generalization of the normal distribution on the surface of a hypersphere, also concentrates
168 its samples within a hypercone. Hence, this distribution provides as an effective proxy for studying
169 the geometric and distributional properties of representations learned by MCL.

170 **Definition 3** (vMF Distribution). *forall $c \in \mathbb{S}^{h-1}$ and $\kappa \geq 0$, the probability density of a random
171 h -dimensional unit vector $z \sim \text{vMF}(c, \kappa)$ is given by:*

$$172 \quad f_h(z; c, \kappa) = D_h(\kappa) e^{\kappa c^\top z}, \text{ where } D_h(\kappa) = \frac{\kappa^\nu}{(2\pi)^{\nu+1} I_\nu(\kappa)}. \quad (5)$$

173 Here, $\nu = h/2 - 1$, and $I_\nu(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ is the modified Bessel function of the first kind of order ν ,
174 which is defined as:

$$175 \quad I_\nu(x) = \sum_{k=0}^{\infty} \frac{1}{k! \Gamma(\nu + k + 1)} \left(\frac{x}{2}\right)^{2k+\nu}. \quad (6)$$

176 c denotes the center vector and $\frac{1}{\kappa}$ denotes the dispersion degree. When $\frac{1}{\kappa} = \infty$, the samples are
177 maximally dispersed and uniformly distributed on \mathbb{S}^{h-1} . As $\frac{1}{\kappa}$ decreases, the samples become
178 increasingly concentrated and cluster within a smaller hypercone. When $\frac{1}{\kappa} = 0$, the samples are fully
179 concentrated and collapse to a single point. Throughout this work, we assume that (X, Y) are iid
180 samples from two vMF distributions, that is, $x_i \sim \text{vMF}(c_x, \kappa_x)$ and $y_i \sim \text{vMF}(c_y, \kappa_y)$.
181

182 3.2 REPRESENTATIONAL CONVERGENCE WITHOUT DISTRIBUTIONAL CONSTRAINT
183

184 First, we assume that the encoders, f_I and f_T , are sufficiently powerful, capable of realizing any
185 representation distribution without any constraints. Theorem 1 reveals that when the limit of \mathcal{L}_{MCL}
186 attains its minimum, the representations of each paired sample (x_i, y_i) converge to the **same point**,
187 while the representations of all pairs converge to the **uniform distribution** in \mathbb{S}^{h-1} (Fig. 1b).

188 **Theorem 1.** *Let (X, Y) be an N -pair configuration, where $X = (x_1, \dots, x_N) \in (\mathbb{S}^{h-1})^N$ are iid
189 samples from μ_x and $Y = (y_1, \dots, y_N) \in (\mathbb{S}^{h-1})^N$ are iid samples from μ_y . Let $\nu = h/2 - 1$, it
190 holds that:*

$$191 \quad \lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}} - 2 \log(N) = \mathbb{E}_{x_i \sim \mu_x} \left[-\frac{x_i \cdot y_i}{\tau} \right] + \mathbb{E}_{x_i \sim \mu_x} \left[\log \mathbb{E}_{y_i \sim \mu_y} \left[\exp \left(\frac{x_i \cdot y_i}{\tau} \right) \right] \right] \\ 192 \quad + \mathbb{E}_{y_i \sim \mu_y} \left[-\frac{x_i \cdot y_i}{\tau} \right] + \mathbb{E}_{y_i \sim \mu_y} \left[\log \mathbb{E}_{x_j \sim \mu_x} \left[\exp \left(\frac{x_i \cdot y_i}{\tau} \right) \right] \right] \quad (7) \\ 193 \quad \geq -2/\tau + 2 \log(\Gamma(\nu + 1) (2\tau)^\nu I_\nu(1/\tau)),$$

194 where equality is attained if and only if there exists a configuration of (X, Y) such that:

195 (A1) $\forall i \in [N], x_i = y_i$.

196 (A2) $\mu_x = \sigma_{h-1}$ and $\mu_y = \sigma_{h-1}$.

197 Here, σ_{h-1} denotes the uniform probability measure on \mathbb{S}^{h-1} . The proof is provided in Sec. E.1.
198 Under the assumption that X and Y are drawn from two vMF distributions, Corollary 1 implies that
199 when the limit of \mathcal{L}_{MCL} attains its minimum, the modality gap converges to **zero** ($\Delta_\theta \rightarrow 0$), and both
200 κ_x and κ_y converge to **zero**. This result follows directly from Theorem 1.

201 **Corollary 1.** *Let (X, Y) be an N -pair configuration, where $X = (x_1, \dots, x_N) \in (\mathbb{S}^{h-1})^N$ are iid
202 samples from $\text{vMF}(c_x, \kappa_x)$, and $Y = (y_1, \dots, y_N) \in (\mathbb{S}^{h-1})^N$ are iid samples from $\text{vMF}(c_y, \kappa_y)$.
203 $\lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}} - 2 \log(N)$ attains its minimum if and only if the following conditions hold:*

204 (A3) $\forall i \in [N], x_i = y_i (\Rightarrow \Delta_\theta = \cos^{-1}(c_x \cdot c_y) = 0)$.

205 (A4) $\kappa_x = \kappa_y = 0$.

206 **Convergence 1:** Without any distributional constraints, X and Y converge to a paired uniform
207 distribution on \mathbb{S}^{h-1} , and the modality gap converges to zero.

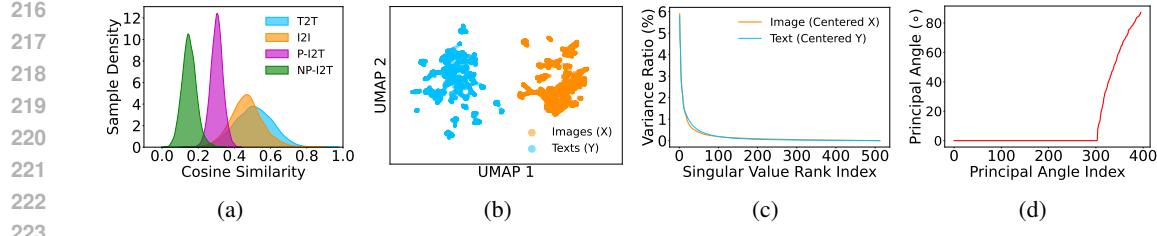


Figure 2: Distributional Constraints. CLIP ViT-B/32 embeddings of MSCOCO validation set. **(a)**: Density plot of cosine similarities between image and image (I2I), text and text (T2T), paired image and text (P-I2T) and unpaired image and text (NP-I2T). **(b)**: UMAP plot. **(c)**: Explained variance ratio of singular values of $X - \mu_X$ and $Y - \mu_Y$. **(d)**: Value of principal angles.

3.3 REPRESENTATIONAL CONVERGENCE UNDER THE CONE CONSTRAINT

However, in practice, sufficiently powerful encoders are not available. Fig. 2b reveals that intra-modal similarities between two modalities are larger than inter-modal similarities. Fig. 2a further shows that (X, Y) separate into two clusters. Both indicate that X and Y lie in two hypercones on \mathbb{S}^{h-1} .

In this subsection, we assume that the encoders, f_I and f_T , are powerful to the extent that (X, Y) are embedded in two hypercones spanning all dimensions of \mathbb{S}^{h-1} , i.e., (X, Y) are subject to the *cone constraint*. In this case, $\kappa_x > 0$ and $\kappa_y > 0$. Since the modality gap depends solely on the angle between the two center vectors, we focus on the configuration of (c_x, c_y) and their corresponding loss terms: $\mathcal{L}_{MCL}^c = \mathcal{L}_{X \rightarrow Y}(c_x; Y) + \mathcal{L}_{Y \rightarrow X}(c_y; X)$. We first define a convergence function \mathcal{J} .

Definition 4. $\forall \kappa, \nu, \tau > 0$, a function $\mathcal{J}(\cdot; \kappa, \nu) : [-1, 1] \rightarrow \mathbb{R}$ is defined as:

$$\mathcal{J}(w; \kappa, \nu) = -\frac{w}{\tau} + \log\left(\frac{I_\nu(M_\kappa(w))}{M_\kappa(w)^\nu}\right) - \log\left(\frac{I_\nu(\kappa)}{\kappa^\nu}\right), \quad (8)$$

where the function $M_\kappa(\cdot) : [-1, 1] \rightarrow \mathbb{R}_0^+$ is defined as:

$$M_\kappa(w) = \sqrt{\kappa^2 + \frac{2\kappa w}{\tau} + \frac{1}{\tau^2}}. \quad (9)$$

Then, Theorem 2 shows that when the limit of \mathcal{L}_{MCL}^c attains its minimum, the modality gap converges to zero ($\Delta_\theta \rightarrow 0$) (Fig. 1c).

Theorem 2. Let (X, Y) be an N -pair configuration, where $X = (x_1, \dots, x_N) \in (\mathbb{S}^{h-1})^N$ are iid samples from $\mu_x = \text{vMF}(c_x, \kappa_x)$, and $Y = (y_1, \dots, y_N) \in (\mathbb{S}^{h-1})^N$ are iid samples from $\mu_y = \text{vMF}(c_y, \kappa_y)$. Let $\nu = h/2 - 1$. Suppose there exists an index $i = c$ such that $x_c = c_x$, $y_c = c_y$. Denote $\Delta_\theta = \cos^{-1}(c_x \cdot c_y)$. For any fixed $\kappa_x, \kappa_y > 0$, it holds that:

$$\begin{aligned} \lim_{N \rightarrow \infty} \mathcal{L}_{MCL}^c - 2 \log(N) &= \mathcal{J}(\cos(\Delta_\theta); \kappa_y, \nu) + \mathcal{J}(\cos(\Delta_\theta); \kappa_x, \nu) \\ &\geq \mathcal{J}(1; \kappa_y, \nu) + \mathcal{J}(1; \kappa_x, \nu), \end{aligned} \quad (10)$$

where equality is attained if and only if there exists a configuration of (X, Y) such that:

$$(A5) \quad \Delta_\theta = \cos^{-1}(c_x \cdot c_y) = 0.$$

The proof is provided in Sec. E.2. Since the distributions of X and Y are symmetric, non-center pairs $(x_i, y_i)_{i \neq c}$ do not affect the configuration of (c_x, c_y) , as confirmed by Theorem 4.

Convergence 2: Under the cone constraint, the modality gap still converges to zero.

3.4 REPRESENTATIONAL CONVERGENCE UNDER THE SUBSPACES CONSTRAINT

To investigate whether X and Y collapse into subspaces of \mathbb{S}^{h-1} , we plot singular values σ_i of the centered X and the centered Y in Fig. 2c. Zero σ_i s confirm dimension collapse. Fig. 2d shows the principal angles γ_i between the subspaces where X and Y collapse. Zero γ_i s imply that the two subspaces share overlapped dimensions. Detailed explanations are provided in Sec. C.1 and Sec. C.2.

In this subsection, we assume that the encoders, f_I and f_T , embed (X, Y) into two partially overlapping subspaces of \mathbb{S}^{h-1} (Fig. 1d), i.e., (X, Y) are subject to the *subspace constraint*. To simplify the analysis, we require that the two subspaces are hyperplanes, as described below:

Definition 5. Let \mathbb{A} and \mathbb{B} be two distinct $(h-1)$ -dimensional linear subspaces (i.e., hyperplanes through the origin) with normal vectors n_A and n_B , projection matrices P_A and P_B . Denote $\mathbb{C} = \mathbb{A} \cap \mathbb{B}$, with P_C as its projection matrix. Define $\phi = \cos^{-1} \left(\frac{n_A \cdot n_B}{\|n_A\| \cdot \|n_B\|} \right)$ as the angle between \mathbb{A} and \mathbb{B} , restricted to $0 < \phi_{\min} \leq \phi < \frac{\pi}{2}$. Then, \mathbb{S}_X and \mathbb{S}_Y can be represented as:

$$\begin{aligned} \mathbb{S}_X &= \mathbb{S}^{h-1} \cap \mathbb{A} = \{x \in \mathbb{R}^h : \|x\| = 1, n_A \cdot x = 0\} \cong \mathbb{S}^{h-2} \in \mathbb{S}^{h-1}, \\ \mathbb{S}_Y &= \mathbb{S}^{h-1} \cap \mathbb{B} = \{y \in \mathbb{R}^h : \|y\| = 1, n_B \cdot y = 0\} \cong \mathbb{S}^{h-2} \in \mathbb{S}^{h-1}. \end{aligned} \quad (11)$$

\mathbb{C} is an $(h-2)$ dimensional linear subspace (Strang, 2022). We now define a convergence function $\tilde{\mathcal{J}}$. Note that function \mathcal{J} in Definition 4 is a special case of $\tilde{\mathcal{J}}$ with $\mathcal{J}(w; \kappa, \nu) = \tilde{\mathcal{J}}(w, w, 1; \kappa, \nu)$.

Definition 6. $\forall \kappa, \nu, \tau > 0$, $\tilde{\mathcal{J}}(\cdot, \cdot, \cdot; \kappa, \nu) : [-1, 1] \times [-1, 1] \times [0, 1] \rightarrow \mathbb{R}$ is defined as:

$$\tilde{\mathcal{J}}(w_1, w_2, t; \kappa, \nu) = -\frac{w_1}{\tau} + \log \left(\frac{I_\nu \left(\tilde{M}_\kappa(w_2, t) \right)}{\tilde{M}_\kappa(w_2, t)^\nu} \right) - \log \left(\frac{I_\nu(\kappa)}{\kappa^\nu} \right), \quad (12)$$

where the function $\tilde{M}_\kappa(\cdot, \cdot) : [-1, 1] \times [0, 1] \rightarrow \mathbb{R}_0^+$ is defined as:

$$\tilde{M}_\kappa(w, t) = \sqrt{\kappa^2 + \frac{2\kappa w}{\tau} + \frac{t^2}{\tau^2}}. \quad (13)$$

Theorem 3 shows that when the limit of $\mathcal{L}_{\text{MCL}}^c$ attains its minimum, c_x, c_y are orthogonal to \mathbb{C} , and the modality gap converges to the **smallest angle between \mathbb{A} and \mathbb{B}** ($\Delta_\theta \rightarrow \phi_{\min}$) (Fig. 1e).

Theorem 3. Let (X, Y) be an N -pair configuration, where $X = (x_1, \dots, x_N) \in (\mathbb{S}_X \setminus \mathbb{C})^N$ are iid samples from $\mu_x = \text{vMF}(c_x, \kappa_x)$, and $Y = (y_1, \dots, y_N) \in (\mathbb{S}_Y \setminus \mathbb{C})^N$ are iid samples from $\mu_y = \text{vMF}(c_y, \kappa_y)$. Let $\tilde{\nu} = (h-1)/2 - 1$. Suppose there exists an index $i = c$ such that $x_c = c_x$, $y_c = c_y$. Denote $\Delta_\theta = \cos^{-1}(c_x \cdot c_y)$ and assume that $c_x, c_y \notin \mathbb{C}$ with $c_x \cdot c_y > 0$. For any fixed $\kappa_x, \kappa_y > 0$, it holds that:

$$\begin{aligned} &\lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}}^c - 2 \log(N) \\ &= \tilde{\mathcal{J}}(\cos(\Delta_\theta), \cos(\Delta_\theta), \|P_B c_x\|; \kappa_y, \tilde{\nu}) + \tilde{\mathcal{J}}(\cos(\Delta_\theta), \cos(\Delta_\theta), \|P_A c_y\|; \kappa_x, \tilde{\nu}) \\ &\geq \tilde{\mathcal{J}}(\cos(\phi_{\min}), \cos(\phi_{\min}), \cos(\phi_{\min}); \kappa_y, \tilde{\nu}) + \tilde{\mathcal{J}}(\cos(\phi_{\min}), \cos(\phi_{\min}), \cos(\phi_{\min}); \kappa_x, \tilde{\nu}), \end{aligned} \quad (14)$$

where equality is attained if and only if there exists a configuration of (X, Y) such that:

$$(A6) \quad c_x \perp \mathbb{C} \text{ and } c_y \perp \mathbb{C} (\Rightarrow \Delta_\theta = \phi).$$

$$(A7) \quad \Delta_\theta = \cos^{-1}(c_x \cdot c_y) = \phi_{\min}.$$

The proof is provided in Sec. E.3. Condition (A6) shows the optimal configuration of (c_x, c_y) for any given ϕ . Condition (A7) establishes that the loss decreases monotonically as ϕ decreases to ϕ_{\min} . Since the distributions of X and Y are symmetric, non-center pairs $(x_i, y_i)_{i \neq c}$ do not affect Condition (A6). Moreover, optimizing of $\mathcal{L}_{\text{MCL}}^{i \neq c}$ also yields Condition (A7), as shown in Theorem 4.

Convergence 3: Under the subspace constraint, the modality gap converges to the smallest angle between the two hyperplanes.

4 REPRESENTATIONAL CONVERGENCE AND MODALITY ALIGNMENT

In Sec. 3.4, we identified the true origin of the modality gap by analyzing the configuration of the center pair. However, the relationship between the modality gap and downstream performance, which depends on the configuration of all pairs, remains unclear. In this section, we show that, under the subspace constraint, non-center pairs cannot be perfectly aligned when the MCL loss is minimized.

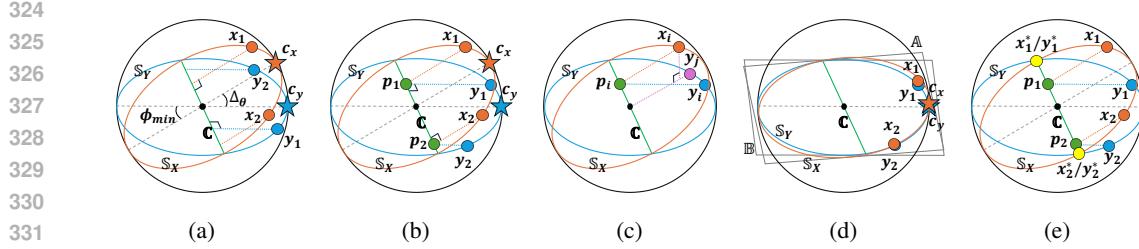


Figure 3: Modality Alignment. Notations follow Fig. 1. **(a)**: Condition (A6) ($c_x, c_y \perp \mathbb{C}$) and IMS ($x_i \cdot c_x = y_i \cdot c_y$) hold. **(b)**: The projections of $(x_i, y_i)_{i \neq c}$ on \mathbb{C} converge to p_i (green point), i.e., $P_C x_i = P_C y_i = p_i$. **(c)**: When Condition (A6) and (A8) ($P_C x_i = P_C y_i$) hold, $P_B x_i \nparallel y_i$, $P_A y_i \nparallel x_i$. Denote $y_j = \frac{P_B x_i}{\|P_B x_i\|}$ (purple dot), then $x_i \cdot y_j > x_i \cdot y_i$ and $(x_i, y_i)_{i \neq c}$ are not perfectly aligned. **(d)**: Rotating X with the hyperplane \mathbb{A} towards \mathbb{B} , X and Y can be aligned perfectly. **(e)**: Project x_i and y_i onto \mathbb{C} and re-normalize, then x_i^* and y_i^* (yellow dots) are perfectly aligned.

4.1 INTRA-MODAL ISOMETRY AND PERFECT ALIGNMENT

The Platonic Representation Hypothesis (Huh et al., 2024) suggests that contrastive learners are optimized by representations of X and Y whose intra-modal kernels (i.e., pairwise similarities) align. Building on this idea, we define the kernel alignment as *Intra-Modal Isometry*.

Definition 7 (Intra-Modal Isometry (IMS)). *Let (X, Y) be an N -pair configuration in \mathbb{R}^h , we say (X, Y) achieves Intra-Modal Isometry if and only if $\forall i, j \in [N], i \neq j, x_i \cdot x_j = y_i \cdot y_j$.*

The Intra-Modal Isometry assumption implies that $\forall i \in [N], x_i \cdot c_x = y_i \cdot c_y$, and thus $\kappa_x = \kappa_y$ (Fig. 3a). However, knowledge of the intra-modal configuration alone is insufficient to determine how the modality gap affects downstream performance. In downstream tasks such as zero-shot image classification, given an input from one modality (e.g., x_i), CLIP retrieves data from the other modality (e.g., y_j) with the largest similarity to the input. Ideally, the output should be $y_j = y_i$. We therefore define an ideal inter-modal configuration as *Perfect Alignment*. And when Perfect Alignment is achieved, downstream performance is maximized.

Definition 8 (Perfect Alignment). *Let (X, Y) be an N -pair configuration in \mathbb{R}^h , we say (x_i, y_i) is perfectly aligned if and only if $\forall j \neq i, x_i \cdot y_j > x_i \cdot y_j$ and $x_i \cdot y_i > x_j \cdot y_x$. If $\forall i \in [N], (x_i, y_i)$ is perfectly aligned, we say (X, Y) achieves Perfect Alignment.*

4.2 REPRESENTATIONAL CONVERGENCE OF NON-CENTER PAIRS

To investigate the alignment between two modalities, we examine the optimal configuration of each data pair. Theorem 4 states that if Condition (A6) (in Theorem 3) is satisfied through the optimization of $\mathcal{L}_{\text{MCL}}^{i \neq c}$, and if (X, Y) achieves Intra-Modal Isometry (Fig. 3a), then when the limit of $\mathcal{L}_{\text{MCL}}^{i \neq c}$ attains its minimum, the projections of any non-center pair $(x_i, y_i)_{i \neq c}$ onto \mathbb{C} converge to the same vector (Fig. 3b).

Theorem 4. *Let (X, Y) be an N -pair configuration, where $X = (x_1, \dots, x_N) \in (\mathbb{S}_X \setminus \mathbb{C})^N$ are iid samples from $\mu_x = \text{vMF}(c_x, \kappa_x)$, and $Y = (y_1, \dots, y_N) \in (\mathbb{S}_Y \setminus \mathbb{C})^N$ are iid samples from $\mu_y = \text{vMF}(c_y, \kappa_y)$. Let $\tilde{\nu} = (h-1)/2 - 1$. Denote $\Delta_\theta = \cos^{-1}(c_x \cdot c_y)$ and assume $c_x, c_y \perp \mathbb{C}$ with $c_x \cdot c_y > 0$. Suppose (X, Y) achieves Intra-Modal Isometry. Then $\forall i \in [N]$, denote $\theta_i^c = \cos^{-1}(x_i \cdot c_x) = \cos^{-1}(y_i \cdot c_y)$, and $\kappa = \kappa_x = \kappa_y$. Let $\theta_i^c \in (0, \frac{\pi}{2})$ and $\kappa > 0$, it holds that:*

$$\begin{aligned}
 & \lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}}^{i \neq c} - 2 \log(N) \\
 &= \tilde{\mathcal{J}}(\cos(\Delta_\theta), \cos(\theta_i^c), \|P_B x_i\|; \kappa, \tilde{\nu}) + \tilde{\mathcal{J}}(\cos(\Delta_\theta), \cos(\theta_i^c), \|P_A y_i\|; \kappa, \tilde{\nu}) \\
 &\geq 2\tilde{\mathcal{J}}\left(\cos^2(\theta_i^c) \cos(\phi_{\min}) + \sin^2(\theta_i^c), \cos(\theta_i^c), \sqrt{\cos^2(\theta_i^c) \cos^2(\phi_{\min}) + \sin^2(\theta_i^c)}; \kappa, \tilde{\nu}\right), \tag{15}
 \end{aligned}$$

where equality is attained if and only if there exists a configuration of (X, Y) such that:

$$(A8) \quad P_C x_i = P_C y_i.$$

378 (A9) $\Delta_\theta = \cos^{-1}(c_x \cdot c_y) = \phi_{\min}$.
 379

380 The proof of Theorem 4 is provided in Sec. E.4. Condition (A8) characterizes the optimal configuration
 381 of $(x_i, y_i)_{i \neq c}$ for any given ϕ . Condition (A9) establishes that the loss decreases monotonically
 382 as ϕ decreases to ϕ_{\min} , consistent with Condition (A7) of Theorem 3. Moreover, Theorem 4 implies
 383 that MCL aims to **maximize the mutual information between the two modalities in the shared**
 384 **space while preserving modality-specific information in the complementary space.**

385 4.3 REPRESENTATIONAL CONVERGENCE DOSE NOT ENSURE PERFECT ALIGNMENT

387 In Lemma 12, we show that $(x_i, y_i)_{i \neq c}$ are perfectly aligned if and only if the projections of
 388 $(x_i, y_i)_{i \neq c}$ onto \mathbb{B} and \mathbb{A} are collinear, i.e., $P_B x_i \parallel y_i$ and $P_A y_i \parallel x_i$. However, when training is
 389 optimized such that conditions (A6) and (A8) hold, $P_B x_i \nparallel y_i$ and $P_A y_i \nparallel x_i$. This implies that
 390 $(x_i, y_i)_{i \neq c}$ are **not perfectly aligned** (Fig. 3c).

391 **Corollary 2.** $\forall i \in [N], i \neq c$, if $c_x, c_y \perp \mathbb{C}$ and $P_C x_i = P_C y_i \neq \vec{0}$ and $\phi > 0$, then it holds:

393 (A10) $(x_i, y_i)_{i \neq c}$ are not perfectly aligned.

395 The proof of Corollary 2 is provided in Sec. E.4.3. Since the limit of \mathcal{L}_{MCL} attains its minimum
 396 when both $\mathcal{L}_{\text{MCL}}^c$ and $\mathcal{L}_{\text{MCL}}^{i \neq c}$ attain their minima, and since all paired samples are non-center pairs
 397 almost surely (the ‘center’ forms a zero measure set in \mathbb{S}_X or \mathbb{S}_Y), then we conclude that:

398 **Convergence 4:** Under the subspace constraint, paired samples cannot be perfectly aligned.

401 5 SHARED SUBSPACE PROJECTION IMPROVES MODALITY ALIGNMENT

403 In Sec. 4, we prove that the representations of paired samples are not perfectly aligned. Despite
 404 this undesirable configuration, in this section we derive potential methods to improve the alignment
 405 between the two modalities.

406 5.1 HOW TO ACHIEVE PERFECT ALIGNMENT

408 In downstream tasks, when $(x_i, y_i)_{i \neq c}$ are not perfectly aligned, x_i can be misaligned to some
 409 $y_{j \neq i}$ (Fig. 3c). A straightforward way to address this is to manually shift (x_i, y_i) in \mathbb{S}^{h-1} . For
 410 example, Liang et al. (2022) translate x_i toward y_i as $x_i^{\text{new}} = x_i + \Delta_u$, followed by renormalization.
 411 This operation clearly alters the distributions of X . Since downstream performance depends on the
 412 number of misaligned y_j in the test set. A change in the distribution of X leads to a change in the
 413 proportion of misaligned y_j , but in an unpredictable direction. Therefore, the impact of translating X
 414 on downstream performance can be arbitrary. An illustrative example is provided in Sec. B.1.

415 As shown in Fig. 3d, if we rotate \mathbb{A} to overlap with \mathbb{B} , then $\mathbb{A} = \mathbb{B} = \mathbb{C}$. In this case, Condition
 416 (A8) implies $x_i = y_i$, and thus x_i and y_j are **perfectly aligned**. Hence, modality alignment can be
 417 improved by rotating the hyperplanes \mathbb{A} and \mathbb{B} until $\mathbb{A} = \mathbb{B}$ ($\Delta_\theta = \phi = 0$).

418 **Corollary 3.** $\forall i \in [N], i \neq c$, if $c_x, c_y \perp \mathbb{C}$, $P_C x_i = P_C y_i$ and $(x_i, y_i)_{i \neq c} \in \mathbb{S}^{h-1} \setminus \mathbb{C}$, then
 419 $(x_i, y_i)_{i \neq c}$ are perfectly aligned if the following condition holds:

421 (A11) $\Delta_\theta = \phi = 0$.

422 The proof of Corollary 3 is provided in Sec. E.4.3. Despite this theoretical guarantee, rotating a
 423 high-dimensional hyperplane can be complicated in practice. As illustrated in Fig. 3e, if we project
 424 x_i and y_i onto \mathbb{C} and then renormalize, we obtain $x_i^* = y_i^*$. And (x_i^*, y_i^*) are **perfectly aligned**.

425 **Corollary 4.** $\forall i \in [N], i \neq c$, if $c_x, c_y \perp \mathbb{C}$ and $P_C x_i = P_C y_i$, then the following holds:

427 (A12) $(\frac{P_C x_i}{\|P_C x_i\|}, \frac{P_C y_i}{\|P_C y_i\|})_{i \neq c}$ are perfectly aligned

429 The proof of Corollary 4 is provided in Sec. E.4.3. Note that in Fig. 3e, \mathbb{C} is a 1D line, so all
 430 transformed paired samples overlap at y_1^* and y_2^* . In practice, however, the dimension of \mathbb{C} is
 431 typically greater than 1 (e.g., 212D for MS-COCO dataset). For instance, in the 4D example in Fig. 6
 of Sec. B.2, \mathbb{C} is a 2D plane, and the samples are distributed along a unit circle.

432 Table 1: Size of θ_Δ and accuracies (%) of zero-shot image classification of ViT-B/32.
433

434 Model	435 CIFAR-10			436 CIFAR-100			437 ImageNet-1K		
	438 Δ_θ	439 R1	440 R5	438 Δ_θ	439 R1	440 R5	438 Δ_θ	439 R1	440 R5
CLIP	74.69°	89.00	99.36	74.19°	65.23	88.88	71.02°	63.34	88.82
CLIP + Translation	7.02°	80.97	96.09	30.50°	54.46	77.25	51.68°	60.37	86.93
CLIP + Removal	72.5°	14.91	56.22	73.16°	16.82	6.44	69.71°	49.50	78.55
CLIP + SSP	5.37°	86.43	99.27	30.39°	64.51	88.79	50.40°	62.45	88.41

441
442 5.2 EXPERIMENT
443

444 Theorem 4, Corollary 3 and Corollary 4 suggest that if projections of X and Y are aligned in the
445 shared space, modality alignment can be improved. This also indicates that the modality gap can be
446 reduced pos hoc without harming downstream performance.

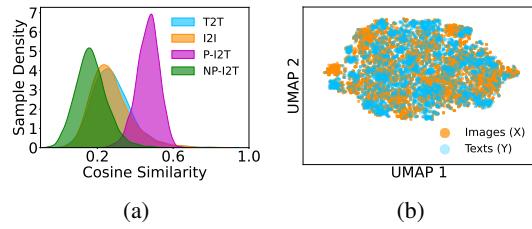
447 **Method.** Following Corollary 4, we apply the shared space projection (SSP) method pos hoc to
448 improve the alignment of the modality. Detailed procedures are described in Sec. C.3.

449 **Modality Alignment.** To validate the effectiveness of our method, we start by visualizing X
450 and Y after applying SSP. We first project X and Y onto an estimated shared space of 212
451 dimensions. Fig. 4a (vs. Fig. 2a) shows the cosine similarities of the projected X and Y . It
452 indicates that our method improves both inter-modal alignment (larger P-I2T) and intra-modal
453 uniformity (smaller T2T and I2I). Since the shared space is not estimated from the original
454 training data, the estimation can be noisy. Hence, we select a 10 dimensional subspace of
455 the estimated shared space to reduce the estimation error (details explained in Sec. C.3). We
456 project X and Y onto this subspace. Fig. 4b (vs. Fig. 2b) shows that the projected X and Y are no
457 longer in separate clusters.

458 **Zero-Shot Image Classification.** We also test our method in the zero-shot image classification task
459 on various datasets. Details of this experiment are provided in Sec. D.1. Our goal is to reduce the
460 size of the modality gap as much as possible without harming downstream performance. In Tab. 1,
461 we list results of the size of the modality gap (Δ_θ), the top-1 accuracy (R1), and the top-5 accuracy
462 (R5). We include two baseline methods: a translation-based approach (Liang et al., 2022) and a
463 dimension-removal approach (Schrodi et al., 2025). Our results show that our method outperforms
464 these baselines by achieving a greater reduction in the modality gap while maintaining comparable
465 downstream performance prior to the post hoc operation. Despite its advantages, our method does not
466 lead to improved downstream performance, as indicated in Corollary 4. We argue that this limitation
467 arises because the intra-model isometry assumption does not hold in CLIP. Prior work has shown that
468 CLIP’s vision and text spaces exhibit different neighborhood structures (Udandarao, 2022; Schrodi
469 et al., 2025). We provide additional experiments of **Zero-Shot Cross-Modal Retrieval** in Sec. D.2.

470
471 6 CONCLUSION
472

473 Our work comprehensively investigates two key questions: (1) What causes the modality gap? (2)
474 How does it affect downstream tasks? Our theorems identify *dimension collapse* as the fundamental
475 origin of the modality gap. Our theorems also demonstrate that paired samples cannot be perfectly
476 aligned under the subspace constraint. We further prove that two approaches, hyperplane rotation and
477 shared space projection, can achieve perfect alignment between two modalities. We apply the latter
478 approach post-hoc and validate its effectiveness in downstream tasks. Besides the pos hoc application,
479 our method has potential to be applied to pretraining. It can directly optimize modality alignment in
480 the shared space to achieve the intra-modal isometry. We will explore it in the next step.



481 Figure 4: Results. CLIP ViT-B/32 embeddings of
482 MSCOCO validation set after applying SSP are
483 used. (a): UMAP plot. (b): Density plot of cosine
484 similarities.

486 REFERENCES
487

488 Árpád Baricz. Turán type inequalities for modified bessel functions. *Bulletin of the Australian*
489 *Mathematical Society*, 2010.

490 Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for
491 contrastive learning of visual representations. In *ICML*, 2020.

492

493 Sanghyuk Chun. Multiplicity is an inevitable and inherent challenge in multimodal learning. *arXiv*,
494 2025.

495 Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale
496 hierarchical image database. In *CVPR*, 2009.

497

498 Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Improving cross-modal alignment in clip. In
499 *ICLR*, 2025.

500

501 Abrar Fahim, Alex Murphy, and Alona Fyshe. It's not a modality gap: Characterizing and addressing
502 the contrastive gap. *arXiv*, 2024.

503

504 Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised con-
505 trastive learning. In *ICML*, 2021.

506

507 Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation
508 hypothesis. In *ICML*, 2024.

509

510 Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in
511 contrastive self-supervised learning. In *ICLR*, 2022.

512

513 Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron
514 Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020.

515

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

517

518 Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image
519 pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

520

521 Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap:
522 Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*,
523 2022.

524

525 Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr
526 Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

527

528 Litian Liu and Yao Qin. Detecting out-of-distribution through the lens of neural collapse. In *CVPR*,
529 2025.

530

531 Kanti V Mardia and Peter E Jupp. *Directional statistics*. John Wiley & Sons, 2009.

532

533 Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets
534 language-image pre-training. In *ECCV*, 2022.

535

536 Frank WJ Olver. *NIST handbook of mathematical functions*. Cambridge university press, 2010.

537

538 Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal
539 phase of deep learning training. *Proceedings of the National Academy of Sciences*, 2020.

540

541 Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and
542 Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer
543 image-to-sentence models. In *ICCV*, 2015.

544

545 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,
546 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual
547 models from natural language supervision. In *ICML*, 2021.

540 Simon Schrodi, David T Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one
541 trigger: on the modality gap, object bias, and information imbalance in contrastive vision-language
542 representation learning. In *ICLR*, 2025.

543

544 PY Shi, M Welle, M Bjørkman, and D Kragic. Understanding the modality gap in clip. *ICLR*
545 *Workshop*, 2023.

546

547 Gilbert Strang. *Introduction to linear algebra*. SIAM, 2022.

548

549 Vishaal Udandarao. Understanding and fixing the modality gap in vision-language models. *Master's*
550 *thesis, University of Cambridge*, 2022.

551

552 Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint*
553 *arXiv:1011.3027*, 2010.

554

555 Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through align-
556 ment and uniformity on the hypersphere. In *ICML*, 2020.

557

558 Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised Feature Learning via
559 Non-parametric Instance Discrimination. In *CVPR*, 2018.

560

561 Can Yaras, Siyi Chen, Peng Wang, and Qing Qu. Explaining and mitigating the modality gap in
562 contrastive multimodal learning. *arXiv*, 2024.

563

564 Lingjie Yi, Tao Sun, Yikai Zhang, Songzhu Zheng, Weimin Lyu, Haibin Ling, and Chao Chen.
565 Pivotalign: Improve semi-supervised learning by learning intra-class heterogeneity and aligning
566 with pivots. In *WACV*, 2025a.

567

568

569

570

571

572

573

574

575

576

577

578

579

580

581

582

583

584

585

586

587

588

589

590

591

592

593

594 **A APPENDIX A: MORE DISCUSSIONS**
595596 **A.1 RELATED WORK**
597598 Due to the page limit of the initial submission (9 pages), we include the related work here. In the
599 final version (10 pages), this section will be moved into the main text.
600601 **A.1.1 REPRESENTATION LEARNING AND REPRESENTATIONAL CONVERGENCE**
602603 Unimodal representations can be learned in an unsupervised manner using self-supervised contrastive
604 learning (SSL) (Chen et al., 2020). When the InfoNCE loss (Wu et al., 2018) reaches its minimum,
605 the representations of differently augmented views of an image converge to a single point, and
606 the representation of all images converge to a uniform distribution on \mathbb{S}^{h-1} (Wang & Isola, 2020).
607 However, Jing et al. (2022) empirically shows that this theoretical optimum may not be realized in
608 practice: the learned representations tend to collapse into a lower-dimensional subspace rather than
609 spanning the entire embedding space.
610611 In the supervised setting, representations can be learned through a neural classifier. When the cross-
612 entropy loss is minimized, representations of samples from different balanced classes converge to the
613 vertices of a regular simplex inscribed in \mathbb{S}^{h-1} , a phenomenon known as *neural collapse* (Papyan
614 et al., 2020). Graf et al. (2021) provide a theoretical explanation of this phenomenon. Representations
615 can also be learned with supervised contrastive learning (SupCon) (Khosla et al., 2020). Graf et al.
616 (2021) prove that the COR of a balanced dataset of SupCon also forms a regular simplex. Yi et al.
617 (2025b) provide a refined proof and further show that, for imbalanced datasets, representations
618 converge to a skewed simplex or even collapse into two distinct points. Other works extend the
619 concept of neural collapse to semi-supervised learning (Yi et al., 2025a) and OOD detection (Liu &
620 Qin, 2025).
621622 Multimodal representations are learned through multimodal contrastive learning (MCL). However,
623 the COR of MCL remains poorly understood. In this work, we address this gap by characterizing
624 the COR of MCL. Our theorems suggest that MCL seeks to maximize the mutual information
625 between the two modalities in the shared space while preserving modality-specific information in the
626 complementary space.
627628 **A.1.2 MODALITY GAP**
629630 Liang et al. (2022) first identified the modality gap, a geometric phenomenon characterized by
631 the complete separation of representations of different modalities in the embedding space. They
632 hypothesize that the gap arises from the cone effect due to random model initialization and is
633 preserved by the contrastive learning objective. Fahim et al. (2024) argues that the modality gap is
634 inherent to contrastive loss. Yaras et al. (2024); Udandarao (2022) examine the role of mismatch pairs
635 and the temperature parameter. Shi et al. (2023) attribute the cause of the modality gap to insufficient
636 training. Schrödi et al. (2025) suggests that problematic training data, which contain information bias,
637 create the gap. Most of these works validate their hypotheses through numerical examples on a small
638 number of data pairs. By contrast, we provide an analysis based on the entire distribution.
639640 In addition, several studies have proposed post-hoc methods to mitigate the modality gap. Liang
641 et al. (2022) attempts to translate the representations of one modality toward those of another using a
642 constant shift. Schrödi et al. (2025) explores removing the few dimensions that primarily drive the
643 modality gap. However, experiments in both works reported that narrowing the modality gap pos hoc
644 may lead to degraded downstream performance. Eslami & de Melo (2025) mitigates the modality gap
645 by retraining CLIP from scratch. Our work focuses on training-free pos-hoc plug-and-play methods
646 that can directly leverage existing pre-trained models.
647648 **A.2 LIMITATIONS**
649650 While our work investigates the origin of the modality gap and attributes it to dimension collapse, we
651 do not address the exact factors that lead to dimension collapse. (Jing et al., 2022) theoretically show
652 that dimension collapse occurs whenever negative eigenvalues appear in the weight matrix of a neural
653 network. (Schrödi et al., 2025) suggests that when training data with information bias are sufficiently
654

aligned, ‘more dimensions’ are required to focus on objects and and ‘less dimensions’ to focus on attributes, ultimately resulting in dimension collapse. (Chun, 2025) provides a more comprehensive study of the inherent challenges within MCL, including intra-modal variability, asymmetries in information, and task-dependent alignment. We suspect that all these factors contribute to dimension collapse in the learned representations. Identifying the causes of dimension collapse thus constitutes a major open problem, parallel to understanding the origin of the modality gap, and represents an important direction for future research.

655 A.3 CONNECTION BETWEEN OUR THEOREMS AND PREVIOUS HYPOTHESES

656 In this subsection, we examine the connection between empirical observations from prior studies and
657 our theoretical conclusions.

658 **Cone Effect:** The cone effect hypothesis (Liang et al., 2022) posits that the representations of X and
659 Y fall into different cones on the hypersphere, thereby causing the modality gap. In our theoretical
660 framework, as described in Sec. 3.1, the cone size of the representations is modeled by the parameter
661 κ . However, in contrast to this hypothesis, Theorem 2 shows that the cone size has no effect on the
662 convergence of the modality gap, even when the representations follow a uniform distribution (i.e.,
663 $\kappa \rightarrow 0$).

664 **Temperature:** It is hypothesized that the choice of temperature contributes to the emergence of
665 the modality gap (Yaras et al., 2024; Udandarao, 2022). However, Theorem 2 suggests that the
666 temperature parameter, τ , has no effect on the convergence of the modality gap. We suspect that if
667 temperature has any impact, it operates indirectly by influencing dimension collapse.

668 **Information Bias:** (Schrodi et al., 2025) argue that information bias, i.e., images containing more
669 information than the corresponding text, leads to the modality gap. The unequal amount of information
670 across modalities prevents Intra-Modal Isometry of the representations (see Definition 7), making it
671 difficult for the model to align representations from the two modalities. This results in sub-optimal
672 inter-modal alignment, which in turn imposes a lower bound on the alignment terms and ensures
673 $\Delta_\theta > 0$. We posit that there is a strong connection between information bias and dimension collapse:
674 information bias induces dimension collapse in the learned representations, thereby causing the
675 modality gap.

676 A.4 DISCLOSURE OF LLM USAGE

677 In the preparation of this paper, we used large language models (LLMs) as general-purpose assistive
678 tools. Specifically, we used an LLM to help with grammar polishing, wording improvements, and
679 proof-reading.

680 Any text or content generated by the LLM have been reviewed and edited by the authors. We take
681 full responsibility for the content of the submission. The LLM was not used to produce novel
682 research claims, data analysis, results formulation, or conclusions. The research ideation, theoretical
683 contributions, experiments, and all core technical work are entirely the work of the authors.

684 B APPENDIX B: MORE SUPPORTING EXAMPLES

685 In this subsection, we provide more examples and illustrations.

686 B.1 ILLUSTRATIVE EXAMPLE OF TRANSLATION-BASED METHOD

687 In this subsection, we provide an illustrative example showing that the impact of translating X pos
688 hoc on downstream performance can be arbitrary. Fig. 5a depicts a set of X and Y where Condition
689 (A6) and Condition (A8) hold. Fig. 5b illustrates how X s are going to be translated. Fig. 5c shows
690 the positions of X^* s after translation. Fig. 5d illustrates how X^* s are going to be normalized. Fig. 5e
691 shows the positions of X^{**} s after normalization. In Fig. 5f, we observe that the distribution of X^{**} s
692 differs substantially from that of X : they no longer reside in the same shared space (the large circle
693 in this example), and their projections onto the shared space diverge from those of Y s. The direction
694 of these changes depends on the specific configuration of X and is therefore unpredictable. Hence,

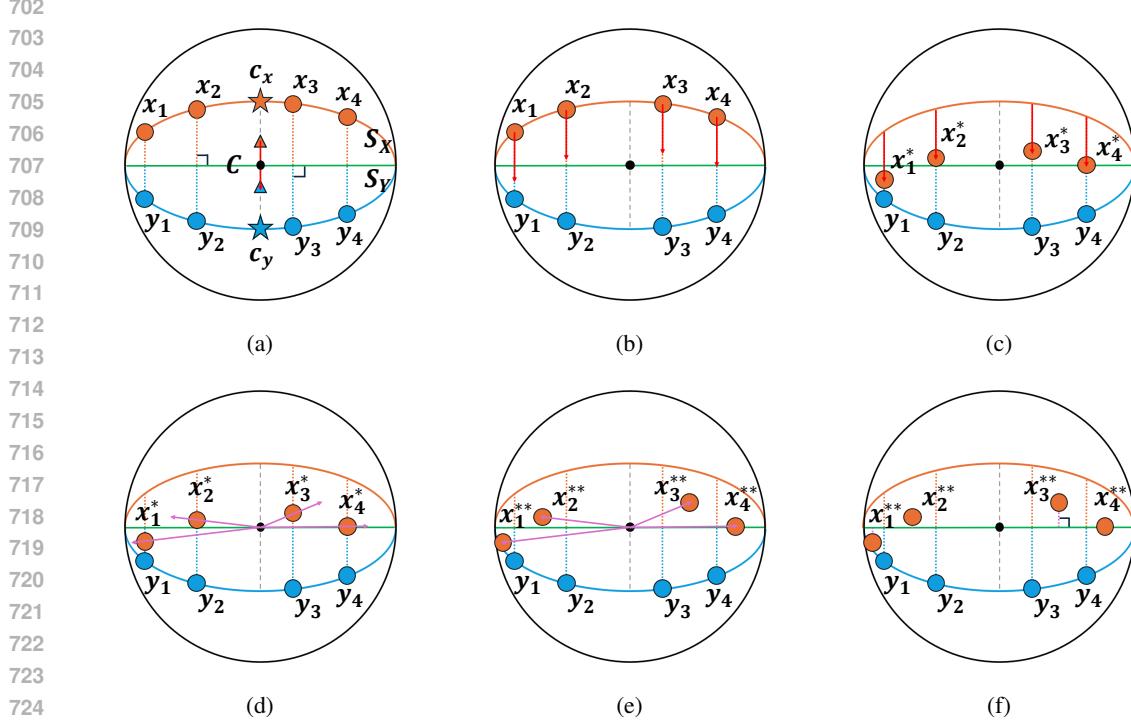


Figure 5: Translation-based method. Notations follow Fig. 1. (a): Condition (A6) ($c_x, c_y \perp \mathbb{C}$) and (A8) ($P_C x_i = P_C y_i$) hold. Orange/blue triangles represents μ_x and μ_y . The red arrows denotes the direction and scale of the constant translation ($\mu_y - \mu_x$). (b): Translating X . (c): X are translated to X^* . (d): X^* are being re-normalized. Purple arrows are denotes the direction and scale of the normalization. (e): X^* are re-normalized to X^{**} . (f): Distribution of X altered after translation with $P_C x_i^{**} \neq P_C y_i$.

the impact of translating X on downstream performance is unpredictable. In practice, the impact is often a negative one.

B.2 ADDITIONAL EXAMPLE OF MODALITY ALIGNMENT

In Sec. 5, we discuss how the shared space projection approach can improve modality alignment. As an illustrative case, Fig. 3e presents an example in a 3D embedding space where \mathbb{C} corresponds to a 1D line. However, this example may be misinterpreted as implying that all transformed paired samples (x_i^*, y_i^*) perfectly overlap at y_1^* and y_2^* . To clarify this point, in this subsection we examine a more intricate example in a 4D embedding space.

First, recall Definition 5 and set $h = 4$:

Definition 5 [Restate with $h = 4$] Let \mathbb{A} and \mathbb{B} be two distinct $(h - 1)$ -dimensional linear subspaces (i.e., hyperplanes through the origin) with normal vectors n_A and n_B , projection matrices P_A and P_B . Denote $\mathbb{C} = \mathbb{A} \cap \mathbb{B}$, with P_C as its projection matrix. Define $\phi = \cos^{-1} \left(\frac{n_A \cdot n_B}{\|n_A\| \cdot \|n_B\|} \right)$ as the angle between \mathbb{A} and \mathbb{B} , restricted to $0 < \phi_{\min} \leq \phi < \frac{\pi}{2}$. Then, \mathbb{S}_X and \mathbb{S}_Y can be represented as:

$$\begin{aligned} \mathbb{S}_X &= \mathbb{S}^3 \cap \mathbb{A} = \{x \in \mathbb{R}^4 : \|x\| = 1, n_A \cdot x = 0\} \cong \mathbb{S}^2 \in \mathbb{S}^3, \\ \mathbb{S}_Y &= \mathbb{S}^3 \cap \mathbb{B} = \{y \in \mathbb{R}^4 : \|y\| = 1, n_B \cdot y = 0\} \cong \mathbb{S}^2 \in \mathbb{S}^3. \end{aligned} \quad (16)$$

Now, \mathbb{S}^3 denotes the 4D unit hypersphere. To analyze this case, we decompose the embedding space. Let $\{e_1, e_2, e_3, e_4\}$ be an orthonormal basis of \mathbb{R}^4 . Suppose that the shared space \mathbb{C} lies within the span of e_1 and e_2 :

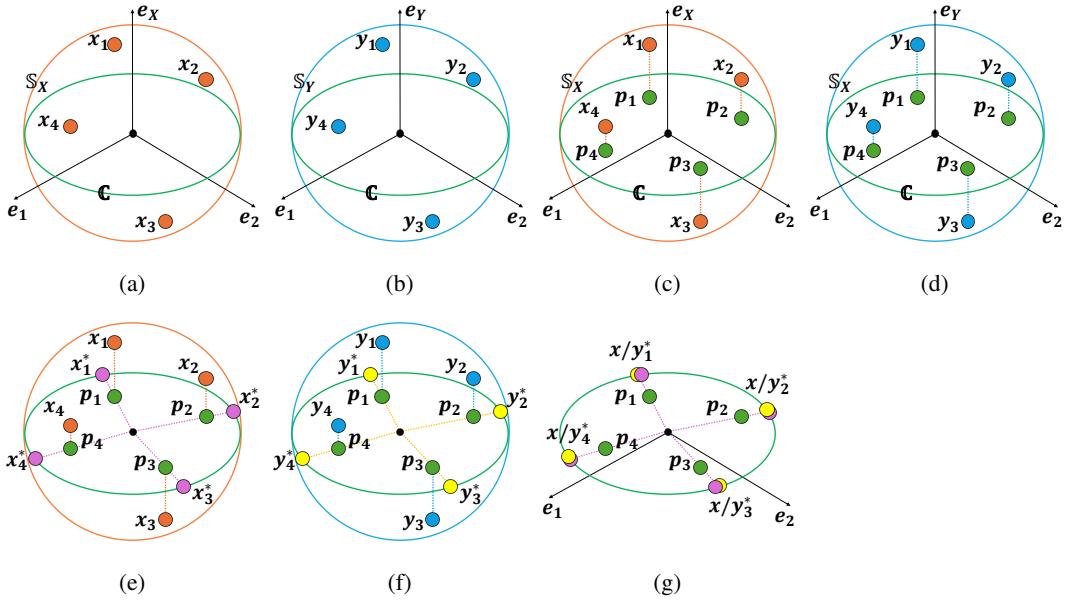


Figure 6: Modality Alignment in 4D Space. \mathbb{S}_X (orange circle) and \mathbb{S}_Y (blue circle) are two 3D unit spheres within \mathbb{S}^3 as described in Eq. (16) and Eq. (22). The shared space \mathbb{C} is a 2D plane as described in Eq. (17). $\mathbb{S}_X \cap \mathbb{C} = \mathbb{S}_Y \cap \mathbb{C}$ is a 2D circle (green circle). (a) 4 samples from X (orange dots). (b) 4 samples from Y (blue dots). (c), (d): The projections of $(x_i, y_i)_{i \neq c}$ on the shared space \mathbb{C} converge to p_i (green point), i.e., $P_C x_i = P_C y_i = p_i$. (e), (f): Re-normalize p_i to get x_i^* (purple dots) and y_i^* (yellow dots) as described in Eq. (23). (g): (x_i^*, y_i^*) are perfectly aligned.

$$\mathbb{C} = \text{span} \{e_1\} \oplus \text{span} \{e_2\}. \quad (17)$$

Therefore, \mathbb{C}^\perp is a 2-dimensional orthogonal complement of \mathbb{C} , and \mathbb{C}^\perp satisfies:

$$\begin{aligned} \mathbb{C}^\perp &= \text{span} \{e_3\} \oplus \text{span} \{e_4\}, \\ \mathbb{R}^4 &= \mathbb{C} \oplus \mathbb{C}^\perp. \end{aligned} \quad (18)$$

Define two unit vectors e_X and e_Y such that:

$$\begin{aligned} e_X &\in \mathbb{S}_X, \text{ and } e_X \perp \mathbb{C}, \\ e_Y &\in \mathbb{S}_Y, \text{ and } e_Y \perp \mathbb{C}. \end{aligned} \quad (19)$$

Since $n_A, n_B \in \mathbb{C}^\perp$, $n_A \perp e_X$ and $n_B \perp e_Y$, we have:

$$\langle e_X, e_Y \rangle = \pm \langle n_A, n_B \rangle, \quad (20)$$

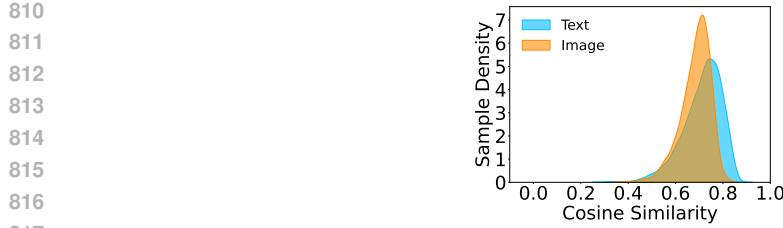
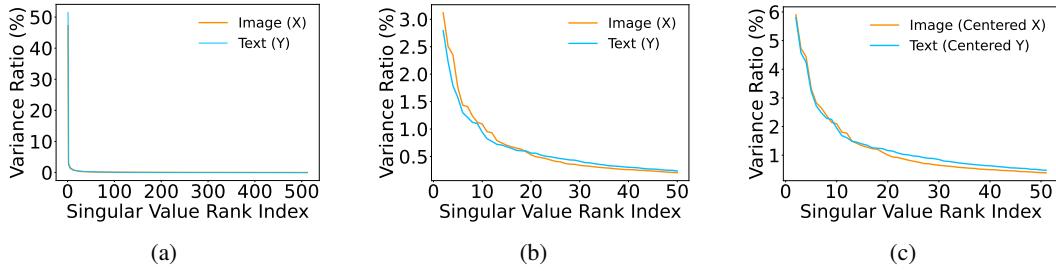
and we choose a pair of e_X and e_Y such that:

$$\langle e_X, e_Y \rangle = \langle n_A, n_B \rangle = \cos(\phi) \in (0, 1). \quad (21)$$

Therefore, \mathbb{S}_X and \mathbb{S}_Y can be represented by two orthonormal bases:

$$\begin{aligned} \mathbb{S}_X &\in \mathbb{A} = \text{span} \{e_1\} \oplus \text{span} \{e_2\} \oplus \text{span} \{e_X\}, \\ \mathbb{S}_Y &\in \mathbb{B} = \text{span} \{e_1\} \oplus \text{span} \{e_2\} \oplus \text{span} \{e_Y\}. \end{aligned} \quad (22)$$

In Theorem 3, we show that when $\mathcal{L}_{\text{MCL}}^c$ is minimized, $c_x, c_y \perp \mathbb{C}$ (Condition (A6)). Accordingly, we can set $c_x = e_X$ and $c_y = e_Y$. These settings are illustrated in Fig. 6a and Fig. 6b.

818 Figure 7: Density plot of θ_i^c of CLIP ViT-B/32 embeddings of MS-COCO validation set.
819828 Figure 8: Singular values. CLIP ViT-B/32 embeddings of MS-COCO validation set are used. **(a)**: All
829 singular values σ_i of X and Y . **(b)**: The 2nd to the 50th σ_i of X and Y . **(c)**: The 1st to the 50th σ_i
830 of the centered X and the centered Y .
831832
833
834 In Theorem 4, we show that when $\mathcal{L}_{\text{MCL}}^{i \neq c}$ is minimized, $P_C x_i = P_C y_i$ (Condition (A8)). This
835 condition is illustrated in Fig. 6c and Fig. 6d.
836837 Re-normalize the projections to obtain the transformed pairs:
838

$$x_i^* = \frac{P_C x_i}{\|P_C x_i\|}, \quad (23)$$

$$y_i^* = \frac{P_C y_i}{\|P_C y_i\|},$$

845 We illustrate x_i^* and y_i^* in Fig. 6e and Fig. 6f. In Corollary 4, we show that (x_i^*, y_i^*) are perfectly
846 aligned, as illustrated in Fig. 6g.
847848

B.3 JUSTIFICATION OF ASSUMPTION IN THEOREM 4

849850 In Theorem 4, we assume that the angle between a modality input and its center, θ_i^c , satisfies
851 $\theta_i^c \in (0, \frac{\pi}{2})$. In Lemma 15, we provide a theoretical justification for this assumption. Furthermore,
852 the density plot of θ_i^c in Fig. 7 shows that almost all θ_i^c indeed lie within $(0, \frac{\pi}{2})$.
853854

C APPENDIX C: DETAILS OF METHOD

856857 In this subsection, we describe in details about how to detect dimension collapse, how to detect the
858 shared space of two subspaces, and how to conduct projection onto the shared space.
859860

C.1 DETECT DIMENSION COLLAPSE

862863 Suppose we have two point clouds, X and Y , each consisting of h -dimensional normalized vectors:
864 $X = (x_1, \dots, x_N) \in (\mathbb{S}^{h-1})^N$ and $Y = (y_1, \dots, y_N) \in (\mathbb{S}^{h-1})^N$. Then we have:

$$\begin{aligned}
864 \quad & \mathbb{A} = \text{span}(X), \quad d_X = \dim(\mathbb{A}), \\
865 \quad & \mathbb{B} = \text{span}(Y), \quad d_Y = \dim(\mathbb{B}), \\
866 \quad & \mathbb{C} = \mathbb{A} \cap \mathbb{B}, \quad d_{\text{overlap}} = \dim(\mathbb{C}). \\
867 \\
868
\end{aligned} \tag{24}$$

869 Apply the Singular Value Decomposition (SVD) to X and Y and we get:

$$\begin{aligned}
871 \quad & X = U_X \Sigma_X V_X^\top, \\
872 \quad & Y = U_Y \Sigma_Y V_Y^\top. \\
873
\end{aligned} \tag{25}$$

874 If X and Y collapse into subspaces of \mathbb{S}^{h-1} , then Σ_X and Σ_Y have $d_X < h$ and $d_Y < h$ significant
875 singular values, respectively.
876

877 In the discussion in Sec. 3.4, X and Y represent the image and text embeddings of the MS-COCO
878 dataset. Since X and Y are not centered at zero, the first singular values, σ_1^x and σ_1^y , dominate
879 when SVD is applied. Correspondingly, the first right singular vectors of X and Y are c_x and c_y ,
880 respectively. As shown in Fig. 8a, these first right singular vectors account for approximately 50%
881 of the explained variance. Therefore, in Fig. 2c, we plot the singular values of the centered X and
882 Y , which better capture the patterns of variation. In Fig. 8b, we present the 2nd to the 50th singular
883 values of X and Y , while in Fig. 8c, we show the 1st to the 50th singular values of the centered X
884 and the centered Y .

885 And dimension collapse in X and Y occurs when zero values appear on the diagonals of Σ_X and Σ_Y .
886

887 C.2 FIND THE SHARED SPACE

888 We then select the first d_X columns from V_X and the first d_Y columns from V_Y whose cumulative
889 explained variance exceeds a predefined threshold c (e.g., $c = 99\%$). We obtain:
890

$$\begin{aligned}
891 \quad & B_X = V_X[:, :d_X] \in \mathbb{R}^{h \times d_X} : \text{orthonormal basis for } \mathbb{A}, \\
892 \quad & B_Y = V_Y[:, :d_Y] \in \mathbb{R}^{h \times d_Y} : \text{orthonormal basis for } \mathbb{B}. \\
893
\end{aligned} \tag{26}$$

894 To investigate whether \mathbb{A} and \mathbb{B} have overlap dimensions, we need to check the principal angles
895 between \mathbb{A} and \mathbb{B} , which are defined as:
896

897 **Definition 9.** *The principal angles $\gamma_1 \leq \gamma_2 \leq \dots \leq \gamma_k$ between \mathbb{A} and \mathbb{B} are recursively defined as:*
898

$$\cos(\gamma_i) = \max_{u \in \mathbb{A}, v \in \mathbb{B}} u^\top v, \quad \|u\| = \|v\| = 1, \quad u^\top u_j = v^\top v_j = 0 \quad (j < i), \tag{27}$$

901 where $k = \min(d_X, d_Y)$.
902

903 The principal angles quantify the alignment between these subspaces:
904

- 905 • The smallest principal angle θ_1 measures how close the two subspaces are: if $\gamma_1 = 0$, there is at
906 least one common direction.
- 907 • If multiple principal angles are zero, then the intersection of the subspaces has a larger dimension.
908

909 The principal angles between subspaces \mathbb{A} and \mathbb{B} can be computed as follows:
910

- 911 1. Compute the singular values of the matrix $G = B_X^\top B_Y \in \mathbb{R}^{d_X \times d_Y}$.
- 912 2. The singular values $\sigma_i^p \in [0, 1]$
- 913 3. Then the principal angles are $\gamma_i = \arccos(\sigma_i^p)$

915 The number of principal angles equal to zero gives the dimension of the intersection:
916

$$917 \quad d_{\text{overlap}} = \#\{i : \gamma_i = 0\}. \tag{28}$$

918 In practice, due to noise or finite precision, we use a threshold: count how many $\sigma_i^p > 1 - \epsilon$ (e.g.,
 919 $\epsilon = 10^{-3}$). Thus:

$$921 \quad d_{\text{overlap}} = \#\{i : \sigma_i^p > 1 - \epsilon\}. \quad (29)$$

923 The empirical result of MS-COCO dataset is provided in Fig. 2d.

925 C.3 PROCEDURES OF SSP METHOD

927 In this subsection, we provide the details of the Shared Space Projection (SSP) algorithm.

929 **Step 1:** Apply the SVD decomposition to X and Y to get V_X and V_Y as Eq. (25).

930 **Step 2:** Select the first d_X and d_Y right singular vectors of X and Y whose cumulative explained
 931 variance are great than 99%. The resulting vectors, B_X and B_Y , form the bases for \mathbb{A} and \mathbb{B} , as
 932 indicated by Eq. (26).

933 **Step 3:** Apply the SVD decomposition $G = B_X^\top B_Y \in \mathbb{R}^{d_X \times d_Y}$.

$$935 \quad G = U_G \Sigma_G V_G^\top \quad (30)$$

937 **Step 4:** Compute d_{overlap} according to Eq. (29) while setting $\epsilon = 10^{-3}$. Compute the basis of the
 938 shared space B_S by:

$$941 \quad B_S = B_X U_G[:, :d_{\text{overlap}}] = B_Y V_G[:, :d_{\text{overlap}}]. \quad (31)$$

943 **Explain:** Since the shared space is estimated from the available data rather than the original training
 944 data (assumed inaccessible), the estimation may be noisy. To mitigate this, we can select $k < d_{\text{overlap}}$
 945 columns from B_S to form B_S^k . The columns of B_S^k constitute an orthonormal basis for a k -dimensional
 946 subspace of the estimated shared space. By removing dimensions that carry minimal information,
 947 the estimation error can be reduced. The following optional step explains how to select these k
 948 dimensions.

949 **Step 5 (Optional):** Project X and Y onto each column of B_S :

$$951 \quad P = B_S^T X^T, \\ 952 \quad X' = \text{einsum}('hk, kn- > knh')(B_S, P), \\ 953 \quad X'' = \text{Normalize}(X') \text{ by the last dimension.} \quad (32)$$

955 Here, einsum denotes Einstein summation notation. Compute the variance of X'' along the last two
 956 dimensions to obtain an array S of length d_{overlap} . Each entry of S is actually the singular value of
 957 projections onto the corresponding column of B_S . S quantifies the amount of information contained
 958 in each column of B_S . By ranking S in descending order, select the top k columns from B_S to form
 959 B_S^K .

960 **Step 6:** Project X and Y onto the column space of B_S^K and get X^* and Y^* .

$$963 \quad X^* = \left(B_S^K B_S^{K^T} X^T \right)^T, \\ 964 \quad Y^* = \left(B_S^K B_S^{K^T} Y^T \right)^T. \quad (33)$$

967 **Step 7:** Normalize X^* and Y^* to get X^{**} and Y^{**} . Use X^{**} and Y^{**} for downstream tasks.

969 Notably, Fig. 8b indicates that fewer than 10 dimensions account for more than 1% of the explained
 970 variance, suggesting that the essential information of X and Y can be effectively captured using only
 971 10 dimensions. Consequently, in Fig. 4b, we project X and Y onto a 10-dimensional subspace that
 preserves the most information.

Table 2: Size of θ_Δ and accuracies (%) of zero-shot image classification of ViT-L/14 on various dataset.

Model	CIFAR-10			CIFAR-100			ImageNet-1K		
	Δ_θ	R1	R5	Δ_θ	R1	R5	Δ_θ	R1	R5
CLIP	77.63°	95.12	99.46	74.19°	65.23	88.88	77.29°	75.56	94.58
CLIP + Translation	14.73°	92.39	98.97	30.50°	54.46	77.25	62.61°	74.05	94.10
CLIP + Removal	79.36°	12.23	62.33	73.16°	16.82	6.44	76.84°	67.04	89.76
CLIP + SSP	13.27°	95.12	99.46	30.39°	64.51	88.79	62.40°	75.26	94.51

Table 3: Size of θ_{Δ} and accuracies (%) of zero-shot cross-modal retrieval of ViT-L/14 on MSCOCO.

Model	MSCOCO						
	Δ_θ	I \rightarrow T			T \rightarrow I		
		R@1	R@5	R@10	R@1	R@5	R@10
CLIP	78.16°	56.06	79.56	86.84	35.33	59.96	70.21
CLIP + Translation	68.49°	54.14	78.32	86.30	35.13	59.79	69.85
CLIP + Removal	76.03°	49.56	73.42	82.18	31.23	54.29	65.00
CLIP + SSP	68.06°	55.54	78.94	86.64	35.22	59.86	70.22

D APPENDIX D: DETAILS OF EXPERIMENTS

In this section, we describe in details about the set up of our experiments.

D.1 ZERO-SHOT IMAGE CLASSIFICATION.

Datasets. We first evaluate our method on the zero-shot image classification task using three widely adopted datasets: two small-scale image dataset **CIFAR-10/100** Krizhevsky et al. (2009) and one large scale image dataset **ImageNet-1k** Deng et al. (2009). For CIFAR-10/100, we adopt the small set of prompts provided by OpenAI for CLIP Radford et al. (2021) (<https://github.com/openai/CLIP.com>). For ImageNet-1k, we adopt the large set of prompts provided by OpenAI for CLIP Radford et al. (2021) (https://colab.research.google.com/github/openai/CLIP/blob/main/notebooks/Prompt_Engineering_for_ImageNet.ipynb).

Implementation Setup. Our implementation refers to Eslami & de Melo (2025). For model backbone, we adopt CLIP’s ViT-B/32 ViT-L/14 models. For the implementation of baseline models, we remove the same number of dimensions in the removal method Schröd et al. (2025) with that of our SSP method. For translation Liang et al. (2022), the hyperparameter λ controls the scale of translation. We choose the smallest value of λ , rounded to two decimal places, that yields an angle reduction larger than SSP.

Additional Results. We report the results using the CLIP ViT-L/14 model as the backbone in Tab. 2. Similar patterns to those in Tab. 1 can be observed, indicating that our conclusions hold across different model backbones.

As shown in both Tab. 1 and Tab. 2, reducing the modality gap becomes more challenging as the number of classes in the test set increases. This is because a larger number of classes introduces a more complex data distribution, thereby enlarging the discrepancy between the test and training distributions. Consequently, our shared space estimation incurs greater estimation error, which limits the capacity of our method to further reduce the modality gap.

D.2 ZERO-SHOT CROSS-MODAL RETRIEVAL

Datasets. In addition to zero-shot image classification, we evaluate our method on zero-shot image-to-text and text-to-image retrieval using the MSCOCO (Lin et al., 2014) and Flickr30K (Plummer

1026 Table 4: Size of θ_Δ and accuracies (%) of zero-shot cross-modal retrieval of ViT-L/14 on Flickr30K.
1027

Model	Δ_θ	Flickr30K					
		$\mathbf{I} \rightarrow \mathbf{T}$			$\mathbf{T} \rightarrow \mathbf{I}$		
		R@1	R@5	R@10	R@1	R@5	R@10
CLIP	78.16°	56.06	79.56	86.84	35.33	59.96	70.21
CLIP + Translation	-	-	-	-	-	-	-
CLIP + Removal	-	-	-	-	-	-	-
CLIP + SSP	-	-	-	-	-	-	-

1038 et al., 2015) datasets. Unlike the common practice of appending a prompt such as ‘a photo of the
1039 caption’, we directly use the raw captions to generate text embeddings. This approach aims to align
1040 the text space more closely with its natural form rather than introducing distortion through artificial
1041 prompts.

1042 **Implementation Setup.** This implementation setup follows Sec. D.1. The only difference is that we
1043 only use CLIP ViT-L/14 as the model backbone.

1045 D.3 RESULTS

1047 The goal of this experiment is to reduce the size of the modality gap as much as possible without
1048 harming downstream performance. In Tab. 3 and Tab. 4, we list results of the size of the modality gap
1049 (Δ_θ), the top-1 accuracy (R@1), the top-5 accuracy (R@5), and the top-10 accuracy (R@10). Similar
1050 patterns to those in Tab. 1 can be observed, indicating that our conclusions hold across different
1051 downstream tasks.

1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079

1080 E APPENDIX E: PROOFS
10811082 E.1 DETAILS OF THEOREM 1
1083

1084 In this section, we provide proofs of Theorem 1 that is proposed in Sec. 3.2. We also provide details
1085 of the auxiliary theorems (Theorem S1 and Theorem S2) and technical lemmas (Lemma 1, Lemma 2,
1086 Lemma 3, Lemma 4) that support the proof of Theorem 1. For convenience in reading, let us recall
1087 some related notions and definitions.

- 1088 • $h, N \in \mathbb{N}$.
- 1089 • $\mathbb{S}^{h-1} = \{z \in \mathbb{R}^h : \|z\| = 1\}$.
- 1090 • σ_{h-1} : the uniform probability measure of \mathbb{S}^{h-1} .

1091 **Definition** (Multimodal Contrastive Loss (MCL Loss)). Let (X, Y) be an N -pair configuration,
1092 where $X = (x_1, \dots, x_N) \in (\mathbb{S}^{h-1})^N$ and $Y = (y_1, \dots, y_N) \in (\mathbb{S}^{h-1})^N$. $\forall \tau > 0$, the multimodal
1093 contrastive loss $\mathcal{L}_{\text{MCL}}(\cdot, \cdot) : (\mathbb{S}^{h-1})^N \times (\mathbb{S}^{h-1})^N \rightarrow \mathbb{R}$ is defined as:

$$1094 \mathcal{L}_{\text{MCL}} = \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{\text{MCL}}^i, \text{ where } \mathcal{L}_{\text{MCL}}^i = \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) + \mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}(y_i; X).$$

1100 Here, $\mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}$ is the \mathcal{X} -to- \mathcal{Y} alignment and $\mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}$ is the \mathcal{Y} -to- \mathcal{X} alignment, which are defined
1101 respectively as:

$$1102 \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) = -\log \frac{\exp(x_i \cdot y_i / \tau)}{\sum_{j=1}^N \exp(x_i \cdot y_j / \tau)}, \quad \mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}(y_i; X) = -\log \frac{\exp(x_i \cdot y_i / \tau)}{\sum_{j=1}^N \exp(x_j \cdot y_i / \tau)}.$$

1103 E.1.1 PROOF OF THEOREM 1
1104

1105 In this subsection, we provide the proof of Theorem 1. For convenience in reading, we first restate
1106 Theorem 1 here.

1107 **Theorem 1.** [Restate] Let (X, Y) be an N -pair configuration, where $X = (x_1, \dots, x_N) \in (\mathbb{S}^{h-1})^N$
1108 are iid samples from μ_x and $Y = (y_1, \dots, y_N) \in (\mathbb{S}^{h-1})^N$ are iid samples from μ_y . Let $\nu =$
1109 $h/2 - 1$, it holds that:

$$1110 \lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}} - 2 \log(N) = \mathbb{E}_{x_i \sim \mu_x} \left[-\frac{x_i \cdot y_i}{\tau} \right] + \mathbb{E}_{x_i \sim \mu_x} \left[\log \mathbb{E}_{y_i \sim \mu_y} \left[\exp \left(\frac{x_i \cdot y_i}{\tau} \right) \right] \right] \\ 1111 + \mathbb{E}_{y_i \sim \mu_y} \left[-\frac{x_i \cdot y_i}{\tau} \right] + \mathbb{E}_{y_i \sim \mu_y} \left[\log \mathbb{E}_{x_j \sim \mu_x} \left[\exp \left(\frac{x_i \cdot y_i}{\tau} \right) \right] \right] \\ 1112 \geq -\frac{2}{\tau} + 2 \log \left(\Gamma(\nu + 1) (2\tau)^\nu I_\nu \left(\frac{1}{\tau} \right) \right)$$

1113 where equality is attained if and only if there exists a configuration of (X, Y) such that:

1114 (A1) $\forall i \in [N], x_i = y_i$.
1115 (A2) $\mu_x = \sigma_{h-1}$ and $\mu_y = \sigma_{h-1}$.

1116 *Proof.* We first decompose $\lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}}^c - 2 \log(N)$ into two parts:

$$1117 \lim_{N \rightarrow \infty} (\mathcal{L}_{\text{MCL}} - 2 \log(N)) = \lim_{N \rightarrow \infty} \left(\frac{1}{N} \sum_{i=1}^N \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) - \log(N) \right) \\ 1118 + \lim_{N \rightarrow \infty} \left(\frac{1}{N} \sum_{i=1}^N \mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}(y_i; X) - \log(N) \right). \quad (34)$$

1134 According to Theorem S2, the convergent function and its lower bound of $\mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}$ are:
 1135

$$\begin{aligned}
 1136 \quad & \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) - \log(N) \\
 1137 \quad &= \mathbb{E}_{x_i \sim \mu_x} \left[-\frac{x_i \cdot y_i}{\tau} \right] + \mathbb{E}_{x_i \sim \mu_x} \left[\log \mathbb{E}_{y_i \sim \mu_y} \left[\exp \left(\frac{x_i \cdot y_i}{\tau} \right) \right] \right] \\
 1138 \quad &\geq -\frac{1}{\tau} + \log \left[\Gamma \left(\frac{h}{2} \right) (2\tau)^{\frac{h}{2}-1} I_{\frac{h}{2}-1} \left(\frac{1}{\tau} \right) \right]
 \end{aligned} \tag{35}$$

1144 where equality is attained if and only if there exists a configuration of (X, Y) such that:
 1145

- 1146 (i) $\forall i \in [N], x_i = y_i$.
- 1147 (ii) $\mu_x = \sigma_{h-1}$ and $\mu_y = \sigma_{h-1}$.

1150 This Theorem also holds for $\mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}$:

$$\begin{aligned}
 1153 \quad & \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}(y_i; X) - \log(N) \\
 1154 \quad &= \mathbb{E}_{y_i \sim \mu_x} \left[-\frac{x_i \cdot y_i}{\tau} \right] + \mathbb{E}_{y_i \sim \mu_y} \left[\log \mathbb{E}_{x_i \sim \mu_x} \left[\exp \left(\frac{x_i \cdot y_i}{\tau} \right) \right] \right] \\
 1155 \quad &\geq -\frac{1}{\tau} + \log \left[\Gamma \left(\frac{h}{2} \right) (2\tau)^{\frac{h}{2}-1} I_{\frac{h}{2}-1} \left(\frac{1}{\tau} \right) \right]
 \end{aligned} \tag{36}$$

1160 where equality is attained if and only if there exists a configuration of (X, Y) such that:
 1161

- 1162 (iii) $\forall i \in [N], x_i = y_i$.
- 1163 (iv) $\mu_x = \sigma_{h-1}$ and $\mu_y = \sigma_{h-1}$.

1166 Combining Eq. (34), Eq. (35) and Eq. (36), we conclude that:

$$\begin{aligned}
 1170 \quad & \lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}} - 2 \log(N) = \mathbb{E}_{x_i \sim \mu_x} \left[-\frac{x_i \cdot y_i}{\tau} \right] + \mathbb{E}_{x_i \sim \mu_x} \left[\log \mathbb{E}_{y_i \sim \mu_y} \left[\exp \left(\frac{x_i \cdot y_i}{\tau} \right) \right] \right] \\
 1171 \quad &+ \mathbb{E}_{y_i \sim \mu_y} \left[-\frac{x_i \cdot y_i}{\tau} \right] + \mathbb{E}_{y_i \sim \mu_y} \left[\log \mathbb{E}_{x_i \sim \mu_x} \left[\exp \left(\frac{x_i \cdot y_i}{\tau} \right) \right] \right] \\
 1172 \quad &\geq -\frac{2}{\tau} + 2 \log \left[\Gamma \left(\frac{h}{2} \right) (2\tau)^{\frac{h}{2}-1} I_{\frac{h}{2}-1} \left(\frac{1}{\tau} \right) \right]
 \end{aligned} \tag{37}$$

1176 where equality is attained if and only if the following conditions hold:
 1177

- 1178 (A1) $\forall i \in [N], x_i = y_i$.
- 1179 (A2) $\mu_x = \sigma_{h-1}$ and $\mu_y = \sigma_{h-1}$.

1182 \square

1185 E.1.2 AUXILIARY THEOREMS PART 1

1186 In this subsection, we provide details and proofs of the auxiliary theorems (Theorem S1 and Theorem S2) that support the proof of Theorem 1.

1188
 1189 **Theorem S1.** Let (X, Y) be an N -pair configuration, where $X = (x_1, \dots, x_N) \in (\mathbb{S}^{h-1})^N$ are iid
 1190 samples from μ_x and $Y = (y_1, \dots, y_N) \in (\mathbb{S}^{h-1})^N$ are iid samples from μ_y . It holds that:

$$\begin{aligned} 1192 \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) - \log(N) &= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N -\log \frac{\exp(x_i \cdot y_i / \tau)}{\sum_{j=1}^N \exp(x_i \cdot y_j / \tau)} - \log(N) \\ 1193 &= \mathbb{E}_{x_i \cdot y_i} \left[-\frac{x_i \cdot y_i}{\tau} \right] + \mathbb{E}_{x_i \sim \mu_x} \left[\log \mathbb{E}_{y_i \sim \mu_y} \left[\exp \left(\frac{x_i \cdot y_i}{\tau} \right) \right] \right] \end{aligned} \quad (38)$$

1198 *Proof.* $\forall x_i \in X$, the \mathcal{X} -to- \mathcal{Y} alignment of x_i can be rewritten as:

$$\begin{aligned} 1200 \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) &= -\log \frac{\exp(x_i \cdot y_i / \tau)}{\sum_j \exp(x_i \cdot y_j / \tau)} \\ 1201 &= -\frac{x_i \cdot y_i}{\tau} + \log \left(N \frac{1}{N} \sum_{j=1}^N \exp \left(\frac{x_i \cdot y_j}{\tau} \right) \right) \\ 1202 &= -\frac{x_i \cdot y_i}{\tau} + \log \left(\frac{1}{N} \sum_{j=1}^N \exp \left(\frac{x_i \cdot y_j}{\tau} \right) \right) + \log(N). \end{aligned} \quad (39)$$

1210 Denote $h_N(x)$ and $h(x)$ as:

$$\begin{aligned} 1212 h_N(x) &= \log \left(\frac{1}{N} \sum_{j=1}^N \exp \left(\frac{x \cdot y_j}{\tau} \right) \right), \\ 1213 \text{and } h(x) &= \log \left(\mathbb{E}_{y \sim \mu_y} \left[\exp \left(\frac{x \cdot y}{\tau} \right) \right] \right). \end{aligned} \quad (40)$$

1217 Lemma 2 reveals that $h_N(x)$ uniformly converges to $h(x)$ almost surely. Thus, we have:

$$\sup_{x \in \mathbb{S}^{h-1}} |h_N(x) - h(x)| \xrightarrow[N \rightarrow \infty]{\text{a.s.}} 0. \quad (41)$$

1222 According to the Strong Law of Large Numbers (SLLN), we have:

$$\frac{1}{N} \sum_{i=1}^N h(x_i) \xrightarrow[N \rightarrow \infty]{\text{a.s.}} \mathbb{E}_{x \sim \mu_x}[h(x)]. \quad (42)$$

1228 Combining Eq. (41) and Eq. (42), we get:

$$\begin{aligned} 1230 \frac{1}{N} \sum_{i=1}^N h_N(x_i) &= \frac{1}{N} \sum_{i=1}^N h(x_i) + \frac{1}{N} \sum_{i=1}^N (h_N(x_i) - h(x_i)) \\ 1231 &\xrightarrow[N \rightarrow \infty]{\text{a.s.}} \mathbb{E}_{x \sim \mu_x}[h(x)]. \end{aligned} \quad (43)$$

1235 Similarly, by the Strong Law of Large Numbers (SLLN), we have:

$$\frac{1}{N} \sum_{i=1}^N -\frac{x_i \cdot y_i}{\tau} \xrightarrow[N \rightarrow \infty]{\text{a.s.}} \mathbb{E}_{x_i \sim \mu_x} \left[-\frac{x_i \cdot y_i}{\tau} \right]. \quad (44)$$

1240 Putting Eq. (39), Eq. (43) and Eq. (44) together, the convergent function of $\frac{1}{N} \sum_{i=1}^N \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y)$
 1241 can be derived as:

$$\begin{aligned}
& \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) - \log(N) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \left(-\frac{x_i \cdot y_i}{\tau} + h_N(x_i) \right) \\
& = \mathbb{E}_{x_i \cdot y_i} \left[-\frac{x_i \cdot y_i}{\tau} \right] + \mathbb{E}_{x_i \sim \mu_x} [h(x_i)] \\
& = \mathbb{E}_{x_i \cdot y_i} \left[-\frac{x_i \cdot y_i}{\tau} \right] + \mathbb{E}_{x_i \sim \mu_x} \left[\log \mathbb{E}_{y_j \sim \mu_y} \left[\exp \left(\frac{x_i \cdot y_j}{\tau} \right) \right] \right].
\end{aligned} \tag{45}$$

□

Theorem S2. Let (X, Y) be an N -pair configuration, where $X = (x_1, \dots, x_N) \in (\mathbb{S}^{h-1})^N$ are iid samples from μ_x and $Y = (y_1, \dots, y_N) \in (\mathbb{S}^{h-1})^N$ are iid samples from μ_y . Let $\nu = h/2 - 1$, it holds that:

$$\begin{aligned}
& \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) - \log(N) \\
& = \mathbb{E}_{x_i \sim \mu_x} \left[-\frac{x_i \cdot y_i}{\tau} \right] + \mathbb{E}_{x_i \sim \mu_x} \left[\log \mathbb{E}_{y_i \sim \mu_y} \left[\exp \left(\frac{x_i \cdot y_i}{\tau} \right) \right] \right] \\
& \geq \log \left(\Gamma(\nu + 1) (2\tau)^\nu I_\nu \left(\frac{1}{\tau} \right) \right)
\end{aligned} \tag{46}$$

where equality is attained if and only if the following conditions hold:

- (B1) $\forall i \in [N], x_i = y_i$.
- (B2) $\mu_x = \sigma_{h-1}$ and $\mu_y = \sigma_{h-1}$.

Proof. **Step 1:** We start the proof by find the convergent function of $\frac{1}{N} \sum_{i=1}^N \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y)$ as $N \rightarrow \infty$. $\forall x_i \in X$, as prove in Theorem S1:

$$\begin{aligned}
& \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) - \log(N) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N -\log \frac{\exp(x_i \cdot y_i / \tau)}{\sum_{j=1}^N \exp(x_i \cdot y_j / \tau)} - \log(N) \\
& = \mathbb{E}_{x_i \cdot y_i} \left[-\frac{x_i \cdot y_i}{\tau} \right] + \mathbb{E}_{x_i \sim \mu_x} \left[\log \mathbb{E}_{y_i \sim \mu_y} \left[\exp \left(\frac{x_i \cdot y_i}{\tau} \right) \right] \right].
\end{aligned} \tag{47}$$

Step 2: Next, we find the minimal value and the optimal condition of convergent function.

According to the Cauchy-Schwarz inequality, the first term in Eq. (47) can be bounded below:

$$\mathbb{E}_{x_i \cdot y_i} \left[-\frac{x_i \cdot y_i}{\tau} \right] \geq \mathbb{E}_{x_i \cdot y_i} \left[-\frac{\|x_i\| \|y_i\|}{\tau} \right] = -\frac{1}{\tau}. \tag{48}$$

where equality is attained if and only if there exists a configuration of (X, Y) such that :

- (B1) $\forall i \in [N], x_i = y_i$.

Note that condition (B1) implies $\mu_x = \mu_y$. Applying this condition to the second term in Eq. (47), we can transform it as:

$$\mathbb{E}_{x \sim \mu_x} \left[\log \left(\mathbb{E}_{y \sim \mu_y} \left[\exp \left(\frac{x \cdot y}{\tau} \right) \right] \right) \right] = \mathbb{E}_{x \sim \mu} \left[\log \left(\mathbb{E}_{y \sim \mu} \left[\exp \left(\frac{x \cdot y}{\tau} \right) \right] \right) \right]. \tag{49}$$

Let $\mathbb{M}(\mathbb{S}^{h-1})$ be the set of Borel probability measures in \mathbb{S}^{h-1} . The RHS of Eq. (49) is then a functional $\mathcal{F}[\cdot] : \mathbb{M}(\mathbb{S}^{h-1}) \rightarrow \mathbb{R}$:

1296

$$\mathcal{F}[\mu] = \mathbb{E}_{x \sim \mu} \left[\log \left(\mathbb{E}_{y \sim \mu} \left[\exp \left(\frac{x \cdot y}{\tau} \right) \right] \right) \right]. \quad (50)$$

1299 According to Lemma 3, $\mathcal{F}[\mu]$ is minimized when $\mu = \sigma_{h-1}$ where σ_{h-1} is the uniform measure of
 1300 \mathbb{S}^{h-1} .
 1301

1302

$$\sigma_{h-1} = \arg \min_{\mu \in \mathbb{M}(\mathbb{S}^{h-1})} \mathcal{F}[\mu]. \quad (51)$$

1304

1305 Therefore, we have:

1306

$$\mathcal{F}[\mu] \geq \mathcal{F}[\sigma_{h-1}]. \quad (52)$$

1308

1309 where equality is attained if and only if there exists a configuration of (X, Y) such that :

1310

$$1311 \quad (B2) \quad \mu_x = \mu_y = \sigma_{h-1}$$

1312

1313 Let $\Gamma(\cdot)$ be the Gamma function, Lemma 4 derives that:

1314

$$\begin{aligned} 1315 \quad \mathcal{F}[\sigma_{h-1}] &= \mathbb{E}_{x \sim \sigma_{h-1}} \left[\mathbb{E}_{y \sim \sigma_{h-1}} \left[\exp \left(\frac{x \cdot y}{\tau} \right) \right] \right] \\ 1316 \\ 1317 &= \log \left[\Gamma \left(\frac{h}{2} \right) (2\tau)^{\frac{h}{2}-1} I_{\frac{h}{2}-1} \left(\frac{1}{\tau} \right) \right] \end{aligned} \quad (53)$$

1319

1320 Combining Eq. (47), Eq. (48), Eq. (49), Eq. (53), we conclude that:

1321

$$\begin{aligned} 1322 \quad \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) - \log(N) \\ 1323 \\ 1324 &= \mathbb{E}_{x_i \sim \mu_x} \left[-\frac{x_i \cdot y_i}{\tau} \right] + \mathbb{E}_{x_i \sim \mu_x} \left[\log \mathbb{E}_{y_i \sim \mu_y} \left[\exp \left(\frac{x_i \cdot y_i}{\tau} \right) \right] \right] \\ 1325 \\ 1326 &\geq -\frac{1}{\tau} + \log \left[\Gamma \left(\frac{h}{2} \right) (2\tau)^{\frac{h}{2}-1} I_{\frac{h}{2}-1} \left(\frac{1}{\tau} \right) \right] \end{aligned} \quad (54)$$

1327

1328 where equality is attained if and only if the following conditions hold:

1329

$$1330 \quad (B1) \quad \forall i \in [N], x_i = y_i.$$

1331

$$1332 \quad (B2) \quad \mu_x = \sigma_{d-1} \text{ and } \mu_y = \sigma_{d-1}.$$

1333

1334

1335

1336

1337

1338

1339

1340

1341

1342

1343

1344

1345

1346

1347

1348

1349

□

1350 E.1.3 TECHNICAL LEMMAS PART 1
13511352 In this section, we provide details and proofs of the technical lemmas (technical lemmas 1,
1353 Lemma 2, Lemma 3, Lemma 4) that support the proof of Theorem 1, Theorem S1 and Theorem S2.1354 **Lemma 1.** Let $x \in \mathbb{S}^{h-1}$ and Y be an N -point configuration, where $Y = (y_1, \dots, y_N) \in (\mathbb{S}^{h-1})^N$
1355 are iid samples from μ_y . $\forall \tau > 0$, define a sequence of functions $\{g_N\} : \mathbb{S}^{h-1} \rightarrow \mathbb{R}$ as:

1356

1357
$$g_N(x) = \frac{1}{N} \sum_{j=1}^N \exp\left(\frac{x \cdot y_j}{\tau}\right). \quad (55)$$

1358
1359

1360 Define a function $g : \mathbb{S}^{h-1} \rightarrow \mathbb{R}$ as:

1361

1362
$$g(x) = \mathbb{E}_{y \sim \mu_y} \left[\exp\left(\frac{x \cdot y}{\tau}\right) \right]. \quad (56)$$

1363
1364

1365 It holds that $\{g_N\}$ converges uniformly to g :

1366

1367
$$g_N(x) \xrightarrow[N \rightarrow \infty]{\text{unif.}} g(x). \quad (57)$$

1368
1369

1370 *Proof. Step 1 Boundedness and Lipschitz Property:*

1371

1372 Consider a function class $\mathcal{F} = \{f_x(y) = \exp\left(\frac{x \cdot y}{\tau}\right) : x, y \in \mathbb{S}^{h-1}\}$. Since $\|x\| = \|y\| = 1, x \cdot y \in$
1373 $[-1, 1]$, hence $\forall f_x \in \mathcal{F}$, we have:

1374

1375
$$|f_x(y)| \leq e^{1/\tau}. \quad (58)$$

1376

1377 Therefore, $f_x(y)$ is uniformly bounded in y , so is its derivative:

1378

1379
$$\|\nabla_x f_x(y)\| = \left\| \frac{y}{\tau} f_x(y) \right\| \leq \frac{1}{\tau} e^{1/\tau}. \quad (59)$$

1380
1381

1382 Then $\forall x_k \in \mathbb{S}^{h-1}$,

1383

1384
$$|f_x(y) - f_{x_k}(y)| \leq \frac{1}{\tau} e^{1/\tau} =: L. \quad (60)$$

1385
1386

1387 Thus, $f_x(y)$ is Lipschitz in x with constant $L = \frac{e^{1/\tau}}{\tau}$, uniformly in y .1388 **Step 2 η -Net:**

1389

1390 According to Lemma 5.2 in (Vershynin, 2010), $\forall \varepsilon > 0$ and $\eta = \frac{\varepsilon}{4L}$, there exists a finite η -net,
1391 $\mathcal{N}_\eta = \{x_1, x_2, \dots, x_K\} \subset \mathbb{S}^{h-1}$, with cardinality:

1392

1393
$$K = |\mathcal{N}_\eta| \leq \left(1 + \frac{2}{\eta}\right)^h < \left(\frac{3}{\eta}\right)^h. \quad (61)$$

1394
1395

1396 $\forall x \in \mathbb{S}^{h-1}$, $\exists x_k \in \mathcal{N}_\eta$ such that $\|x - x_k\| < \eta$. Because $f_x(y)$ is L -Lipschitz in x , we have:

1397

1398
$$|f_x(y) - f_{x_k}(y)| \leq L \|x - x_k\| = L\eta. \quad (62)$$

1399

1400

1401 And we also have:

1402

1403
$$\begin{aligned} |g_N(x) - g_N(x_k)| &\leq L\eta, \\ |g(x) - g(x_k)| &\leq L\eta. \end{aligned} \quad (63)$$

1404
1405

1404 **Step 3 Probability Bound:**

1405
 1406 $\forall x_k \in \mathcal{N}_\eta$, the random variables $Z_j := f_{x_k}(y_j)$ are iid and lie in $[e^{-1/\tau}, e^{1/\tau}]$. According to the
 1407 Hoeffding's inequality:

1408
 1409
$$P\left(|g_N(x_k) - g(x_k)| > \frac{\varepsilon}{2}\right) \leq 2 \exp\left(-\frac{2N(\varepsilon/2)^2}{(2e^{1/\tau})^2}\right) = 2e^{-cN\varepsilon^2}, \quad (64)$$

1410
 1411 where $c = \frac{1}{8e^{2/\tau}} > 0$. Taking a union bound over the η -net:

1412
 1413
$$P\left(\max_{x_k \in \mathcal{N}_\eta} |g_N(x_k) - g(x_k)| > \frac{\varepsilon}{2}\right) \leq 2Ke^{-cN\varepsilon^2}. \quad (65)$$

1414 **Step 4 Uniform Convergence:**

1415 Since $\forall x \in \mathbb{S}^{h-1}$, $|g_N(x) - g(x)|$ can be decomposed as:

1416
 1417
$$\begin{aligned} |g_N(x) - g(x)| &\leq |g_N(x) - g_N(x_k)| + |g_N(x_k) - g(x_k)| + |g(x_k) - g(x)| \\ &\leq 2L\eta + \max_{x_k \in \mathcal{N}_\eta} |g_N(x_k) - g(x_k)| \\ &= \frac{\varepsilon}{2} + \max_{x_k \in \mathcal{N}_\eta} |g_N(x_k) - g(x_k)|. \end{aligned} \quad (66)$$

1418 Plugging Eq. (65) into Eq. (66), we have:

1419
 1420
$$\begin{aligned} P\left(\sup_{x \in \mathbb{S}^{h-1}} |g_N(x) - g(x)| > \varepsilon\right) &\leq P\left(\max_{x_k \in \mathcal{N}_\eta} |g_N(x_k) - g(x_k)| > \frac{\varepsilon}{2}\right) \\ &\leq 2Ke^{-cN\varepsilon^2}, \end{aligned} \quad (67)$$

1421 and therefore:

1422
 1423
$$\sup_{x \in \mathbb{S}^{h-1}} |g_N(x) - g(x)| \xrightarrow[N \rightarrow \infty]{P} 0. \quad (68)$$

1424 Eq. (67) justifies that:

1425
 1426
$$\sum_{N=1}^{\infty} P\left(\sup_{x \in \mathbb{S}^{h-1}} |g_N(x) - g(x)| > \varepsilon\right) \leq 2K \sum_{N=1}^{\infty} e^{-cN\varepsilon^2} < \infty. \quad (69)$$

1427 According to the Borel–Cantelli lemma:

1428
 1429
$$P\left(\limsup_{N \rightarrow \infty} \sup_{x \in \mathbb{S}^{h-1}} |g_N(x) - g(x)| > \varepsilon\right) = 0. \quad (70)$$

1430 Therefore:

1431
 1432
$$\sup_{x \in \mathbb{S}^{h-1}} |g_N(x) - g(x)| \xrightarrow[N \rightarrow \infty]{\text{a.s.}} 0. \quad (71)$$

1433 We conclude now the empirical averages $g_N(\cdot)$ converge uniformly in \mathbb{S}^{h-1} to $g(\cdot)$:

1434
 1435
$$g_N(x) \xrightarrow[N \rightarrow \infty]{\text{unif.}} g(x). \quad (72)$$

1436 \square

1458
 1459 **Lemma 2.** Let $x \in \mathbb{S}^{h-1}$ and Y be an N -point configuration, where $Y = (y_1, \dots, y_N) \in (\mathbb{S}^{h-1})^N$
 1460 are iid samples from μ_y . $\forall \tau > 0$, define a sequence of functions $\{h_N\} : \mathbb{S}^{h-1} \rightarrow \mathbb{R}$ as:

1461
 1462
$$h_N(x) = \log \left(\frac{1}{N} \sum_{j=1}^N \exp \left(\frac{x \cdot y_j}{\tau} \right) \right). \quad (73)$$

 1463
 1464

1465 Define a function $h : \mathbb{S}^{h-1} \rightarrow \mathbb{R}$ as:

1466
 1467
$$h(x) = \log \left(\mathbb{E}_{y \sim \mu_y} \left[\exp \left(\frac{x \cdot y}{\tau} \right) \right] \right). \quad (74)$$

 1468
 1469

1470 It holds that $\{h_N\}$ converges uniformly to h :

1471
 1472
$$\lim_{N \rightarrow \infty} h_N(x) \xrightarrow[N \rightarrow \infty]{\text{unif.}} h(x). \quad (75)$$

 1473
 1474

1475 *Proof.* According to Lemma 1:

1476
 1477
$$\sum_{j=1}^N \exp \left(\frac{x \cdot y_j}{\tau} \right) = g_N(x) \xrightarrow[N \rightarrow \infty]{\text{unif.}} g(x) = \mathbb{E}_{y \sim \mu_y} \left[\exp \left(\frac{x \cdot y}{\tau} \right) \right], \quad (76)$$

 1478
 1479

1480 and

1481
 1482
$$\sup_{x \in \mathbb{S}^{h-1}} |g_N(x) - g(x)| \xrightarrow[N \rightarrow \infty]{\text{a.s.}} 0. \quad (77)$$

 1483
 1484

1485 Because $\langle x, y \rangle \in [-1, 1]$ for unit vectors, $\exp(x \cdot y / \tau)$ satisfies:

1486
 1487
$$e^{-1/\tau} \leq \exp \left(\frac{x \cdot y}{\tau} \right) \leq e^{1/\tau}. \quad (78)$$

 1488
 1489

1490 Hence $\forall x, g_N(x), g(x) \in [a, b]$ with $a = e^{-1/\tau} > 0, b = e^{1/\tau} > 0$. In the compact interval $[a, b]$,
 1491 by the mean value theorem, $\forall u < v \in [a, b], \exists u < \xi < v$ such that:

1492
 1493
$$|\log u - \log v| = \frac{|u - v|}{\xi} \leq \frac{1}{a} |u - v| = e^{1/\tau} |u - v|. \quad (79)$$

 1494
 1495

1496 Thus, the function $\log(\cdot)$ is Lipschitz. Therefore:

1497
 1498
$$\sup_{x \in \mathbb{S}^{h-1}} |h_N(x) - h(x)| = \sup_{x \in \mathbb{S}^{h-1}} |\log g_N(x) - \log g(x)| \leq \frac{1}{a} \sup_{x \in \mathbb{S}^{h-1}} |g_N(x) - g(x)| \xrightarrow[N \rightarrow \infty]{\text{a.s.}} 0 \quad (80)$$

 1499
 1500

1501 We conclude now $h_N(\cdot)$ converge uniformly in \mathbb{S}^{h-1} to $h(\cdot)$:

1502
 1503
$$\lim_{N \rightarrow \infty} h_N(x) \xrightarrow{\text{unif.}} h(x) \quad (81)$$

 1504
 1505

1506 \square

1507
 1508
 1509

1510
 1511

1512 **Lemma 3.** Let $M(\mathbb{S}^{h-1})$ be the set of Borel probability measures in \mathbb{S}^{h-1} . Let $\sigma_{h-1} \in M(\mathbb{S}^{h-1})$
 1513 be the uniform probability measure in \mathbb{S}^{h-1} . $\forall x, y \in \mathbb{S}^{h-1}$ and $\tau > 0$, a function $f : \mathbb{S}^{h-1} \times \mathbb{S}^{h-1} \rightarrow$
 1514 \mathbb{R}^+ is defined as:
 1515

$$1516 \quad 1517 \quad f(x, y) = \exp\left(\frac{x \cdot y}{\tau}\right). \quad (82)$$

1518 $\forall \mu \in M(\mathbb{S}^{h-1})$, a functional $\mathcal{G} : M(\mathbb{S}^{h-1}) \rightarrow \mathbb{R}^+$ is defined as:
 1519

$$1520 \quad 1521 \quad \mathcal{F}_f[\mu] = \int_{\mathbb{S}^{h-1}} \log\left(\int_{\mathbb{S}^{h-1}} f(x, y) d\mu(y)\right) d\mu(x). \quad (83)$$

1522 It holds that σ_{h-1} is the unique minimizer of \mathcal{F} :
 1523

$$1524 \quad 1525 \quad \min_{\mu \in M(\mathbb{S}^{h-1})} \mathcal{F}_f[\mu] = \min_{\mu \in M(\mathbb{S}^{h-1})} \int_{\mathbb{S}^{h-1}} \log\left(\int_{\mathbb{S}^{h-1}} f(x, y) d\mu(y)\right) d\mu(x). \quad (84)$$

1526 **Proof. Step 1: A change of probability measure.**
 1527

1528 Let $\sigma := \sigma_{h-1}$ be the uniform probability in \mathbb{S}^{h-1} . By rotational invariance there is a constant c such
 1529 that:
 1530

$$1531 \quad 1532 \quad c := c_{h, \tau}(x) = \int_{y \in \mathbb{S}^{h-1}} f(x, y) d\sigma(y), \quad (85)$$

1533 which is independent of x . $\forall x \in \mathbb{S}^{h-1}$. Define a kernel K as:
 1534

$$1535 \quad 1536 \quad K(x, dy) := \frac{f(x, y)}{c} d\sigma(y), \quad (86)$$

1537 so that:
 1538

$$1539 \quad 1540 \quad \int_{y \in \mathbb{S}^{h-1}} K(x, dy) = 1. \quad (87)$$

1541 Since $f(x, y) = f(y, x)$, exchanging x and y , the following holds:
 1542

$$1543 \quad 1544 \quad \sigma(dx)K(x, dy) = \sigma(dy)K(y, dx). \quad (88)$$

1545 For any measurable $A \subset \mathbb{S}^{h-1}$, we have:
 1546

$$1547 \quad 1548 \quad K(x, A) := \int_{y \in A} \frac{f(x, y)}{c} d\sigma(y). \quad (89)$$

1549 Consider a probability distribution μ in \mathbb{S}^{h-1} , define:
 1550

$$1551 \quad 1552 \quad (\mu K)(A) := \int_{x \in \mathbb{S}^{h-1}} K(x, A) d\mu(x) \\ 1553 \quad 1554 \quad = \int_{x \in \mathbb{S}^{h-1}} \int_{y \in A} \frac{f(x, y)}{c} d\sigma(y) d\mu(x) \\ 1555 \quad 1556 \quad = \int_{y \in A} \int_{x \in \mathbb{S}^{h-1}} \frac{f(x, y)}{c} d\mu(x) d\sigma(y). \quad (90)$$

1557 Therefore, $\mu K \ll \sigma$, i.e., μK is absolutely continuous with respect to σ . By Radon–Nikodym
 1558 theorem, we have:
 1559

1566

$$\frac{d(\mu K)}{d\sigma}(y) = \frac{1}{c} \int_{x \in \mathbb{S}^{h-1}} f(x, y) d\mu(x). \quad (91)$$

1569

1570 Note that:

1571

$$\begin{aligned} 1572 \quad (\sigma K)(A) &= \int_{x \in \mathbb{S}^{h-1}} K(x, A) d\sigma(x) \\ 1573 &= \int_{x \in \mathbb{S}^{h-1}} \int_{y \in A} \frac{f(x, y)}{c} d\sigma(y) d\sigma(x) \\ 1574 &= \int_{y \in A} \int_{x \in \mathbb{S}^{h-1}} \frac{f(x, y)}{c} d\sigma(x) d\sigma(y) \\ 1575 &= \int_{y \in A} d\sigma(y) \\ 1576 &= \sigma(A) \\ 1577 \\ 1578 \\ 1579 \\ 1580 \\ 1581 \\ 1582 \end{aligned} \quad (92)$$

1583

According to Eq. (88), exchanging x and y in Eq. (91), we get:

1584

$$\frac{d(\mu K)}{d\sigma}(x) = \frac{1}{c} \int_{y \in \mathbb{S}^{h-1}} f(y, x) d\mu(y). \quad (93)$$

1585

And since $f(y, x) = f(x, y)$, then:

1586

$$\frac{d(\mu K)}{d\sigma}(x) = \frac{1}{c} \int_{y \in \mathbb{S}^{h-1}} f(x, y) d\mu(y). \quad (94)$$

1587

Step 2: An exact identity for \mathcal{F}_f

1588

Define a (normalized) zonal integral operator T on $L^2(\sigma)$ as:

1589

$$(T\rho)(x) = \frac{1}{c} \int_{\mathbb{S}^{h-1}} f(x, y) \rho(y) d\sigma(y), \quad (95)$$

1590

where:

1591

1600

1601

1602

1603

1604

1605

$$\begin{aligned} 1602 \quad \rho(x) &= \frac{d\mu}{d\sigma}(x) \\ 1603 \quad T &= \rho \frac{d\mu K}{d\sigma}, \\ 1604 \\ 1605 \end{aligned} \quad (96)$$

1606

with $\rho \geq 0$ and $\int \rho d\sigma = 1$. Here, $L^2(\sigma)$ is the Hilbert space of (equivalence classes of) square-integrable functions on the sphere with respect to the measure σ . Then $\mathcal{F}_f[\mu]$ can be rewritten as:

1607

1608

1609

1610

1611

1612

1613

$$\mathcal{F}_f[\mu] = \log c + \int \rho \log(T\rho) d\sigma. \quad (97)$$

Denote:

1614

1615

1616

1617

1618

1619

$$F[\rho] := \int_{\mathbb{S}^{h-1}} \rho(x) \log(T\rho(x)) d\sigma(x), \quad (98)$$

then we have:

$$\mathcal{F}_f[\mu] = \log c + F[\rho]. \quad (99)$$

1620 **Step 3: Minimize $F[\rho]$.**

1621
 1622 We will minimize $F[\rho]$ over the probability simplex $\{\rho \geq 0, \int \rho = 1\}$. Basic facts about T : the
 1623 kernel $f(x, y) = e^{(x \cdot y)/\tau}$ is smooth, symmetric, strictly positive and depends only on $x \cdot y$. Hence:

1624
 1625 • T is a positive, self-adjoint, compact operator on $L^2(\sigma)$;
 1626 • $T1 = 1$ (since c normalizes it);
 1627
 1628 • By the Funk-Hecke theorem/Jentzsch-Perron-Frobenius, the eigensystem of T is constituted
 1629 of the spherical harmonics $\{Y_{\ell m}\}$ with eigenvalues $\lambda_0 = 1 > \lambda_1 \geq \lambda_2 \geq \dots > 0$. The
 1630 eigenspace corresponding to λ_0 has dimension 1 and contains only constant functions.
 1631
 1632 • In particular, on the mean-zero subspace $L_0^2(\sigma) = \{f : \int f d\sigma = 0\}$ all the eigenvalues
 1633 $\lambda_\ell, \ell \geq 1$ are strictly positive and bounded from above by $\lambda_1 < 1$.
 1634
 1635 • As a consequence, for any $\eta \in L_0^2(\sigma)$ we have $\|T\eta\|_{L^2} \leq \lambda_1 \|\eta\|_{L^2}$.

1636 **(3.1) First order variation and Euler-Lagrange equation**

1637 Consider a mass-preserving perturbation $\rho_\varepsilon = \rho + \varepsilon\eta$ with $\int \eta d\sigma = 0$. Because T is linear,
 1638

1639
 1640
$$\frac{d}{d\varepsilon} F[\rho_\varepsilon] \Big|_{\varepsilon=0} = \int \eta \log(T\rho) d\sigma + \int \rho \frac{T\eta}{T\rho} d\sigma = \int \eta \left[\log(T\rho) + T\left(\frac{\rho}{T\rho}\right) \right] d\sigma, \quad (100)$$

 1641

1642 where we used self-adjointness: $\int \rho \frac{T\eta}{T\rho} = \int \eta T(\rho/T\rho)$. Introduce a Lagrange multiplier λ for the
 1643 constraint $\int \rho = 1$.
 1644

1645 The stationarity $\delta(F - \lambda \int \rho) = 0$ for all mean-zero η yields the Euler-Lagrange (EL) equation:
 1646

1647
 1648
$$\log(T\rho)(x) + T\left(\frac{\rho}{T\rho}\right)(x) = \lambda \quad \text{for } \sigma\text{-a.e. } x. \quad (101)$$

 1649

1650 We easily check that $\rho \equiv 1$ is a critical point.

1651 If $\rho \equiv 1$, then $T\rho \equiv 1$, hence $\log(T\rho) \equiv 0$ and $T(\rho/T\rho) = T1 = 1$. Thus Eq. (101) holds with
 1652 $\lambda = 1$.
 1653

1654 **(3.2) Second order variation at the uniform density**

1655 Let $\rho \equiv 1$ and perturb $\rho_\varepsilon = 1 + \varepsilon\eta$ with $\int \eta = 0$.
 1656

1657 Differentiate the first-variation formula once more in the same direction η :

1658
 1659 • The directional derivative of $\log(T\rho)$ is $(T\eta)/(T\rho)$, so at $\rho = 1$ it is $T\eta$.
 1660 • The map $\rho \mapsto T(\rho/T\rho)$ has derivative at $\rho = 1$:

1661
 1662
 1663
$$D[T(\rho/T\rho)]|_{\rho=1}[\eta] = T(\eta - T\eta) = T\eta - T(T\eta). \quad (102)$$

 1664

1665 Hence the (constrained) second variation is

1666
 1667
$$\delta^2 F[1; \eta] = \int \eta (T\eta + T\eta - T(T\eta)) d\sigma = 2\langle \eta, T\eta \rangle - \langle T\eta, T\eta \rangle. \quad (103)$$

 1668

1669 Use the spectral decomposition $\eta = \sum_{\ell \geq 1, m} a_{\ell m} Y_{\ell m}$ (no $\ell = 0$ term because $\int \eta = 0$). Since
 1670 $TY_{\ell m} = \lambda_\ell Y_{\ell m}$,
 1671

1672
 1673
$$\delta^2 F[1; \eta] = \sum_{\ell \geq 1, m} (2\lambda_\ell - \lambda_\ell^2) a_{\ell m}^2 = \sum_{\ell \geq 1, m} \lambda_\ell (2 - \lambda_\ell) a_{\ell m}^2. \quad (104)$$

1674 Because $0 < \lambda_\ell < 1$ for $\ell \geq 1$, each factor $\lambda_\ell (2 - \lambda_\ell)$ is strictly positive. Therefore:

$$1675 \quad 1676 \quad 1677 \quad \delta^2 F[1; \eta] > 0 \quad \text{for every } \eta \in L_0^2(\sigma), \eta \neq 0. \quad (105)$$

1678 So $\rho \equiv 1$ is a strict local minimizer of F under the mass constraint $\int \rho d\sigma = 1$.

1680 (3.3) Uniqueness of the critical point

1681 Suppose ρ satisfies Eq. (101). Expand ρ in spherical harmonics: $\rho = 1 + \sum_{\ell \geq 1, m} a_{\ell m} Y_{\ell m}$. Since
1682 $T\rho = 1 + \sum_{\ell \geq 1, m} \lambda_\ell a_{\ell m} Y_{\ell m}$ with $0 < \lambda_\ell < 1$, the left side of Eq. (101) has a constant term 1 and
1683 non-constant part

$$1684 \quad 1685 \quad 1686 \quad \underbrace{\left(\log \left(1 + \sum \lambda_\ell a_{\ell m} Y_{\ell m} \right) \right)_{\text{non-const}}}_{\text{all harmonics } \ell \geq 1} + \underbrace{\sum \lambda_\ell a_{\ell m} Y_{\ell m}}_{T(\rho/T\rho) \text{ to first order}}. \quad (106)$$

1687 Project Eq. (101) onto each harmonic $Y_{\ell m}$ with $\ell \geq 1$. A standard contraction/implicit-function
1688 argument (or just comparing coefficients to first order and using that higher-order terms can't cancel
1689 all modes simultaneously because $|\lambda_\ell| < 1$) forces all $a_{\ell m} = 0$. Thus any solution of (EL) is
1690 constant; with mass 1, the only solution is $\rho \equiv 1$.

1691 So $\rho \equiv 1$ is the unique critical point of F on the simplex of probability measures on \mathbb{S}^{h-1} .

1692 (3.4) Global minimality

1693 Since F is lower semi-continuous for the weak topology on the set $M(\mathbb{S}^{h-1})$ of probability measures
1694 on the sphere, a global minimizer exists by compactness. Since any minimizer must satisfy (EL) and
1695 the only critical point is $\rho \equiv 1$, the global minimizer is $\rho \equiv 1$, i.e. $\mu = \sigma$. \square

1700
1701 **Lemma 4.** Let $M(\mathbb{S}^{h-1})$ be the set of Borel probability measures in \mathbb{S}^{h-1} . Let $\sigma_{h-1} \in M(\mathbb{S}^{h-1})$
1702 be the uniform probability measure in \mathbb{S}^{h-1} . $\forall x, y \in \mathbb{S}^{h-1}$ and $\tau > 0$, a function $f : \mathbb{S}^{h-1} \times \mathbb{S}^{h-1} \rightarrow$
1703 \mathbb{R}^+ is defined as:

$$1704 \quad 1705 \quad 1706 \quad f(x, y) = \exp \left(\frac{x \cdot y}{\tau} \right). \quad (107)$$

1707
1708 $\forall \mu \in M(\mathbb{S}^{h-1})$, a functional $\mathcal{F} : M(\mathbb{S}^{h-1}) \rightarrow \mathbb{R}^+$ is defined as:

$$1709 \quad 1710 \quad 1711 \quad \mathcal{F}_f[\mu] = \int_{\mathbb{S}^{h-1}} \log \left(\int_{\mathbb{S}^{h-1}} f(x, y) d\mu(y) \right) d\mu(x). \quad (108)$$

1712
1713 Let $\Gamma(\cdot)$ be the Gamma function and $\nu = h/2 - 1$, it holds that:

$$1714 \quad 1715 \quad 1716 \quad \mathcal{F}_f[\sigma_{h-1}] = \log \left(\Gamma(\nu + 1) (2\tau)^\nu I_\nu \left(\frac{1}{\tau} \right) \right) \quad (109)$$

1717 **Proof. Step 1: Rotational Invariance**

1718 Since the measure σ_{h-1} is invariant under orthogonal transformations. For any fixed $x \in \mathbb{S}^{h-1}$, the
1719 inner integral:

$$1720 \quad 1721 \quad 1722 \quad \int_{\mathbb{S}^{h-1}} \exp \left(\frac{x \cdot y}{\tau} \right) d\sigma_{h-1}(y), \quad (110)$$

1723
1724 depends only on the distribution of $(x \cdot y)$, and by rotational symmetry, this integral is independent of
1725 x . Thus, define:

1728

$$1729 \quad Z_\tau := \int_{\mathbb{S}^{h-1}} \exp\left(\frac{x \cdot y}{\tau}\right) d\sigma_{h-1}(y), \quad (111)$$

1730

1731 and Z_τ is constant for all x . Since $\log Z_\tau$ is constant and σ_{h-1} is a probability measure, we have:

1732

$$1733 \quad \mathcal{F}_f [\sigma_{h-1}] = \int_{\mathbb{S}^{h-1}} \log Z_\tau d\sigma_{h-1}(x) = \log Z_\tau, \quad (112)$$

1734

1735 **Step 2: Compute Z_τ**

1736

1737 Without the loss of generality, we assume the coordinate of x as:

1738

1739

1740

1741

$$x = e_h = (0, \dots, 0, 1). \quad (113)$$

1742

1743 Then $x \cdot y = y_h$, the last coordinate of y . So:

1744

1745

1746

$$Z_\tau = \int_{\mathbb{S}^{h-1}} \exp\left(\frac{y_h}{\tau}\right) d\sigma_{h-1}(y). \quad (114)$$

1747

1748

1749 Let $t = y_h = x \cdot y \in [-1, 1]$. The pushforward of σ_{h-1} under the map $y \mapsto x \cdot y$ has probability density:

1750

1751

1752

1753

$$p_h(t) = \frac{\Gamma\left(\frac{h}{2}\right)}{\Gamma\left(\frac{h-1}{2}\right)\sqrt{\pi}} (1-t^2)^{\frac{h-3}{2}}, \quad t \in [-1, 1]. \quad (115)$$

1754

1755

1756 Then:

1757

1758

1759

$$Z_\tau = \int_{-1}^1 \exp\left(\frac{t}{\tau}\right) p_h(t) dt = \frac{\Gamma\left(\frac{h}{2}\right)}{\Gamma\left(\frac{h-1}{2}\right)\sqrt{\pi}} \int_{-1}^1 e^{t/\tau} (1-t^2)^{\frac{h-3}{2}} dt. \quad (116)$$

1760

1761

1762 A classical integral (equivalently, an integral representation of the modified Bessel I_ν) is:

1763

1764

$$\int_{-1}^1 e^{\kappa t} (1-t^2)^{\nu-\frac{1}{2}} dt = \sqrt{\pi} \Gamma\left(\nu + \frac{1}{2}\right) \left(\frac{2}{\kappa}\right)^\nu I_\nu(\kappa), \quad \kappa > 0, \nu > -\frac{1}{2}. \quad (117)$$

1765

1766

1767 Set: $\kappa = \frac{1}{\tau}$ and $\nu = \frac{h-2}{2}$, so that $\nu - \frac{1}{2} = \frac{h-3}{2}$. Then:

1768

1769

1770

$$\int_{-1}^1 e^{t/\tau} (1-t^2)^{\frac{h-3}{2}} dt = \sqrt{\pi} \Gamma\left(\frac{h-1}{2}\right) (2\tau)^{\frac{h}{2}-1} I_{\frac{h}{2}-1}\left(\frac{1}{\tau}\right). \quad (118)$$

1771

1772 Substitute into Z_τ :

1773

1774

1775

$$Z_\tau = \frac{\Gamma\left(\frac{h}{2}\right)}{\Gamma\left(\frac{h-1}{2}\right)\sqrt{\pi}} \cdot \sqrt{\pi} \Gamma\left(\frac{h-1}{2}\right) (2\tau)^{\frac{h}{2}-1} I_{\frac{h}{2}-1}\left(\frac{1}{\tau}\right). \quad (119)$$

1776

1777

1778 Simplify:

1779

1780

1781

$$Z_\tau = \Gamma\left(\frac{h}{2}\right) (2\tau)^{\frac{h}{2}-1} I_{\frac{h}{2}-1}\left(\frac{1}{\tau}\right). \quad (120)$$

1782 **Step 3: Compute $\mathcal{F}_f [\sigma_{h-1}]$**

1782

1783

$$\begin{aligned}
 \mathcal{F}_f [\sigma_{h-1}] &= \log Z_\tau \\
 &= \log \left[\Gamma \left(\frac{h}{2} \right) (2\tau)^{\frac{h}{2}-1} I_{\frac{h}{2}-1} \left(\frac{1}{\tau} \right) \right] \\
 &= \log \left(\Gamma (\nu + 1) (2\tau)^\nu I_\nu \left(\frac{1}{\tau} \right) \right)
 \end{aligned} \tag{121}$$

1789

1790

1791

1792

1793

1794

1795

1796

1797

1798

1799

1800

1801

1802

1803

1804

1805

1806

1807

1808

1809

1810

1811

1812

1813

1814

1815

1816

1817

1818

1819

1820

1821

1822

1823

1824

1825

1826

1827

1828

1829

1830

1831

1832

1833

1834

1835

□

1836 E.2 DETAILS OF THEOREM 2
1837

1838 In this section, we provide proofs of Theorem 2 that is proposed in Sec. 3.3. We also provide
1839 details and proofs of the auxiliary theorems (Theorem S3 and Theorem S4) and the technical lemmas
1840 (Lemma 5, Lemma 6, Lemma 7, Lemma 8 and Lemma 9) that support the proof Theorem 2. For
1841 convenience in reading, let us recall some related notions and definitions.

- 1842 • $h, N \in \mathbb{N}$.
- 1843 • $\mathbb{S}^{h-1} = \{z \in \mathbb{R}^h : \|z\| = 1\}$.
- 1844 • $X = (x_1, \dots, x_N) \in (\mathbb{S}^{h-1})^N$.
- 1845 • $Y = (y_1, \dots, y_N) \in (\mathbb{S}^{h-1})^N$.
- 1846 • $\mu_x = \frac{1}{N} \sum_{i=1}^N x_i$.
- 1847 • $\mu_y = \frac{1}{N} \sum_{i=1}^N y_i$.
- 1848 • $c_x = \frac{\mu_x}{\|\mu_x\|}$.
- 1849 • $c_y = \frac{\mu_y}{\|\mu_y\|}$.

1854 **Definition** (Multimodal Contrastive Loss (MCL Loss)). Let (X, Y) be an N -pair configuration,
1855 where $X = (x_1, \dots, x_N) \in (\mathbb{S}^{h-1})^N$ and $Y = (y_1, \dots, y_N) \in (\mathbb{S}^{h-1})^N$. $\forall \tau > 0$, the multimodal
1856 contrastive loss $\mathcal{L}_{\text{MCL}}(\cdot, \cdot) : (\mathbb{S}^{h-1})^N \times (\mathbb{S}^{h-1})^N \rightarrow \mathbb{R}$ is defined as:

$$1858 \quad \mathcal{L}_{\text{MCL}} = \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{\text{MCL}}^i, \text{ where } \mathcal{L}_{\text{MCL}}^i = \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) + \mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}(y_i; X).$$

1861 Here, $\mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}$ is the \mathcal{X} -to- \mathcal{Y} alignment and $\mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}$ is the \mathcal{Y} -to- \mathcal{X} alignment, which are defined
1862 respectively as:

$$1865 \quad \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) = -\log \frac{\exp(x_i \cdot y_i / \tau)}{\sum_{j=1}^N \exp(x_i \cdot y_j / \tau)}, \quad \mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}(y_i; X) = -\log \frac{\exp(x_i \cdot y_i / \tau)}{\sum_{j=1}^N \exp(x_j \cdot y_i / \tau)}.$$

1868 **Definition** (Modality Gap) Let (X, Y) be an N -pair configuration, where $X = (x_1, \dots, x_N) \in$
1869 $(\mathbb{S}^{h-1})^N$ and $Y = (y_1, \dots, y_N) \in (\mathbb{S}^{h-1})^N$. The modality gap between X and Y can be expressed
1870 as the angle between the center representations:

$$1872 \quad \Delta_\theta = \cos^{-1}(c_x \cdot c_y).$$

1874 **Definition** (vMF Distribution). $\forall c \in \mathbb{S}^{h-1}$ and $\kappa \geq 0$, the probability density of a random h -
1875 dimensional unit vector $z \sim \text{vMF}(c, \kappa)$ is given by:

$$1877 \quad f_h(z; c, \kappa) = D_h(\kappa) e^{\kappa c^\top z}, \text{ where } D_h(\kappa) = \frac{\kappa^\nu}{(2\pi)^\nu + 1 I_\nu(\kappa)}.$$

1880 Here, $\nu = h/2 - 1$, and $I_\nu(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ is the modified Bessel function of the first kind of order ν ,
1881 which is defined as:

$$1883 \quad I_\nu(x) = \sum_{k=0}^{\infty} \frac{1}{k! \Gamma(\nu + k + 1)} \left(\frac{x}{2}\right)^{2k+\nu}.$$

1886 **Definition** (Function \tilde{M}). $\forall \kappa, \tau > 0$, a function $\tilde{M}_\kappa(\cdot, \cdot) : [-1, 1] \times [0, 1] \rightarrow \mathbb{R}_0^+$ is defined as:

$$1889 \quad \tilde{M}_\kappa(w, t) = \sqrt{\kappa^2 + \frac{2\kappa w}{\tau} + \frac{t^2}{\tau^2}}.$$

1890 **Definition** (Function $\tilde{\mathcal{J}}$). $\forall \kappa, \nu, \tau > 0$, $\tilde{\mathcal{J}}(\cdot, \cdot, \cdot; \kappa, \nu) : [-1, 1] \times [-1, 1] \times [0, 1] \rightarrow \mathbb{R}$ is defined as:
 1891

$$1892 \quad 1893 \quad \tilde{\mathcal{J}}(w_1, w_2, t; \kappa, \nu) = -\frac{w_1}{\tau} + \log \left(\frac{I_\nu \left(\tilde{M}_\kappa(w_2, t) \right)}{\tilde{M}_\kappa(w_2, t)^\nu} \right) - \log \left(\frac{I_\nu(\kappa)}{\kappa^\nu} \right).$$

$$1894$$

$$1895$$

1896 **Definition** (Function M). $\forall \kappa, \tau > 0$, a function $M_\kappa(\cdot) : [-1, 1] \rightarrow \mathbb{R}_0^+$ is defined as:
 1897

$$1898 \quad 1899 \quad M_\kappa(w) = \sqrt{\kappa^2 + \frac{2\kappa w}{\tau} + \frac{1}{\tau^2}}$$

$$1900 \quad 1901 \quad = \tilde{M}_\kappa(w, 1).$$

$$1902$$

1903 **Definition** (Function \mathcal{J}). $\forall \kappa, \nu, \tau > 0$, a function $\mathcal{J}(\cdot; \kappa, \nu) : [-1, 1] \rightarrow \mathbb{R}$ is defined as:
 1904

$$1905 \quad 1906 \quad \mathcal{J}(w; \kappa, \nu) = -\frac{w}{\tau} + \log \left(\frac{I_\nu(M_\kappa(w))}{M_\kappa(w)^\nu} \right) - \log \left(\frac{I_\nu(\kappa)}{\kappa^\nu} \right)$$

$$1907 \quad 1908 \quad = \tilde{\mathcal{J}}(w, w, 1; \kappa, \nu).$$

$$1909$$

1910 **Definition** (Function $\hat{\mathcal{J}}$). $\forall \kappa, \nu, \tau > 0$, a function $\hat{\mathcal{J}}(\cdot, \cdot; \kappa, \nu) : [-1, 1] \times [0, 1] \rightarrow \mathbb{R}$ is defined as:
 1911

$$1912 \quad 1913 \quad \hat{\mathcal{J}}(w, t; \kappa, \nu) = -\frac{w}{\tau} + \log \left(\frac{I_\nu \left(\tilde{M}_\kappa(w, t) \right)}{\tilde{M}_\kappa(w, t)^\nu} \right) - \log \left(\frac{I_\nu(\kappa)}{\kappa^\nu} \right)$$

$$1914 \quad 1915 \quad = \tilde{\mathcal{J}}(w, w, t; \kappa, \nu).$$

$$1916$$

1917 E.2.1 PROOF OF THEOREM 2

1919 In this subsection, we provide the proof of Theorem 2. For convenience in reading, we first restate
 1920 Theorem 2 here.
 1921

1922 **Theorem 2.** [Restate] Let (X, Y) be an N -pair configuration, where $X = (x_1, \dots, x_N) \in (\mathbb{S}^{h-1})^N$
 1923 are iid samples from $\mu_x = \text{vMF}(c_x, \kappa_x)$, and $Y = (y_1, \dots, y_N) \in (\mathbb{S}^{h-1})^N$ are iid samples from
 1924 $\mu_y = \text{vMF}(c_y, \kappa_y)$. Let $\nu = h/2 - 1$. Suppose there exists an index $i = c$ such that $x_c = c_x$,
 1925 $y_c = c_y$. Denote $\Delta_\theta = \cos^{-1}(c_x \cdot c_y)$. For any fixed $\kappa_x, \kappa_y > 0$, it holds that:
 1926

$$1927 \quad \lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}}^c - 2 \log(N) = \mathcal{J}(\cos(\Delta_\theta); \kappa_y, \nu) + \mathcal{J}(\cos(\Delta_\theta); \kappa_x, \nu)$$

$$1928 \quad = \tilde{\mathcal{J}}(\cos(\Delta_\theta), \cos(\Delta_\theta), 1; \kappa_y, \nu) + \tilde{\mathcal{J}}(\cos(\Delta_\theta), \cos(\Delta_\theta), 1; \kappa_x, \nu)$$

$$1929 \quad \geq \mathcal{J}(1; \kappa_y, \nu) + \mathcal{J}(1; \kappa_x, \nu)$$

$$1930 \quad = \tilde{\mathcal{J}}(1, 1, 1; \kappa_y, \nu) + \tilde{\mathcal{J}}(1, 1, 1; \kappa_x, \nu),$$

$$1931$$

$$1932$$

1933 where equality is attained if and only if there exists a configuration of (X, Y) such that:
 1934

$$1935 \quad (A5) \quad \Delta_\theta = \cos^{-1}(c_x \cdot c_y) = 0.$$

$$1936$$

1937 *Proof.* We first decompose $\lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}}^c - 2 \log(N)$ into two parts:
 1938

$$1939 \quad 1940 \quad \lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}}^c - 2 \log(N) = \lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(c_x; Y) - \log(N)$$

$$1941 \quad + \lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}(c_y; X) - \log(N). \tag{122}$$

$$1942$$

1943 According to Theorem S4, the convergent function and its lower bound of $\mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}$ are:
 1944

1944

$$\lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(c_x; Y) - \log(N) = \mathcal{J}(\cos(\Delta_\theta); \kappa_y, \nu) \geq \mathcal{J}(1; \kappa_y, \nu), \quad (123)$$

1945 where equality is attained if and only if there exists a configuration of (X, Y) such that:

$$(i) \Delta_\theta = \cos^{-1}(c_x \cdot c_y) = 0.$$

1946 This Theorem also holds for $\mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}$:

$$\lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}(c_y; X) - \log(N) = \mathcal{J}(\cos(\Delta_\theta); \kappa_x, \nu) \geq \mathcal{J}(1; \kappa_x, \nu), \quad (124)$$

1947 where equality is attained if and only if there exists a configuration of (X, Y) such that:

$$(ii) \Delta_\theta = \cos^{-1}(c_x \cdot c_y) = 0.$$

1948 Combining Eq. (123), Eq. (124), and consider $\mathcal{J}(w; \kappa, \nu) = \tilde{\mathcal{J}}(w, w, 1; \kappa, \nu)$, we reach the conclusion that:

$$\begin{aligned} \lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}}^c - 2 \log(N) &= \mathcal{J}(\cos(\Delta_\theta); \kappa_y, \nu) + \mathcal{J}(\cos(\Delta_\theta); \kappa_x, \nu) \\ &= \tilde{\mathcal{J}}(\cos(\Delta_\theta), \cos(\Delta_\theta), 1; \kappa_y, \nu) + \tilde{\mathcal{J}}(\cos(\Delta_\theta), \cos(\Delta_\theta), 1; \kappa_x, \nu) \\ &\geq \mathcal{J}(1; \kappa_y, \nu) + \mathcal{J}(1; \kappa_x, \nu) \\ &= \tilde{\mathcal{J}}(1, 1, 1; \kappa_y, \nu) + \tilde{\mathcal{J}}(1, 1, 1; \kappa_x, \nu), \end{aligned} \quad (125)$$

1949 where equality is attained if and only if there exists a configuration of (X, Y) such that:

$$(A5) \Delta_\theta = \cos^{-1}(c_x \cdot c_y) = 0.$$

□

E.2.2 AUXILIARY THEOREMS PART 2

1950 In this subsection, we provide details and proofs of the auxiliary theorems (Theorem S3 and Theorem S4) that support the proof of Theorem 2.

1951 **Theorem S3.** Let (X, Y) be an N -pair configuration, where $X = (x_1, \dots, x_N) \in (\mathbb{S}^{h-1})^N$ are 1952 iid samples from $\mu_x = \text{vMF}(c_x, \kappa_x)$, and $Y = (y_1, \dots, y_N) \in (\mathbb{S}^{h-1})^N$ are iid samples from 1953 $\mu_y = \text{vMF}(c_y, \kappa_y)$. Let $\nu = h/2 - 1$ and $\kappa_y > 0$.

1954 $\forall x_i \in X$, denote $w_i = x_i \cdot y_i$ and $w_{x_i, c_y} = x_i \cdot c_y$. It holds that:

$$\begin{aligned} \lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) - \log(N) &= \lim_{N \rightarrow \infty} -\log \frac{\exp(x_i \cdot y_i / \tau)}{\sum_{j=1}^N \exp(x_i \cdot y_j / \tau)} - \log(N) \\ &= -\frac{w_i}{\tau} + \log \left(\frac{I_\nu(M_{\kappa_y}(w_{x_i, c_y}))}{M_{\kappa_y}(w_{x_i, c_y})^\nu} \right) - \log \left(\frac{I_\nu(\kappa_y)}{\kappa_y^\nu} \right) \\ &= \tilde{\mathcal{J}}(w_i, w_{x_i, c_y}, 1; \kappa_y, \nu), \end{aligned} \quad (126)$$

1955 where $\forall \kappa \geq 0, \tau > 0, M_\kappa(\cdot) : [-1, 1] \rightarrow \mathbb{R}_0^+$ is defined as:

$$M_\kappa(w) = \sqrt{\kappa^2 + \frac{2\kappa w}{\tau} + \frac{1}{\tau^2}}. \quad (127)$$

1998 and I_ν is the modified Bessel function of the first kind of order ν , which is defined as:
1999

2000
2001
$$I_\nu(m) = \sum_{k=0}^{\infty} \frac{1}{k! \Gamma(\nu + k + 1)} \left(\frac{m}{2}\right)^{2k+\nu}. \quad (128)$$

2002

2003 Suppose there exists an index $i = c$ such that $x_c = c_x$, $y_c = c_y$. Denote $w_c = c_x \cdot c_y$. It holds that:
2004

2005
2006
2007
$$\lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(c_x; Y) - \log(N) = -\frac{w_c}{\tau} + \log \left(\frac{I_\nu(M_{\kappa_y}(w_c))}{M_{\kappa_y}(w_c)^\nu} \right) - \log \left(\frac{I_\nu(\kappa_y)}{\kappa_y^\nu} \right) \quad (129)$$

2008
2009
2010

2011 *Proof.* **Step 1:** We start the proof by find the convergent function of $\mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y)$ as $N \rightarrow \infty$.
2012 Same with Eq. (39) of Theorem S1, $\forall x_i \in X$, the \mathcal{X} -to- \mathcal{Y} alignment of x_i can be rewritten as:
2013

2014
2015
$$\mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) = -\log \frac{\exp(x_i \cdot y_i / \tau)}{\sum_j \exp(x_i \cdot y_j / \tau)}$$

2016
2017
2018
$$= -\frac{x_i \cdot y_i}{\tau} + \log \left(N \frac{1}{N} \sum_{j=1}^N \exp \left(\frac{x_i \cdot y_j}{\tau} \right) \right) \quad (130)$$

2019
2020
2021
2022
2023

$$= -\frac{x_i \cdot y_i}{\tau} + \log \left(\frac{1}{N} \sum_{j=1}^N \exp \left(\frac{x_i \cdot y_j}{\tau} \right) \right) + \log(N).$$

2024 Lemma 2 shows that:
2025

2026
2027
$$\lim_{N \rightarrow \infty} \log \left(\frac{1}{N} \sum_{j=1}^N \exp \left(\frac{x_i \cdot y_j}{\tau} \right) \right) = \log \left(\mathbb{E}_{y \sim \mu_y} \left[\exp \left(\frac{x_i \cdot y}{\tau} \right) \right] \right). \quad (131)$$

2028
2029

2030 According to the moment-generating function of the vMF distribution:
2031

2032
2033
$$\mathbb{E}_{y \sim \mu_y} \left[\exp \left(\frac{x_i \cdot y}{\tau} \right) \right] = \mathbb{E}_{y \sim \mu_y} \left[\exp \left(\frac{x_i}{\tau} \cdot y \right) \right] = \frac{I_\nu(\kappa'_y)}{I_\nu(\kappa_y)} \left(\frac{\kappa_y}{\kappa'_y} \right)^\nu, \quad (132)$$

2034
2035 where $\kappa'_y = \|\kappa_y c_y + \frac{x_i}{\tau}\|_2$.
2036

2037 Then we have:
2038

2039
2040
$$\lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) - \log(N) = -\frac{x_i \cdot y_i}{\tau} + \log \left(\frac{I_\nu(\kappa'_y)}{\kappa'_y^\nu} \right) - \log \left(\frac{I_\nu(\kappa_y)}{\kappa_y^\nu} \right). \quad (133)$$

2041

2042 **Step 2:** we will transform $\mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}$ from a function of vectors to a function of angles between vectors.
2043

2044 Without loss of generality, we assume the coordinate of c_y as
2045

2046
$$c_y = (1, 0, \dots, 0). \quad (134)$$

2047

2048 Denote $\cos(\theta_{x_i, c_y}) = x_i \cdot c_y$. Then x_i can be represented as:
2049

2050
2051
$$\begin{aligned} x_i &= (\cos(\theta_{x_i, c_y}), u \sin(\theta_{x_i, c_y})) \\ &= (\cos(\theta_{x_i, c_y}), u_2 \sin(\theta_{x_i, c_y}), u_3 \sin(\theta_{x_i, c_y}), \dots, u_h \sin(\theta_{x_i, c_y})), \end{aligned} \quad (135)$$

2052 where $u = (0, u_2, u_3, \dots, u_h) \cong \mathbb{S}^{h-2} \in \mathbb{S}^{h-1}$ is a unit vector orthogonal to the first axis with:
 2053

$$\|u\| = u_1^2 + u_2^2 + u_3^2 + \cdots + u_h^2 = 1. \quad (136)$$

According to Eq. (134), Eq. (135) and Eq. (136), κ'_y (in Eq. (132)) can be re-written as:

$$\begin{aligned}
\kappa'_y &= \left\| \kappa_y c_y + \frac{x_i}{\tau} \right\|_2 \\
&= \sqrt{\left(\kappa_y + \frac{\cos(\theta_{x_i, c_y})}{\tau} \right)^2 + \sum_{i=2}^h \left(\frac{\sin(\theta_{x_i, c_y}) u_i}{\tau} \right)^2} \\
&= \sqrt{\left(\kappa_y + \frac{\cos(\theta_{x_i, c_y})}{\tau} \right)^2 + \frac{\sin^2(\theta_{x_i, c_y})}{\tau^2}} \\
&= \sqrt{\kappa_y^2 + \frac{2\kappa_y \cos(\theta_{x_i, c_y})}{\tau} + \frac{1}{\tau^2}} \\
&= M_{\kappa_y}(\cos(\theta_{x_i, c_y})). \tag{137}
\end{aligned}$$

2072 Consider that $w_i = x_i \cdot y_i$, $w_{x_i, c_y} = \cos(\theta_{x_i, c_y}) = x_i \cdot c_y$, putting Eq. (133) and Eq. (137) together,
 2073 we have:

$$\begin{aligned}
\lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) - \log(N) &= -\frac{x_i \cdot y_i}{\tau} + \log \left(\frac{I_\nu(\kappa'_y)}{\kappa_y^\nu} \right) - \log \left(\frac{I_\nu(\kappa_y)}{\kappa_y^\nu} \right) \\
&= -\frac{w_i}{\tau} + \log \left(\frac{I_\nu(M_{\kappa_y}(w_{x_i, c_y}))}{M_{\kappa_y}(w_{x_i, c_y})^\nu} \right) - \log \left(\frac{I_\nu(\kappa_y)}{\kappa_y^\nu} \right) \\
&= \tilde{\mathcal{J}}(w_i, w_{x_i, c_y}, 1; \kappa_y, \nu).
\end{aligned} \tag{138}$$

2084 When there exists a data pair $i = c$ such that $x_c = c_x$, $y_c = c_y$, $w_i = w_{x_i, c_n} = w_c$, then we have:

$$\begin{aligned} \lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(c_x; Y) - \log(N) &= -\frac{w_c}{\tau} + \log \left(\frac{I_\nu(M_{\kappa_y}(w_c))}{M_{\kappa_y}(w_c)^\nu} \right) - \log \left(\frac{I_\nu(\kappa_y)}{\kappa_y^\nu} \right) \\ &= \mathcal{J}(w_c; \kappa_y, \nu) = \tilde{\mathcal{J}}(w_c, w_c, 1; \kappa_y, \nu). \end{aligned} \quad (139)$$

Theorem S4. Let (X, Y) be an N -pair configuration, where $X = (x_1, \dots, x_N) \in (\mathbb{S}^{h-1})^N$ are iid samples from $\mu_x = \text{vMF}(c_x, \kappa_x)$, and $Y = (y_1, \dots, y_N) \in (\mathbb{S}^{h-1})^N$ are iid samples from $\mu_y = \text{vMF}(c_y, \kappa_y)$. Let $\nu = h/2 - 1$. Suppose there exists an index $i = c$ such that $x_c = c_x$, $y_c = c_y$. Denote $\Delta_\theta = \cos^{-1}(c_x \cdot c_y)$. For any fixed $\kappa_y > 0$, it holds that:

$$\lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(c_x; Y) - \log(N) = \mathcal{J}(\cos(\Delta_\theta); \kappa_y, \nu) \geq \mathcal{J}(1; \kappa_y, \nu), \quad (140)$$

where equality is attained if and only if there exists a configuration of (X, Y) such that:

$$(B3) \quad \Delta_\theta \equiv \cos^{-1} (c_x \cdot c_y) = 0.$$

2105 **Proof. Step 1:** We start the proof by find the convergent function of $\mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(c_x; Y)$ as $N \rightarrow \infty$. Denote $w_c = c_x \cdot c_y$. $\forall \kappa_y > 0$, as prove in Theorem S3:

$$\begin{aligned}
\lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(c_x; Y) - \log(N) &= \lim_{N \rightarrow \infty} -\log \frac{\exp(c_x \cdot c_y / \tau)}{\sum_{j=1}^N \exp(c_x \cdot y_j / \tau)} - \log(N) \\
&= -\frac{w_c}{\tau} + \log \left(\frac{I_\nu(M_{\kappa_y}(w_c))}{M_{\kappa_y}(w_c)^\nu} \right) - \log \left(\frac{I_\nu(\kappa_y)}{\kappa_y^\nu} \right) \\
&= \mathcal{J}(w_c; \kappa_y, \nu).
\end{aligned} \tag{141}$$

where $\forall \kappa \geq 0, \tau > 0$, $\mathcal{J}(\cdot; \kappa, \nu)$ is a function on $[-1, 1]$ and $M_\kappa(\cdot) : [-1, 1] \rightarrow \mathbb{R}_0^+$ is defined as:

$$M_\kappa(w) = \sqrt{\kappa^2 + \frac{2\kappa w}{\tau} + \frac{1}{\tau^2}}, \tag{142}$$

and I_ν is the modified Bessel function of the first kind of order ν , which is defined as:

$$I_\nu(m) = \sum_{k=0}^{\infty} \frac{1}{k! \Gamma(\nu + k + 1)} \left(\frac{m}{2}\right)^{2k+\nu}. \tag{143}$$

Step 2: Next, we find the minimal value and the optimal condition of convergent function.

As shown in Lemma 5 (set $s = 1$), $\mathcal{J}(w; \kappa, \nu) = \tilde{J}(w, w, 1; \kappa, \nu)$ is a concave function of w . When a function is concave, its minimal value occurs at the endpoints of its domain. Therefore :

$$\mathcal{J}(w_c; \kappa_y, \nu) \geq \min\{\mathcal{J}(-1; \kappa_y, \nu), \mathcal{J}(1; \kappa_y, \nu)\}. \tag{144}$$

According to Lemma 6:

$$\mathcal{J}(-1; \kappa_y, \nu) \geq \mathcal{J}(1; \kappa_y, \nu). \tag{145}$$

Therefore, we conclude:

$$\lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(c_x; Y) - \log(N) = \mathcal{J}(\cos(\Delta_\theta); \kappa_y, \nu) \geq \mathcal{J}(1; \kappa_y, \nu), \tag{146}$$

where equality is attained if and only if the following conditions hold:

$$(B3) \quad \Delta_\theta = \cos^{-1}(c_x \cdot c_y) = 0.$$

□

E.2.3 TECHNICAL LEMMAS PART 2

In this subsection, we provide details and proofs of technical lemmas (Lemma 5, Lemma 6, Lemma 7, Lemma 8 and Lemma 9) that support the proof of Theorem 2, Theorem S3 and Theorem S4.

Lemma 5. $\forall \kappa, \nu, \tau > 0$ and $s \in [0, 1]$, a function $\hat{\mathcal{J}}_{t=s}(\cdot; \kappa, \nu) : (-1, 1] \rightarrow \mathbb{R}$ is defined as:

$$\begin{aligned}
\hat{\mathcal{J}}_{t=s}(w; \kappa, \nu) &= -\frac{w}{\tau} + \log \left(\frac{I_\nu(\tilde{M}_{t=s}(w))}{\tilde{M}_{t=s}(w)^\nu} \right) - \log \left(\frac{I_\nu(\kappa)}{\kappa^\nu} \right) \\
&= \hat{\mathcal{J}}(w, t = s; \kappa, \nu) = \tilde{J}(w, w, t = s; \kappa, \nu),
\end{aligned} \tag{147}$$

where $\tilde{M}_{t=s}(\cdot) : (-1, 1] \rightarrow \mathbb{R}^+$ is defined as:

$$\tilde{M}_{t=s}(w) = \sqrt{\kappa^2 + \frac{2\kappa w}{\tau} + \frac{s^2}{\tau^2}} = \tilde{M}_\kappa(w, t = s), \tag{148}$$

2160 and I_ν is the modified Bessel function of the first kind of order ν , which is defined as:
 2161

$$2162 \\ 2163 I_\nu(m) = \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(\nu+k+1)} \left(\frac{m}{2}\right)^{2k+\nu}. \quad (149) \\ 2164 \\ 2165$$

2166 It holds that, for any fixed s , $\hat{\mathcal{J}}_{t=s}(\cdot)$ is a strictly decreasing function when $w \in [0, s]$ and a concave
 2167 function $w \in (-1, 1]$.
 2168

2169
 2170 *Proof.* Let us first decompose the function $\hat{\mathcal{J}}_{t=s}$. Denote two functions $G_1(w)$ and $G_2(w)$ as:
 2171

$$2172 \\ 2173 G_1(w) = -\frac{w}{\tau}, \\ 2174 \\ 2175 G_3(m) = \log(I_\nu(m)) - \nu \log(m), \\ 2176 \\ 2177 G_2(w) = G_3(\tilde{M}_{t=s}(w)) \\ 2178 = \log(I_\nu(\tilde{M}_{t=s}(w))) - \nu \log(\tilde{M}_{t=s}(w)). \quad (150) \\ 2179$$

2180 Denote the function $G(w)$ and the constant C as:
 2181

$$2182 \\ 2183 G(w) = G_1(w) + G_2(w), \\ 2184 \\ 2185 C = -\log\left(\frac{I_\nu(\kappa)}{\kappa^\nu}\right). \quad (151) \\ 2186$$

2187 Then the function $\hat{\mathcal{J}}_{t=s}$ can be written as:
 2188

$$2189 \\ 2190 \\ 2191 \hat{\mathcal{J}}_{t=s}(w; \kappa, \nu) = -\frac{w}{\tau} + \log\left(\frac{I_\nu(\tilde{M}_{t=s}(w))}{\tilde{M}_{t=s}(w)^\nu}\right) - \log\left(\frac{I_\nu(\kappa)}{\kappa^\nu}\right) \\ 2192 \\ 2193 = G(w) + C. \quad (152) \\ 2194$$

2195 Now, we investigate derivatives of $\hat{\mathcal{J}}_{t=s}$.
 2196

2197 The first derivative of G_1 is:
 2198

$$2199 \\ 2200 G'_1(w) = -\frac{1}{\tau} < 0. \quad (153) \\ 2201$$

2202 The second derivative of G_1 is:
 2203

$$2204 \\ 2205 G''_1(w) = 0. \quad (154) \\ 2206$$

2207 According to Lemma 7, the first derivative of $G_3(m)$ is:
 2208

$$2209 \\ 2210 G'_3(m) = \frac{I_{\nu+1}(m)}{I_\nu(m)} \in (0, 1). \quad (155) \\ 2211 \\ 2212$$

2213 The derivative of $\tilde{M}_{t=s}$ is:
 2214

2214
 2215 $\tilde{M}'_{t=s}(w) = \frac{d}{dw} \left(\kappa_y^2 + \frac{s^2}{\tau^2} + 2\frac{\kappa}{\tau}w \right)^{1/2}$
 2216
 2217
 2218 $= \frac{1}{2} \left(\kappa_y^2 + \frac{s^2}{\tau^2} + 2\frac{\kappa}{\tau}w \right)^{-1/2} \cdot 2\frac{\kappa}{\tau}$
 2219
 2220 $= \frac{\kappa}{\tau} \frac{1}{\tilde{M}_{t=s}(w)}$
 2221
 2222 $> 0.$
 2223

2224 Then, the first derivative of G_2 is:
 2225

2226 $G'_2(w) = G'_3(\tilde{M}_{t=s}(w)) \tilde{M}'_{t=s}(w)$
 2227
 2228 $= \frac{I_{\nu+1}(\tilde{M}_{t=s}(w))}{I_{\nu}(\tilde{M}_{t=s}(w))} \tilde{M}'_{t=s}(w)$
 2229
 2230 $= \frac{\kappa}{\tau} \frac{1}{\tilde{M}_{t=s}(w)} \frac{I_{\nu+1}(\tilde{M}_{t=s}(w))}{I_{\nu}(\tilde{M}_{t=s}(w))}.$
 2231
 2232
 2233
 2234
 2235

2236 Combining Eq. (153) and Eq. (157), we have:
 2237

2238 $\hat{\mathcal{J}}'_{t=s}(w; \kappa, \nu) = G'(w)$
 2239
 2240 $= -\frac{1}{\tau} + \frac{\kappa}{\tau} \frac{1}{\tilde{M}_{t=s}(w)} \frac{I_{\nu+1}(\tilde{M}_{t=s}(w))}{I_{\nu}(\tilde{M}_{t=s}(w))}.$
 2241
 2242

2243 Since:
 2244

2245 $\tilde{M}_{t=s}(w) \geq \kappa \Leftrightarrow \frac{2\kappa w}{\tau} + \frac{s^2}{\tau^2} \geq 0$
 2246
 2247 $\Leftrightarrow w \geq -\frac{s^2}{2\kappa\tau}$
 2248
 2249 $\Leftrightarrow w \geq 0.$
 2250

2251 thus, when $w \in [0, 1]$, $\tilde{M}_{t=s}(w) \geq \kappa$ holds. Combining this and Eq. (155), we have:
 2252

2253
 2254 $G'(w) \leq -\frac{1}{\tau} + \frac{\kappa}{\tau} \frac{1}{\kappa} \frac{I_{\nu+1}(\tilde{M}_{t=s}(w))}{I_{\nu}(\tilde{M}_{t=s}(w))}$
 2255
 2256 $< -\frac{1}{\tau} + \frac{1}{\tau}$
 2257
 2258 $= 0.$
 2259

2260 So we can conclude that, for any fixed s , $\hat{\mathcal{J}}_{t=s}(\cdot)$ is a strictly decreasing function on $[0, s]$.
 2261

2262 Denote:
 2263

2264
 2265 $H(m) = \frac{1}{m} \frac{I_{\nu+1}(m)}{I_{\nu}(m)},$
 2266
 2267

according to Lemma 8,

2268
2269
2270

$$H'(m) < 0. \quad (162)$$

2271 Since $G'_2(w)$ can be written as:
2272

2273
$$G'_2(w) = \frac{\kappa}{\tau} H\left(\tilde{M}_{t=s}(w)\right), \quad (163)$$

2275 combining Eq. (156) and Eq. (163), we have
2276

2277
2278
$$G''_2(w) = \frac{\kappa}{\tau} H'\left(\tilde{M}_{t=s}(w)\right) \tilde{M}'_{t=s}(w) \quad (164)$$

2279
2280

2281 Given Eq. (157) and Eq. (164), we can conclude that G_2 is an increasing and concave function.
2282 Combining Eq. (154) and Eq. (164), we have:
2283

2284
2285
$$\hat{\mathcal{J}}''_{t=s}(w; \kappa, \nu) = G''(w) \quad (165)$$

2286
2287
2288

2289 So we can conclude that, for any fixed s , $\hat{\mathcal{J}}_{t=s}(\cdot)$ is a concave function on $(-1, 1]$.
2290

□

2291 **Lemma 6.** $\forall \kappa, \nu, \tau > 0$, a function $\mathcal{J}(\cdot) : [-1, 1] \rightarrow \mathbb{R}$ is defined as:
2292

2293
2294
$$\mathcal{J}(w) = -\frac{w}{\tau} + \log\left(\frac{I_\nu(M(w))}{M(w)^\nu}\right) - \log\left(\frac{I_\nu(\kappa)}{\kappa^\nu}\right) + \log(N), \quad (166)$$

2295
2296

2297 where $M_\kappa(\cdot) : [-1, 1] \rightarrow \mathbb{R}$ is defined as:
2298

2299
2300
$$M_\kappa(w) = \sqrt{\kappa^2 + \frac{2\kappa w}{\tau} + \frac{1}{\tau^2}}, \quad (167)$$

2301

2302 and I_ν is the modified Bessel function of the first kind of order ν , which is defined as:
2303

2304
2305
$$I_\nu(m) = \sum_{k=0}^{\infty} \frac{1}{k! \Gamma(\nu + k + 1)} \left(\frac{m}{2}\right)^{2k+\nu}. \quad (168)$$

2306
2307

2308 $\forall 0 < w \leq 1$, it holds that:
2309

2310
$$\mathcal{J}(w) < \mathcal{J}(-w). \quad (169)$$

2311

2312 *Proof.* Let us first re-write Eq. (169) as:
2313

2314
$$\mathcal{J}(w) < \mathcal{J}(-w) \Leftrightarrow \mathcal{J}(-w) - \mathcal{J}(w) > 0, \quad (170)$$

2315

2316 and we will prove the inequality on RHS. Denote:
2317

2318
2319
$$a = M(-w) = \sqrt{\kappa^2 + \frac{1}{\tau^2} - \frac{2\kappa w}{\tau}}, \quad (171)$$

2320
2321
$$b = M(w) = \sqrt{\kappa^2 + \frac{1}{\tau^2} + \frac{2\kappa w}{\tau}}.$$

2322 In (Eq. (155) of) Lemma 5, it is shown that $M(\cdot)$ is a strictly increasing function. Then, we have:
 2323

$$2324 \quad 0 < a < b, \quad (172)$$

2325 and then we have:
 2326

$$\begin{aligned} 2327 \quad \mathcal{J}(-w) - \mathcal{J}(w) &= \frac{w}{\tau} - \left(-\frac{w}{\tau}\right) + \log\left(\frac{I_\nu(a)}{I_\nu(b)}\right) - \nu \log\left(\frac{a}{b}\right) \\ 2328 \quad &= \frac{2w}{\tau} + \log\frac{I_\nu(a)}{I_\nu(b)} - \nu \log\left(\frac{a}{b}\right). \end{aligned} \quad (173)$$

2329 According to Lemma 9:
 2330

$$\begin{aligned} 2331 \quad \log\left(\frac{I_\nu(a)}{I_\nu(b)}\right) - \nu \log\left(\frac{a}{b}\right) &> (a - b). \end{aligned} \quad (174)$$

2332 Plugging Eq. (174) into Eq. (173), we get:
 2333

$$\begin{aligned} 2334 \quad \mathcal{J}(-w) - \mathcal{J}(w) &> \frac{2w}{\tau} + (a - b) = f(w). \end{aligned} \quad (175)$$

2335 Combining Eq. (170) and Eq. (175), we have:
 2336

$$\begin{aligned} 2337 \quad \mathcal{J}(w) \leq \mathcal{J}(-w) \Leftrightarrow f(w) \geq 0. \end{aligned} \quad (176)$$

2338 Denote:
 2339

$$\begin{aligned} 2340 \quad A &= \kappa^2 + \frac{1}{\tau^2}, \\ 2341 \quad B &= \frac{2\kappa}{\tau}, \end{aligned} \quad (177)$$

2342 then we have:
 2343

$$\begin{aligned} 2344 \quad a &= M(-w) = \sqrt{A - Bw}, \\ 2345 \quad b &= M(w) = \sqrt{A + Bw}. \end{aligned} \quad (178)$$

2346 Observe that:
 2347

$$\begin{aligned} 2348 \quad b - a &= M(w) - M(-w) = \frac{(A + Bw) - (A - Bw)}{\sqrt{A + Bw} + \sqrt{A - Bw}} \\ 2349 \quad &= \frac{2Bw}{\sqrt{A + Bw} + \sqrt{A - Bw}}, \end{aligned} \quad (179)$$

2350 and then:
 2351

$$\begin{aligned} 2352 \quad f(w) &= \frac{2w}{\tau} \left[1 - \frac{2\kappa}{\sqrt{A + Bw} + \sqrt{A - Bw}} \right]. \end{aligned} \quad (180)$$

2353 Therefore, we have:
 2354

$$\begin{aligned}
2376 \quad & f(w) \geq 0 \Leftrightarrow \sqrt{A+Bw} + \sqrt{A-Bw} \geq 2\kappa \\
2377 \quad & \Leftrightarrow \left(\sqrt{A+Bw} + \sqrt{A-Bw} \right)^2 \geq 4\kappa^2 \\
2378 \quad & \Leftrightarrow 2A + 2\sqrt{A^2 - B^2 w^2} \geq 4\kappa^2 \\
2379 \quad & \Leftrightarrow \sqrt{A^2 - B^2 w^2} \geq 2\kappa^2 - A \\
2380 \quad & \Leftrightarrow \sqrt{A^2 - B^2 w^2} \geq \kappa^2 - \frac{1}{\tau^2} \\
2381 \quad & \Leftrightarrow \sqrt{A^2 - B^2 w^2} \geq \kappa^2 - \frac{1}{\tau^2}
\end{aligned} \tag{181}$$

2386 **Case 1:** $0 < \kappa < \frac{1}{\tau}$.

2387 $\kappa^2 - \frac{1}{\tau^2} < 0$ and the last equation in Eq. (181) holds.

2388 **Case 2:** $0 < \frac{1}{\tau} \leq \kappa$.

2389 The Eq. (181) becomes:

$$\begin{aligned}
2392 \quad & f(w) \geq 0 \Leftrightarrow A^2 - B^2 w^2 \geq \left(\kappa^2 - \frac{1}{\tau^2} \right)^2 \\
2393 \quad & \Leftrightarrow \frac{4\kappa^2}{\tau^2} (1 - w^2) \geq 0 \\
2394 \quad & \Leftrightarrow |w| \leq 1.
\end{aligned} \tag{182}$$

2395 Since $0 < w \leq 1$, $f(w) \geq 0$ holds. According to Eq. (176), we conclude that:

$$0 < w \leq 1 \Rightarrow \mathcal{J}(w) \leq \mathcal{J}(-w). \tag{183}$$

2403 \square

2404 **Lemma 7.** $\forall \nu > 0$, a function $G_3 : \mathbb{R}_0^+ \rightarrow \mathbb{R}$ is defined as:

$$G_3(m) = \log(I_\nu(m)) - \nu \log(m). \tag{184}$$

2405 where I_ν is the modified Bessel function of the first kind of order ν , which is defined as:

$$I_\nu(m) = \sum_{k=0}^{\infty} \frac{1}{k! \Gamma(\nu + k + 1)} \left(\frac{m}{2} \right)^{2k+\nu}. \tag{185}$$

2406 It holds that $G_3(\cdot)$ is a strictly increasing function with $G'_3(\cdot) \in (0, 1)$

2407 *Proof.* The first derivative of G_3 is:

$$G'_3(m) = \frac{I'_\nu(m)}{I_\nu(m)} - \frac{\nu}{m}. \tag{186}$$

2408 According to (Olver, 2010):

$$I'_\nu(m) = I_{\nu+1}(m) + \frac{\nu}{m} I_\nu(m), \tag{187}$$

2409 then we have:

$$\frac{I'_\nu(m)}{I_\nu(m)} - \frac{\nu}{m} = \frac{I_{\nu+1}(m)}{I_\nu(m)}. \tag{188}$$

2410 Plugging Eq. (188) into Eq. (186), we get:

2430
 2431
$$G'_3(m) = \frac{I_{\nu+1}(m)}{I_\nu(m)}. \quad (189)$$

 2432
 2433

2434 Since:

2435
 2436
$$0 < I_{\nu+1}(m) < I_\nu(m), \quad (190)$$

 2437

2438 therefore:

2439
 2440
$$G'_3(m) = \frac{I_{\nu+1}(m)}{I_\nu(m)} \in (0, 1). \quad (191)$$

 2441
 2442

2443 This shows that $G_3(\cdot)$ is a strictly increasing function with $G'_3(\cdot) \in (0, 1)$.

2444 \square

2445 **Lemma 8.** $\forall \nu > 0$, a function $H(\cdot) : R^+ \rightarrow \mathbb{R}$ is defined as:

2446
 2447
$$H(m) = \frac{1}{m} \frac{I_{\nu+1}(m)}{I_\nu(m)}, \quad (192)$$

 2448

2449 where I_ν is the modified Bessel function of the first kind of order ν , which is defined as:

2450
 2451
$$I_\nu(m) = \sum_{k=0}^{\infty} \frac{1}{k! \Gamma(\nu + k + 1)} \left(\frac{m}{2}\right)^{2k+\nu}. \quad (193)$$

 2452
 2453

2454 It holds that $H(m)$ is a strictly decreasing function.

2455 *Proof.* $\forall \nu, m \in R^+$, denote $R_\nu(m)$ as:

2456
 2457
$$R_\nu(m) = \frac{I_{\nu+1}(m)}{I_\nu(m)}. \quad (194)$$

 2458

2459 According to (Olver, 2010), we have:

2460
 2461
$$I'_\nu(m) = I_{\nu+1}(m) + \frac{\nu}{m} I_\nu(m), \quad (195)$$

 2462

2463 then:

2464
 2465
$$\begin{aligned} R'_\nu(m) &= \frac{I'_{\nu+1}(m) I_\nu(m) - I_{\nu+1}(m) I'_\nu(m)}{I_\nu(m)^2} \\ &= \frac{\left(I_{\nu+2}(m) + \frac{\nu+1}{m} I_{\nu+1}(m)\right) I_\nu(m) - I_{\nu+1}(m) \left(I_{\nu+1}(m) + \frac{\nu}{m} I_\nu(m)\right)}{I_\nu(m)^2} \\ &= \frac{I_{\nu+2}(m) I_\nu(m) - I_{\nu+1}^2(m) + \frac{1}{m} I_{\nu+1}(m) I_\nu(m)}{I_\nu(m)^2} \\ &= \frac{I_{\nu+2}(m) I_\nu(m) - I_{\nu+1}^2(m)}{I_\nu(m)^2} + \frac{1}{m} R_\nu(m). \end{aligned} \quad (196)$$

 2466
 2467
 2468
 2469

2470 Since $H(m)$ can be rewritten as:

2471
 2472
$$H(m) = \frac{R_\nu(m)}{m}, \quad (197)$$

 2473
 2474
 2475
 2476
 2477
 2478
 2479

2484 then:
 2485

$$\begin{aligned}
 H'(m) &= \frac{R'_\nu(m)m - R_\nu(m)}{m^2} \\
 &= \frac{1}{m} \left(R'_\nu(m) - \frac{1}{m} R_\nu(m) \right) \\
 &= \frac{1}{m} \left(\frac{I_{\nu+2}(m)I_\nu(m) - I_{\nu+1}^2(m)}{I_\nu(m)^2} \right).
 \end{aligned} \tag{198}$$

2494 According to the Turán type inequalities for modified Bessel functions (Baricz, 2010), when $m > 0$:
 2495

$$\frac{I_{\nu+2}(m)I_\nu(m) - I_{\nu+1}^2(m)}{I_\nu(m)^2} < 0, \tag{199}$$

2499 so

$$H'(m) < 0. \tag{200}$$

2503 Then we can conclude that $H(m)$ is a strictly decreasing function.

2504 \square

2505 **Lemma 9.** $\forall \nu > 0$ and $0 < a < b$, it holds that:

$$\log \left(\frac{I_\nu(a)}{I_\nu(b)} \right) > \nu \log \left(\frac{a}{b} \right) + (a - b), \tag{201}$$

2509 where I_ν is the modified Bessel function of the first kind of order ν , which is defined as:

$$I_\nu(m) = \sum_{k=0}^{\infty} \frac{1}{k! \Gamma(\nu + k + 1)} \left(\frac{m}{2} \right)^{2k+\nu}. \tag{202}$$

2514 *Proof.* According to (Olver, 2010), $\forall x > 0$ and $0 < \nu_1 < \nu_2 < \infty$, we have:

$$I_{\nu_1}(x) > I_{\nu_2}(x). \tag{203}$$

2518 Denote a function L as:

$$L(x) = \log I_\nu(x) - \nu \log(x) - x. \tag{204}$$

2522 According to (Olver, 2010):

$$I'_\nu(m) = I_{\nu+1}(m) + \frac{\nu}{m} I_\nu(m), \tag{205}$$

2526 then we have:

$$\frac{I'_\nu(m)}{I_\nu(m)} - \frac{\nu}{m} = \frac{I_{\nu+1}(m)}{I_\nu(m)}. \tag{206}$$

2531 Taking Eq. (203) and Eq. (206) into account, the derivative of L is:

$$\begin{aligned}
 L'(x) &= \frac{I'_\nu(x)}{I_\nu(x)} - \frac{\nu}{x} - 1 \\
 &= \frac{I_{\nu+1}(x)}{I_\nu(x)} - 1 \\
 &< 0.
 \end{aligned} \tag{207}$$

2538 Therefore, $\forall \nu > 0, 0 < b < a$, it holds that:
 2539

$$\begin{aligned} 2540 \quad \log(I_\nu(a)) - \nu \log(a) - a &= L(a) \\ 2541 \quad &> L(a) \\ 2542 \quad &= \log(I_\nu(b)) - \nu \log(b) - b, \\ 2543 \end{aligned} \tag{208}$$

2544 then we have:
 2545

$$2546 \quad \log\left(\frac{I_\nu(a)}{I_\nu(b)}\right) > \nu \log\left(\frac{a}{b}\right) + (a - b). \tag{209}$$

2549 \square
 2550
 2551
 2552
 2553
 2554
 2555
 2556
 2557
 2558
 2559
 2560
 2561
 2562
 2563
 2564
 2565
 2566
 2567
 2568
 2569
 2570
 2571
 2572
 2573
 2574
 2575
 2576
 2577
 2578
 2579
 2580
 2581
 2582
 2583
 2584
 2585
 2586
 2587
 2588
 2589
 2590
 2591

2592 E.3 DETAILS OF THEOREM 3
2593

2594 In this section, we provide proofs of Theorem 3 that is proposed in Sec. 3.4. We also provide
2595 details and proofs of the auxiliary theorems (Theorem S5 and Theorem S6) and the technical
2596 lemmas (Lemma 10, Lemma 11, Lemma 12 and Lemma 13) that support the proof Theorem 3. For
2597 convenience in reading, let us recall some related notions and definitions.

- 2598 • $h, N \in \mathbb{N}$.
- 2599 • $\mathbb{S}^{h-1} = \{z \in \mathbb{R}^h : \|z\| = 1\}$.
- 2600 • $\mathbb{A} = \{x \in \mathbb{R}^h : n_A \cdot x = 0\}$ where n_A is the normal vector of \mathbb{A} .
- 2601 • $\mathbb{B} = \{y \in \mathbb{R}^h : n_B \cdot y = 0\}$ where n_A is the normal vector of \mathbb{B} .
- 2602 • $\phi = \cos^{-1} \left(\frac{n_x \cdot n_y}{\|n_x\| \cdot \|n_y\|} \right)$ and $0 < \phi_{\min} \leq \phi < \frac{\pi}{2}$.
- 2603 • $\mathbb{S}_X = \mathbb{S}^{h-1} \cap \mathbb{A} = \{x \in \mathbb{R}^h : \|x\| = 1, n_A \cdot x = 0\} \cong S^{h-2} \in \mathbb{S}^{h-1}$.
- 2604 • $\mathbb{S}_Y = \mathbb{S}^{h-1} \cap \mathbb{B} = \{y \in \mathbb{R}^h : \|y\| = 1, n_B \cdot y = 0\} \cong S^{h-2} \in \mathbb{S}^{h-1}$.
- 2605 • $\mathbb{C} = \mathbb{A} \cap \mathbb{B}$.
- 2606 • $h_X = h_Y = h - 1$.
- 2607 • $h_C = h - 2$.
- 2608 • P_A : the projection matrix of \mathbb{A} .
- 2609 • P_B : the projection matrix of \mathbb{B} .
- 2610 • P_C : the projection matrix of \mathbb{C} .
- 2611 • $e_A = \{z \in \mathbb{S}_X : z \perp \mathbb{C}\}$.
- 2612 • $e_B = \{z \in \mathbb{S}_Y : z \perp \mathbb{C}\}$.
- 2613 • $\mathbb{C}^\perp = \text{span}\{e_A\} \oplus \text{span}\{e_B\}$
- 2614 • $\mathbb{R}^h = \mathbb{C} \oplus \mathbb{C}^\perp$.
- 2615 • $X = (x_1, \dots, x_N) \in (\mathbb{S}_X)^N$.
- 2616 • $Y = (y_1, \dots, y_N) \in (\mathbb{S}_Y)^N$.
- 2617 • $\mu_x = \frac{1}{N} \sum_{i=1}^N x_i$.
- 2618 • $\mu_y = \frac{1}{N} \sum_{i=1}^N y_i$.
- 2619 • $c_x = \frac{\mu_x}{\|\mu_x\|}$.
- 2620 • $c_y = \frac{\mu_y}{\|\mu_y\|}$.

2621 **Definition** (Multimodal Contrastive Loss (MCL Loss)). Let (X, Y) be an N -pair configuration,
2622 where $X = (x_1, \dots, x_N) \in (\mathbb{S}^{h-1})^N$ and $Y = (y_1, \dots, y_N) \in (\mathbb{S}^{h-1})^N$. $\forall \tau > 0$, the multimodal
2623 contrastive loss $\mathcal{L}_{\text{MCL}}(\cdot, \cdot) : (\mathbb{S}^{h-1})^N \times (\mathbb{S}^{h-1})^N \rightarrow \mathbb{R}$ is defined as:

$$2624 \mathcal{L}_{\text{MCL}} = \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{\text{MCL}}^i, \text{ where } \mathcal{L}_{\text{MCL}}^i = \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) + \mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}(y_i; X).$$

2625 Here, $\mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}$ is the \mathcal{X} -to- \mathcal{Y} alignment and $\mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}$ is the \mathcal{Y} -to- \mathcal{X} alignment, which are defined
2626 respectively as:

$$2627 \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) = -\log \frac{\exp(x_i \cdot y_i / \tau)}{\sum_{j=1}^N \exp(x_i \cdot y_j / \tau)}, \quad \mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}(y_i; X) = -\log \frac{\exp(x_i \cdot y_i / \tau)}{\sum_{j=1}^N \exp(x_j \cdot y_i / \tau)}.$$

2628 **Definition** (Modality Gap) Let (X, Y) be an N -pair configuration, where $X = (x_1, \dots, x_N) \in$
2629 $(\mathbb{S}^{h-1})^N$ and $Y = (y_1, \dots, y_N) \in (\mathbb{S}^{h-1})^N$. The modality gap between X and Y can be expressed
2630 as the angle between the center representations:

2646
2647
2648

$$\Delta_\theta = \cos^{-1}(c_x \cdot c_y).$$

2649 **Definition** (vMF Distribution). $\forall c \in \mathbb{S}^{h-1}$ and $\kappa \geq 0$, the probability density of a random h -
2650 dimensional unit vector $z \sim \text{vMF}(c, \kappa)$ is given by:
2651

2652
$$f_h(z; c, \kappa) = D_h(\kappa) e^{\kappa c^\top z}, \text{ where } D_h(\kappa) = \frac{\kappa^\nu}{(2\pi)^{\nu+1} I_\nu(\kappa)}.$$

2653
2654

2655 Here, $\nu = h/2 - 1$, and $I_\nu(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ is the modified Bessel function of the first kind of order ν ,
2656 which is defined as:
2657

2658
$$I_\nu(x) = \sum_{k=0}^{\infty} \frac{1}{k! \Gamma(\nu + k + 1)} \left(\frac{x}{2}\right)^{2k+\nu}.$$

2659
2660

2661 **Definition** (Function \tilde{M}). $\forall \kappa, \tau > 0$, a function $\tilde{M}_\kappa(\cdot, \cdot) : [-1, 1] \times [0, 1] \rightarrow \mathbb{R}_0^+$ is defined as:
2662

2663
2664
$$\tilde{M}_\kappa(w, t) = \sqrt{\kappa^2 + \frac{2\kappa w}{\tau} + \frac{t^2}{\tau^2}}.$$

2665

2666 **Definition** (Function $\tilde{\mathcal{J}}$). $\forall \kappa, \nu, \tau > 0$, $\tilde{\mathcal{J}}(\cdot, \cdot, \cdot; \kappa, \nu) : [-1, 1] \times [-1, 1] \times [0, 1] \rightarrow \mathbb{R}$ is defined as:
2667

2668
2669
2670
$$\tilde{\mathcal{J}}(w_1, w_2, t; \kappa, \nu) = -\frac{w_1}{\tau} + \log \left(\frac{I_\nu(\tilde{M}_\kappa(w_2, t))}{\tilde{M}_\kappa(w_2, t)^\nu} \right) - \log \left(\frac{I_\nu(\kappa)}{\kappa^\nu} \right).$$

2671
2672

2673 **Definition** (Function M). $\forall \kappa, \tau > 0$, a function $M_\kappa(\cdot) : [-1, 1] \rightarrow \mathbb{R}_0^+$ is defined as:
2674

2675
2676
$$M_\kappa(w) = \sqrt{\kappa^2 + \frac{2\kappa w}{\tau} + \frac{1}{\tau^2}}$$

2677
2678
$$= \tilde{M}_\kappa(w, 1).$$

2679 **Definition** (Function \mathcal{J}). $\forall \kappa, \nu, \tau > 0$, a function $\mathcal{J}(\cdot; \kappa, \nu) : [-1, 1] \rightarrow \mathbb{R}$ is defined as:
2680

2681
2682
$$\mathcal{J}(w; \kappa, \nu) = -\frac{w}{\tau} + \log \left(\frac{I_\nu(M_\kappa(w))}{M_\kappa(w)^\nu} \right) - \log \left(\frac{I_\nu(\kappa)}{\kappa^\nu} \right)$$

2683
2684
$$= \tilde{\mathcal{J}}(w, w, 1; \kappa, \nu).$$

2685

2686 **Definition** (Function \tilde{M}). $\forall \kappa, \tau > 0$, a function $\tilde{M}_\kappa(\cdot, \cdot) : [-1, 1] \times [0, 1] \rightarrow \mathbb{R}_0^+$ is defined as:
2687

2688
$$\tilde{M}_\kappa(w, t) = \tilde{M}_\kappa(w, t).$$

2689

2690 **Definition** (Function $\hat{\mathcal{J}}$). $\forall \kappa, \nu, \tau > 0$, a function $\hat{\mathcal{J}}(\cdot, \cdot; \kappa, \nu) : [-1, 1] \times [0, 1] \rightarrow \mathbb{R}$ is defined as:
2691

2692
2693
$$\hat{\mathcal{J}}(w, t; \kappa, \nu) = -\frac{w}{\tau} + \log \left(\frac{I_\nu(\tilde{M}_\kappa(w, t))}{\tilde{M}_\kappa(w, t)^\nu} \right) - \log \left(\frac{I_\nu(\kappa)}{\kappa^\nu} \right)$$

2694
2695
$$= \tilde{\mathcal{J}}(w, w, t; \kappa, \nu).$$

2696
2697

2698
2699

2700 E.3.1 PROOF OF THEOREM 3
27012702 In this subsection, we provide the proof of Theorem 3. For convenience in reading, we first restate
2703 Theorem 3 here.2704 **Theorem 3.** [Restate] Let (X, Y) be an N -pair configuration, where $X = (x_1, \dots, x_N) \in (\mathbb{S}_X \setminus \mathbb{C})^N$
2705 are iid samples from $\mu_x = \text{vMF}(c_x, \kappa_x)$, and $Y = (y_1, \dots, y_N) \in (\mathbb{S}_Y \setminus \mathbb{C})^N$ are iid samples from
2706 $\mu_y = \text{vMF}(c_y, \kappa_y)$. Let $\tilde{\nu} = (h-1)/2 - 1$. Suppose there exists an index $i = c$ such that $x_c = c_x$,
2707 $y_c = c_y$. Denote $\Delta_\theta = \cos^{-1}(c_x \cdot c_y)$ and assume that $c_x, c_y \notin \mathbb{C}$ with $c_x \cdot c_y > 0$. For any fixed
2708 $\kappa_x, \kappa_y > 0$, it holds that:

2709
2710
$$\begin{aligned} & \lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}}^c - 2 \log(N) \\ &= \tilde{\mathcal{J}}(\cos(\Delta_\theta), \cos(\Delta_\theta), \|P_B c_x\|; \kappa_y, \tilde{\nu}) + \tilde{\mathcal{J}}(\cos(\Delta_\theta), \cos(\Delta_\theta), \|P_A c_y\|; \kappa_x, \tilde{\nu}) \\ &\geq \tilde{\mathcal{J}}(\cos(\phi_{\min}), \cos(\phi_{\min}), \cos(\phi_{\min}); \kappa_y, \tilde{\nu}) + \tilde{\mathcal{J}}(\cos(\phi_{\min}), \cos(\phi_{\min}), \cos(\phi_{\min}); \kappa_x, \tilde{\nu}), \end{aligned}$$

2711
2712
2713
2714
2715

2716 where equality is attained if and only if there exists a configuration of (X, Y) such that:2717
2718 (A6) $c_x \perp \mathbb{C}$ and $c_y \perp \mathbb{C}$.
2719 (A7) $\Delta_\theta = \cos^{-1}(c_x \cdot c_y) = \phi_{\min}$.
27202721 *Proof.* We first decompose $\lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}}^c - 2 \log(N)$ into two parts:
2722

2723
$$\begin{aligned} \lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}}^c - 2 \log(N) &= \lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(c_x; Y) - \log(N) \\ &+ \lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}(c_y; X) - \log(N). \end{aligned} \tag{210}$$

2724
2725
2726

2727 Set:

2728
2729
$$\begin{aligned} \hat{\mathcal{J}}(w, t; \kappa, \nu) &= \tilde{\mathcal{J}}(w, w, t; \kappa, \nu), \\ \tilde{\nu} &= \tilde{\nu}, \end{aligned} \tag{211}$$

2730
2731

2732 According to Theorem S6, the convergent function and its lower bound of $\mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}$ are:
2733

2734
2735
$$\begin{aligned} \lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(c_x; Y) - \log(N) &= \hat{\mathcal{J}}(\cos(\Delta_\theta), \|P_B c_x\|; \kappa_y, \tilde{\nu}) \\ &\geq \hat{\mathcal{J}}(\|P_A c_y\|, \|P_A c_y\|, \cos(\phi); \kappa_y, \tilde{\nu}). \end{aligned} \tag{212}$$

2736
2737

2738 where equality is attained if and only if there exists a configuration of (X, Y) such that:
27392740 (i) $c_x \perp \mathbb{C}$.
2741 (ii) $c_x = \frac{P_A c_y}{\|P_A c_y\|}$.
2742
27432744 This Theorem also holds for $\mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}$:

2745
2746
$$\begin{aligned} \lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}(c_y; X) - \log(N) &= \hat{\mathcal{J}}(\cos(\Delta_\theta), \|P_A c_y\|; \kappa_x, \tilde{\nu}) \\ &\geq \hat{\mathcal{J}}(\|P_B c_x\|, \|P_B c_x\|, \cos(\phi); \kappa_x, \tilde{\nu}). \end{aligned} \tag{213}$$

2747
2748
2749

2750 where equality is attained if and only if there exists a configuration of (X, Y) such that:
27512752 (iii) $c_y \perp \mathbb{C}$.
2753 (iv) $c_y = \frac{P_B c_x}{\|P_B c_x\|}$.

According to Lemma 13, for some $\lambda_x, \lambda_y > 0$ such that the projections of x and y are collinear with the other vector:

(1) The orthogonal projection of x on \mathbb{B} is a scalar multiple of y :

$$P_B x = \lambda_x y, \quad \lambda_x \neq 0,$$

(2) The orthogonal projection of y on \mathbb{A} is a scalar multiple of x :

$$P_A y = \lambda_y x, \quad \lambda_y \neq 0,$$

if and only if the following condition holds:

(v) Either $x \perp \mathbb{C}$ and $y \perp \mathbb{C}$, or $x = \pm y \in \mathbb{C}$.

Since $c_x, c_y \notin \mathbb{C}$, there is only one configuration in (v) that satisfies (ii) + (iv), that is $c_x \perp \mathbb{C}$ and $c_y \perp \mathbb{C}$. In this case, Lemma 13 shows that:

$$\begin{aligned} \cos(\Delta_\theta) &= \cos(\phi) \geq \cos(\phi_{\min}), \\ \|P_A c_y\| &= \|P_B c_x\| = \cos(\phi), \\ P_B c_x &= \cos(\phi) c_y, \\ P_A c_y &= \cos(\phi) c_x. \end{aligned} \tag{214}$$

Combining Eq. (212), Eq. (213) and Eq. (214), we have:

$$\begin{aligned} \lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}}^c - 2 \log(N) &= \hat{\mathcal{J}}(\cos(\Delta_\theta), \|P_B c_x\|; \kappa_y, \tilde{\nu}) + \hat{\mathcal{J}}(\cos(\Delta_\theta), \|P_A c_y\|; \kappa_x, \tilde{\nu}) \\ &\geq \hat{\mathcal{J}}(\cos(\phi), \cos(\phi); \kappa_y, \tilde{\nu}) + \hat{\mathcal{J}}(\cos(\phi), \cos(\phi); \kappa_x, \tilde{\nu}). \end{aligned} \tag{215}$$

where equality is attained if and only if there exists a configuration of (X, Y) such that:

(A6) $c_x \perp \mathbb{C}$ and $c_y \perp \mathbb{C}$.

Since Lemma 11 shows that $\hat{\mathcal{J}}(\cos(\phi), \cos(\phi); \kappa, \tilde{\nu})$ is a strictly decreasing function of $\cos(\phi)$, we have:

$$\begin{aligned} \lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}}^c - 2 \log(N) &= \hat{\mathcal{J}}(\cos(\Delta_\theta), \|P_B c_x\|; \kappa_y, \tilde{\nu}) + \hat{\mathcal{J}}(\cos(\Delta_\theta), \|P_A c_y\|; \kappa_x, \tilde{\nu}) \\ &\geq \hat{\mathcal{J}}(\cos(\phi), \cos(\phi); \kappa_y, \tilde{\nu}) + \hat{\mathcal{J}}(\cos(\phi), \cos(\phi); \kappa_x, \tilde{\nu}) \\ &\geq \hat{\mathcal{J}}(\cos(\phi_{\min}), \cos(\phi_{\min}); \kappa_y, \tilde{\nu}) + \hat{\mathcal{J}}(\cos(\phi_{\min}), \cos(\phi_{\min}); \kappa_x, \tilde{\nu}), \end{aligned} \tag{216}$$

where equality is attained if and only if there exists a configuration of (X, Y) such that:

(A7) $\Delta_\theta = \cos^{-1}(c_x \cdot c_y) = \phi_{\min}$.

Replacing $\hat{\mathcal{J}}(w, t; \kappa, \nu)$ with $\tilde{\mathcal{J}}(w, w, t; \kappa, \nu)$, we conclude that:

$$\begin{aligned} \lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}}^c - 2 \log(N) &= \tilde{\mathcal{J}}(\cos(\Delta_\theta), \cos(\Delta_\theta), \|P_B c_x\|; \kappa_y, \tilde{\nu}) + \tilde{\mathcal{J}}(\cos(\Delta_\theta), \cos(\Delta_\theta), \|P_A c_y\|; \kappa_x, \tilde{\nu}) \\ &\geq \tilde{\mathcal{J}}(\cos(\Delta_\theta), \cos(\Delta_\theta), \cos(\Delta_\theta); \kappa_y, \tilde{\nu}) + \tilde{\mathcal{J}}(\cos(\Delta_\theta), \cos(\Delta_\theta), \cos(\Delta_\theta); \kappa_x, \tilde{\nu}) \\ &\geq \tilde{\mathcal{J}}(\cos(\phi_{\min}), \cos(\phi_{\min}), \cos(\phi_{\min}); \kappa_y, \tilde{\nu}) + \tilde{\mathcal{J}}(\cos(\phi_{\min}), \cos(\phi_{\min}), \cos(\phi_{\min}); \kappa_x, \tilde{\nu}), \end{aligned} \tag{217}$$

where equality is attained if and only if there exists a configuration of (X, Y) such that:

2808 (A6) $c_x \perp \mathbb{C}$ and $c_y \perp \mathbb{C}$.
 2809
 2810 (A7) $\Delta_\theta = \cos^{-1}(c_x \cdot c_y) = \phi_{\min}$.
 2811
 2812 \square
 2813

2814 E.3.2 AUXILIARY THEOREMS PART 3

2815 In this subsection, we provide details and proofs of the auxiliary theorems (Theorem S5 and Theorem S6) that support the proof of Theorem 3.

2816 **Theorem S5.** Let (X, Y) be an N -pair configuration, where $X = (x_1, \dots, x_N) \in (\mathbb{S}_X \setminus \mathbb{C})^N$ are
 2817 iid samples from $\mu_x = \text{vMF}(c_x, \kappa_x)$, and $Y = (y_1, \dots, y_N) \in (\mathbb{S}_Y \setminus \mathbb{C})^N$ are iid samples from
 2818 $\mu_y = \text{vMF}(c_y, \kappa_y)$. Let $\tilde{\nu} = (h-1)/2 - 1$ and $\kappa_y > 0$.

2819 $\forall x_i \in X$, denote $w_i = x_i \cdot y_i$ and $w_{x_i, c_y} = x_i \cdot c_y$. It holds that:

$$\begin{aligned} \lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) - \log(N) &= \lim_{N \rightarrow \infty} -\log \frac{\exp(x_i \cdot y_i / \tau)}{\sum_{j=1}^N \exp(x_i \cdot y_j / \tau)} - \log(N) \\ &= -\frac{w_i}{\tau} + \log \left(\frac{I_{\tilde{\nu}}(\tilde{M}_{\kappa_y}(w_{x_i, c_y}, \|P_B x_i\|))}{\tilde{M}_{\kappa_y}(w_{x_i, c_y}, \|P_B x_i\|)^{\tilde{\nu}}} \right) - \log \left(\frac{I_{\tilde{\nu}}(\kappa_y)}{\kappa_y^{\tilde{\nu}}} \right) \quad (218) \\ &= \tilde{\mathcal{J}}(w_i, w_{x_i, c_y}, \|P_B x_i\|; \kappa, \tilde{\nu}), \end{aligned}$$

2820 where $\forall \kappa, \tau > 0$, $\tilde{M}_\kappa(\cdot, \cdot) : [-1, 1] \times [0, 1] \rightarrow \mathbb{R}_0^+$ is defined as:

$$\tilde{M}_\kappa(w, t) = \sqrt{\kappa^2 + \frac{2\kappa w}{\tau} + \frac{t^2}{\tau^2}}, \quad (219)$$

2821 and I_ν is the modified Bessel function of the first kind of order ν , which is defined as:

$$I_\nu(m) = \sum_{k=0}^{\infty} \frac{1}{k! \Gamma(\nu + k + 1)} \left(\frac{m}{2}\right)^{2k+\nu}. \quad (220)$$

2822 Suppose there exists an index $i = c$ such that $x_c = c_x$, $y_c = c_y$. Denote $w_c = c_x \cdot c_y$. It holds that:

$$\begin{aligned} \lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(c_x; Y) - \log(N) &= -\frac{w_c}{\tau} + \log \left(\frac{I_{\tilde{\nu}}(\tilde{M}_{\kappa_y}(w_c, \|P_B c_x\|))}{\tilde{M}_{\kappa_y}(w_c, \|P_B c_x\|)^{\tilde{\nu}}} \right) - \log \left(\frac{I_{\tilde{\nu}}(\kappa_y)}{\kappa_y^{\tilde{\nu}}} \right) \\ &= \hat{\mathcal{J}}(w_c, \|P_B c_x\|; \kappa_y, \tilde{\nu}) = \tilde{\mathcal{J}}(w_c, w_c, \|P_B x_i\|; \kappa, \tilde{\nu}). \quad (221) \end{aligned}$$

2823 *Proof. Step 1:* We start the proof by find the convergent function of $\mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y)$ as $N \rightarrow \infty$.
 2824 Same with Eq. (39) of Theorem S1, $\forall x_i \in X$, the \mathcal{X} -to- \mathcal{Y} alignment of x_i can be rewritten as:

$$\begin{aligned} \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) &= -\log \frac{\exp(x_i \cdot y_i / \tau)}{\sum_j \exp(x_i \cdot y_j / \tau)} \\ &= -\frac{x_i \cdot y_i}{\tau} + \log \left(N \frac{1}{N} \sum_{j=1}^N \exp\left(\frac{x_i \cdot y_j}{\tau}\right) \right) \\ &= -\frac{x_i \cdot y_i}{\tau} + \log \left(\frac{1}{N} \sum_{j=1}^N \exp\left(\frac{x_i \cdot y_j}{\tau}\right) \right) + \log(N). \quad (222) \end{aligned}$$

2862 Lemma 2 shows that:
 2863

$$2864 \quad \lim_{N \rightarrow \infty} \log \left(\frac{1}{N} \sum_{j=1}^N \exp \left(\frac{x_i \cdot y_j}{\tau} \right) \right) = \log \left(\mathbb{E}_{y \sim \mu_y} \left[\exp \left(\frac{x_i \cdot y}{\tau} \right) \right] \right). \quad (223)$$

2868 According to the moment-generating function of the vMF distribution:
 2869

$$2870 \quad \mathbb{E}_{y \sim \mu_y} \left[\exp \left(\frac{x_i \cdot y}{\tau} \right) \right] = \mathbb{E}_{y \sim \mu_y} \left[\exp \left(\frac{x_i}{\tau} \cdot y \right) \right] = \frac{I_{\tilde{\nu}}(\tilde{\kappa}'_y)}{I_{\tilde{\nu}}(\kappa_y)} \left(\frac{\kappa_y}{\tilde{\kappa}'_y} \right)^{\tilde{\nu}}, \quad (224)$$

$$2874 \quad \text{where } \tilde{\kappa}'_y = \|\kappa_y c_y + \frac{P_B x_i}{\tau}\|_2.$$

2875 Then we have:
 2876

$$2878 \quad \lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) - \log(N) = -\frac{x_i \cdot y_i}{\tau} + \log \left(\frac{I_{\tilde{\nu}}(\tilde{\kappa}'_y)}{\tilde{\kappa}'_y^{\tilde{\nu}}} \right) - \log \left(\frac{I_{\tilde{\nu}}(\kappa_y)}{\kappa_y^{\tilde{\nu}}} \right). \quad (225)$$

2881 **Step 2:** we will transform $\mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}$ from a function of vectors to a function of angles between vectors.
 2882

2883 Without loss of generality, we assume the coordinate of c_y as
 2884

$$2885 \quad c_y = (1, 0, \dots, 0), \quad (226)$$

2887 the hyperplane \mathbb{B} as:
 2888

$$2889 \quad \mathbb{B} = \{x \in \mathbb{R}^h : n_A \cdot x = 0\}, \quad \text{where } n_B = (0, 0, \dots, 1). \quad (227)$$

2891 Let $\hat{x}_i = P_B x_i$, then we have:
 2892

$$2893 \quad \cos(\theta_{x_i, c_y}) = x_i \cdot c_y = P_B x_i \cdot c_y = \hat{x}_i \cdot c_y. \quad (228)$$

2895 Define:
 2896

$$2898 \quad \cos(\hat{\theta}_{x_i, c_y}) = \frac{\hat{x}_i}{\|\hat{x}_i\|} \cdot c_y = \frac{P_B x_i}{\|P_B x_i\|} \cdot c_y, \quad (229)$$

2900 then we have:
 2901

$$2903 \quad \|P_B x_i\| \cos(\hat{\theta}_{x_i, c_y}) = P_B x_i \cdot c_y = \cos(\theta_{x_i, c_y}). \quad (230)$$

2905 And \hat{x}_i can be represented as:
 2906

$$2908 \quad \hat{x}_i = \|P_B x_i\| \left(\cos(\hat{\theta}_{x_i, c_y}), u \sin(\hat{\theta}_{x_i, c_y}) \right)$$

$$2910 \quad = \|P_B x_i\| \left(\cos(\hat{\theta}_{x_i, c_y}), u_2 \sin(\hat{\theta}_{x_i, c_y}), u_3 \sin(\hat{\theta}_{x_i, c_y}), \dots, u_{h-1} \sin(\hat{\theta}_{x_i, c_y}), 0 \right), \quad (231)$$

2912 where $u = (0, u_2, u_3, \dots, u_{h-1}, 0) \cong \mathbb{S}^{h-3} \in \mathbb{S}^{h-1}$ is a unit vector orthogonal to the first and the
 2913 last axes with:
 2914

$$2915 \quad \|u\| = 0 + u_2^2 + u_3^2 + \dots + u_{h-1}^2 + 0 = 1. \quad (232)$$

According to Eq. (226), Eq. (231) and Eq. (232), $\tilde{\kappa}'_y$ (in Eq. (224)) can be re-written as:

$$\begin{aligned}
 \tilde{\kappa}'_y &= \left\| \kappa_y c_y + \frac{x_i}{\tau} \right\|_2 \\
 &= \sqrt{\left(\kappa_y + \frac{\|P_B x_i\| \cos(\hat{\theta}_{x_i, c_y})}{\tau} \right)^2 + \sum_{i=2}^{h-1} \left(\frac{\|P_B x_i\| \sin(\hat{\theta}_{x_i, c_y}) u_i}{\tau} \right)^2} \\
 &= \sqrt{\left(\kappa_y + \frac{\|P_B x_i\| \cos(\hat{\theta}_{x_i, c_y})}{\tau} \right)^2 + \frac{\|P_B x_i\|^2 \sin^2(\hat{\theta}_{x_i, c_y})}{\tau^2}} \\
 &= \sqrt{\kappa_y^2 + \frac{2\kappa_y \|P_B x_i\| \cos(\hat{\theta}_{x_i, c_y})}{\tau} + \frac{\|P_B x_i\|^2}{\tau^2}} \\
 &= \sqrt{\kappa_y^2 + \frac{2\kappa_y \cos(\theta_{x_i, c_y})}{\tau} + \frac{\|P_B x_i\|^2}{\tau^2}} \\
 &= \tilde{M}_{\kappa_y}(\cos(\theta_{x_i, c_y}), \|P_B x_i\|).
 \end{aligned} \tag{233}$$

Consider that $w_i = x_i \cdot y_i$, $w_{x_i, c_y} = \cos(\theta_{x_i, c_y}) = x_i \cdot c_y$, putting Eq. (225) and Eq. (233) together, we have:

$$\begin{aligned}
 \lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) - \log(N) &= -\frac{x_i \cdot y_i}{\tau} + \log \left(\frac{I_{\tilde{\nu}}(\tilde{\kappa}'_y)}{\tilde{\kappa}'_y^{\tilde{\nu}}} \right) - \log \left(\frac{I_{\tilde{\nu}}(\kappa_y)}{\kappa_y^{\tilde{\nu}}} \right) \\
 &= -\frac{w_i}{\tau} + \log \left(\frac{I_{\tilde{\nu}}(\tilde{M}_{\kappa_y}(w_{x_i, c_y}, \|P_B x_i\|))}{\tilde{M}_{\kappa_y}(w_{x_i, c_y}, \|P_B x_i\|)^{\tilde{\nu}}} \right) - \log \left(\frac{I_{\tilde{\nu}}(\kappa_y)}{\kappa_y^{\tilde{\nu}}} \right) \\
 &= \tilde{\mathcal{J}}(w_i, w_{x_i, c_y}, \|P_B x_i\|; \kappa, \tilde{\nu}).
 \end{aligned} \tag{234}$$

When there exists a data pair $i = c$ such that $x_c = c_x$, $y_c = c_y$, $w_i = w_{x_i, c_y} = w_c$, then we have:

$$\begin{aligned}
 \lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(c_x; Y) - \log(N) &= -\frac{w_c}{\tau} + \log \left(\frac{I_{\tilde{\nu}}(\tilde{M}_{\kappa_y}(w_c, \|P_B c_x\|))}{\tilde{M}_{\kappa_y}(w_c, \|P_B c_x\|)^{\tilde{\nu}}} \right) - \log \left(\frac{I_{\tilde{\nu}}(\kappa_y)}{\kappa_y^{\tilde{\nu}}} \right) \\
 &= \tilde{\mathcal{J}}(w_c, \|P_B c_x\|; \kappa_y, \tilde{\nu}) = \tilde{\mathcal{J}}(w_c, w_c, \|P_B x_i\|; \kappa, \tilde{\nu}).
 \end{aligned} \tag{235}$$

□

Theorem S6. Let (X, Y) be an N -pair configuration, where $X = (x_1, \dots, x_N) \in (\mathbb{S}_X \setminus \mathbb{C})^N$ are iid samples from $\mu_x = \text{vMF}(c_x, \kappa_x)$, and $Y = (y_1, \dots, y_N) \in (\mathbb{S}_Y \setminus \mathbb{C})^N$ are iid samples from $\mu_y = \text{vMF}(c_y, \kappa_y)$. Let $\tilde{\nu} = (h-1)/2 - 1$. Suppose there exists an index $i = c$ such that $x_c = c_x$, $y_c = c_y$. Denote $\Delta_\theta = \cos^{-1}(c_x \cdot c_y)$ and assume that $c_x, c_y \notin \mathbb{C}$ with $c_x \cdot c_y > 0$. For any fixed $\kappa_x, \kappa_y > 0$ and $\forall \phi \in [0, \frac{\pi}{2}]$, it holds that:

$$\lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(c_x; Y) - \log(N) = \hat{\mathcal{J}}(w_c, \|P_B c_x\|; \kappa_y, \tilde{\nu}) \geq \hat{\mathcal{J}}(\|P_A c_y\|, \cos(\phi); \kappa_y, \tilde{\nu}), \tag{236}$$

where equality is attained if and only if there exists a configuration of (X, Y) such that:

(B4) $c_x \perp \mathbb{C}$.

2970 (B5) $c_x = \frac{P_A c_y}{\|P_A c_y\|}$.
 2971

2972 **Proof. Step 1:** Similarly to the proof of Theorem S4 in Sec. E.2.2, we start the proof by finding
 2973 the convergent function of $\mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(c_x; Y)$ as $N \rightarrow \infty$. Denote $w_c = c_x \cdot c_y$. $\forall \kappa_y > 0$, as proven
 2974 in Theorem S5:

2975
 2976
 2977
$$\lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(c_x; Y) - \log(N) = \lim_{N \rightarrow \infty} -\log \frac{\exp(c_x \cdot c_y / \tau)}{\sum_{j=1}^N \exp(c_x \cdot y_j / \tau)} - \log(N)$$

 2978
 2979
 2980
$$= -\frac{w_c}{\tau} + \log \left(\frac{I_{\tilde{\nu}}(\tilde{M}_{\kappa_y}(w_c, \|P_B c_x\|))}{\tilde{M}_{\kappa_y}(w_c, \|P_B c_x\|)^{\tilde{\nu}}} \right) - \log \left(\frac{I_{\tilde{\nu}}(\kappa_y)}{\kappa_y^{\tilde{\nu}}} \right)$$

 2981
 2982
 2983
 2984
 2985
 2986
$$= \hat{\mathcal{J}}(w_c, \|P_B c_x\|; \kappa_y, \tilde{\nu}), \quad (237)$$

 2987

2988 where $\forall \kappa, \tau > 0$, $\hat{\mathcal{J}}(\cdot, \cdot; \kappa, \tilde{\nu})$ is a function on $[-1, 1] \times [0, 1]$ and $\tilde{M}_{\kappa}(\cdot, \cdot) : [-1, 1] \times [0, 1] \rightarrow \mathbb{R}_0^+$
 2989 is defined as:

2990
 2991
$$\tilde{M}_{\kappa}(w, t) = \sqrt{\kappa^2 + \frac{2\kappa w}{\tau} + \frac{t^2}{\tau^2}}. \quad (238)$$

 2992

2993 and I_{ν} is the modified Bessel function of the first kind of order ν , which is defined as:

2994
 2995
$$I_{\nu}(m) = \sum_{k=0}^{\infty} \frac{1}{k! \Gamma(\nu + k + 1)} \left(\frac{m}{2} \right)^{2k+\nu}. \quad (239)$$

 2996

2997 **Step 2:** Next, we find the minimal value and the optimal condition of convergent function.

2998 $\forall c_x \in \mathbb{S}_X, \phi \in [0, \frac{\pi}{2}]$ it holds that:

3000
 3001
$$0 \leq \cos(\phi) \leq \|P_B c_x\| \leq 1. \quad (240)$$

 3002

3003 As shown in Lemma 10, $\forall w_c \in [0, 1]$, $\hat{\mathcal{J}}(w = w_c, t; \kappa_y, \tilde{\nu})$ is a strictly increasing function of t on
 3004 $(0, 1]$. Therefore, it holds that:

3005
 3006
$$\hat{\mathcal{J}}(w_c, \cos(\phi); \kappa_y, \tilde{\nu}) \leq \hat{\mathcal{J}}(w_c, \|P_B c_x\|; \kappa_y, \tilde{\nu}) \leq \hat{\mathcal{J}}(w_c, 1; \kappa_y, \tilde{\nu}). \quad (241)$$

 3007

3008 where equality in the above chain holds if and only if the following conditions are satisfied:

3009
 3010 (i) The first inequality becomes equality: $c_x \perp \mathbb{C}$.
 3011
 3012 (ii) The second inequality becomes equality: $c_x \in \mathbb{C}$.
 3013

3014 According to Lemma 5 (set $s = \cos(\phi)$), $\hat{\mathcal{J}}(w_c, \cos(\phi); \kappa_y, \tilde{\nu})$ is a strictly decreasing function on
 3015 w_c when $w_c \geq 0$. Also, Lemma 12 shows that:

3016
 3017
$$-\|P_A c_y\| \leq w_c \leq \|P_A c_y\|, \quad (242)$$

 3018

3019 where
 3020

3021
 3022
$$0 \leq \cos(\phi) < \|P_A c_y\| < 1. \quad (243)$$

 3023

Therefore, when $0 \leq w_c \leq \|P_A c_y\|$, it holds that:

3024
3025
3026

$$\hat{\mathcal{J}}(w_c, \cos(\phi); \kappa_y, \tilde{\nu}) \geq \hat{\mathcal{J}}(\|P_A c_y\|, \cos(\phi); \kappa_y, \tilde{\nu}), \quad (244)$$

3027 where equality is attained if and only if there exists a configuration of (X, Y) such that:3028
3029
3030
3031

$$(iii) \quad c_x = \frac{P_A c_y}{\|P_A c_y\|}.$$

3032 Combining Eq. (237), Eq. (241) and Eq. (244), we conclude:

3033
3034
3035
3036

$$\lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(c_x; Y) - \log(N) = \hat{\mathcal{J}}(w_c, \|P_B c_x\|; \kappa_y, \tilde{\nu}) \geq \hat{\mathcal{J}}(\|P_A c_y\|, \cos(\phi); \kappa_y, \tilde{\nu}), \quad (245)$$

3037 and equality is attained if and only if there exists a configuration of (X, Y) such that:3038
3039
3040
3041
3042
3043
3044

$$(B4) \quad c_x \perp \mathbb{C}.$$

$$(B5) \quad c_x = \frac{P_A c_y}{\|P_A c_y\|}.$$

□

3045 E.3.3 TECHNICAL LEMMAS PART 3

3046

3047 In this subsection, we provide details and proofs of technical lemmas (Lemma 10, Lemma 11, 3048 Lemma 12 and Lemma 13) that support the proof of Theorem 3, Theorem S5 and Theorem S6.

3049 **Lemma 10.** $\forall \kappa, \nu, \tau > 0$ and $w_c \in [0, 1]$, a function $\hat{\mathcal{J}}_{t=s}(\cdot; \kappa, \nu) : (0, 1] \rightarrow \mathbb{R}$ is defined as:3050
3051
3052
3053
3054

$$\begin{aligned} \hat{\mathcal{J}}_{w=w_s}(t; \kappa, \nu) &= -\frac{w_s}{\tau} + \log \left(\frac{I_\nu \left(\tilde{M}_{w=w_s}(t) \right)}{\tilde{M}_\kappa(t)^\nu} \right) - \log \left(\frac{I_\nu(\kappa)}{\kappa^\nu} \right) \\ &= \hat{\mathcal{J}}(w=w_s, t; \kappa, \nu) = \tilde{J}(w=w_s, w=w_s, t; \kappa, \nu), \end{aligned} \quad (246)$$

3055 where $\tilde{M}_\kappa(\cdot) : (0, 1] \rightarrow \mathbb{R}^+$ is defined as:3056
3057
3058
3059
3060
3061

$$\tilde{M}_{w=w_s}(t) = \sqrt{\kappa^2 + \frac{2\kappa w_s}{\tau} + \frac{t^2}{\tau^2}} = \tilde{M}_\kappa(w=w_s, t), \quad (247)$$

3062 and I_ν is the modified Bessel function of the first kind of order ν , which is defined as:3063
3064
3065
3066
3067

$$I_\nu(m) = \sum_{k=0}^{\infty} \frac{1}{k! \Gamma(\nu + k + 1)} \left(\frac{m}{2} \right)^{2k+\nu}. \quad (248)$$

3068 It holds that, for any fixed w_s , $\hat{\mathcal{J}}_{w=w_s}(\cdot)$ is a strictly increasing function on $(0, 1]$.3069
3070
30713070 *Proof.* Let us first decompose the function \mathcal{J} . Denote a constant and a function C_1 and $G_2(t)$ as:3072
3073
3074
3075
3076
3077

$$\begin{aligned} C_1 &= -\frac{w_s}{\tau}, \\ G_3(m) &= \log(I_\nu(m)) - \nu \log(m), \\ G_2(t) &= G_3(\tilde{M}_{w=w_s}(t)) \\ &= \log(I_\nu(\tilde{M}_{w=w_s}(t))) - \nu \log(\tilde{M}_{w=w_s}(t)). \end{aligned} \quad (249)$$

3078 Denote the function $G(t)$ and the constant C as:
 3079

$$3080 \quad G(t) = C_1 + G_2(t), \\ 3081 \quad C = -\log\left(\frac{I_\nu(\kappa)}{\kappa^\nu}\right). \quad (250)$$

3084 Then the function $\hat{\mathcal{J}}_{w=w_s}$ can be written as:
 3085

$$3086 \quad \hat{\mathcal{J}}_{w=w_s}(t; \kappa, \nu) = -\frac{w_s}{\tau} + \log\left(\frac{I_\nu(\tilde{M}_{w=w_s}(t))}{\tilde{M}_{w=w_s}(t)^\nu}\right) - \log\left(\frac{I_\nu(\kappa)}{\kappa^\nu}\right) \\ 3087 \quad = G(t) + C. \quad (251)$$

3092 Now, we investigate derivatives of $\hat{\mathcal{J}}_{w=w_s}$.
 3093

3094 According to Lemma 7, the first derivative of $G_3(m)$ is:
 3095

$$3096 \quad G'_3(m) = \frac{I_{\nu+1}(m)}{I_\nu(m)} \in (0, 1). \quad (252)$$

3098 The derivative of $\tilde{M}_{w=w_s}$ is:
 3099

$$3100 \quad \tilde{M}'_{w=w_s}(t) = \frac{d}{dt} \left(\kappa^2 + \frac{2\kappa w_s}{\tau} + \frac{t^2}{\tau^2} \right)^{1/2} \\ 3101 \quad = \frac{1}{2} \left(\kappa^2 + \frac{2\kappa w_s}{\tau} + \frac{t^2}{\tau^2} \right)^{-1/2} \cdot 2 \frac{t}{\tau^2} \\ 3102 \quad = \frac{t}{\tau^2} \frac{1}{\tilde{M}_{w=w_s}(t)} \\ 3103 \quad > 0. \quad (253)$$

3104 Then, the first derivative of G_2 is:
 3105

$$3106 \quad G'_2(t) = G'_3(\tilde{M}_{w=w_s}(t)) \tilde{M}'_{w=w_s}(t) \\ 3107 \quad = \frac{I_{\nu+1}(\tilde{M}_{w=w_s}(t))}{I_\nu(\tilde{M}_{w=w_s}(t))} \tilde{M}'_{w=w_s}(t) \\ 3108 \quad = \frac{t}{\tau^2} \frac{1}{\tilde{M}_{w=w_s}(t)} \frac{I_{\nu+1}(\tilde{M}_{w=w_s}(t))}{I_\nu(\tilde{M}_{w=w_s}(t))} \\ 3109 \quad > 0. \quad (254)$$

3110 Therefore, we have:
 3111

$$3112 \quad \hat{\mathcal{J}}'_{w=w_s}(t; \kappa, \nu) = G'(t) = G'_2(t) \\ 3113 \quad = \frac{t}{\tau^2} \frac{1}{\tilde{M}_{w=w_s}(t)} \frac{I_{\nu+1}(\tilde{M}_{w=w_s}(t))}{I_\nu(\tilde{M}_{w=w_s}(t))} \\ 3114 \quad > 0. \quad (255)$$

3115 So we can conclude that, for any fixed w_s , $\hat{\mathcal{J}}_{w=w_s}(\cdot)$ is a strictly increasing function on $(0, 1]$.
 3116

3132

3133

3134 **Lemma 11.** $\forall \kappa, \nu, \tau > 0$, a function $\hat{\mathcal{J}}(\cdot; \kappa, \nu) : [-1, 1] \rightarrow \mathbb{R}$ is defined as:

3135

3136

3137
$$\hat{\mathcal{J}}_{t=w}(w; \kappa, \nu) = -\frac{w}{\tau} + \log \left(\frac{I_\nu(\tilde{M}_{t=w}(w))}{\tilde{M}_{t=w}(w)^\nu} \right) - \log \left(\frac{I_\nu(\kappa)}{\kappa^\nu} \right) \quad (256)$$

3138

3139

3140
$$= \hat{\mathcal{J}}(w, t = w; \kappa, \nu) = \tilde{J}(w, w, t = w; \kappa, \nu),$$

3141

3142

3143 where $\tilde{M}_{t=w}(\cdot) : [-1, 1] \rightarrow \mathbb{R}^+$ is defined as:

3144

3145

3146
$$\tilde{M}_{t=w}(w) = \sqrt{\kappa^2 + \frac{2\kappa w}{\tau} + \frac{w^2}{\tau^2}} = |\kappa + \frac{w}{\tau}| = \tilde{M}_\kappa(w, t = w), \quad (257)$$

3147

3148

3149 and I_ν is the modified Bessel function of the first kind of order ν , which is defined as:

3150

3151

3152
$$I_\nu(m) = \sum_{k=0}^{\infty} \frac{1}{k! \Gamma(\nu + k + 1)} \left(\frac{m}{2} \right)^{2k+\nu}. \quad (258)$$

3153

3154

3155 It holds that $\hat{\mathcal{J}}_{t=w}(\cdot)$ is a strictly decreasing function when $w \in [0, 1]$.

3156

3157

3158 *Proof.* Let us first decompose the function $\hat{\mathcal{J}}_{t=w}$. Denote the functions $G_1(w)$ and $G_2(w)$ as:

3159

3160

3161

3162

3163

3164

3165

3166

3167

3168

3169

3170

3171

3172

3173

3174

3175

3176

3177

3178

3179

3180

3181

3182

3183

3184

3185

3165 Denote the function $G(w)$ and the constant C as:

3166
$$G(w) = G_1(w) + G_2(w), \quad (260)$$

3167

3168

3169

3170

3171

3172

3173

3174

3175

3176

3177

3178

3179

3180

3181

3182

3183

3184

3185

3171 Then the function $\hat{\mathcal{J}}_{t=w}$ can be written as:

3172
$$\hat{\mathcal{J}}_{t=w}(w; \kappa, \nu) = -\frac{w}{\tau} + \log \left(\frac{I_\nu(\tilde{M}_{t=w}(w))}{\tilde{M}_{t=w}(w)^\nu} \right) - \log \left(\frac{I_\nu(\kappa)}{\kappa^\nu} \right) \quad (261)$$

3173

3174

3175

3176

3177

3178

3179

3180

3181

3182

3183

3184

3185

3179 Now, we investigate derivatives of $\hat{\mathcal{J}}_{t=w}$.

3180 The first derivative of G_1 is:

3181
$$G'_1(w) = -\frac{1}{\tau} < 0. \quad (262)$$

3182

3183

3184

3185

3185 According to Lemma 7, the first derivative of $G_3(m)$ is:

59

3186

$$3187 \quad G'_3(m) = \frac{I_{\nu+1}(m)}{I_{\nu}(m)} \in (0, 1). \quad (263)$$

3188

3189 When $w \in [0, 1]$, the derivative of $\tilde{M}_{t=w}$ is:

3190

$$3191 \quad \tilde{M}'_{t=w}(w) = \frac{1}{\tau}. \quad (264)$$

3192

3193 Then, the first derivative of G_2 is:

3194

$$3195 \quad G'_2(w) = G'_3(\tilde{M}_{t=w}(w)) \tilde{M}'_{t=w}(w)$$

$$3196 \quad = \frac{I_{\nu+1}(\tilde{M}_{t=w}(w))}{I_{\nu}(\tilde{M}_{t=w}(w))} \tilde{M}'_{t=w}(w)$$

$$3197 \quad = \frac{1}{\tau} \frac{I_{\nu+1}(\tilde{M}_{t=w}(w))}{I_{\nu}(\tilde{M}_{t=w}(w))}. \quad (265)$$

3198

3199 Combining Eq. (262), Eq. (263) and Eq. (265), we have:

3200

$$3201 \quad \hat{\mathcal{J}}'_{t=w}(w; \kappa, \nu) = G'(w)$$

$$3202 \quad = -\frac{1}{\tau} + \frac{1}{\tau} \frac{I_{\nu+1}(\tilde{M}_{t=w}(w))}{I_{\nu}(\tilde{M}_{t=w}(w))} = \frac{1}{\tau} \left(-1 + \frac{I_{\nu+1}(\tilde{M}_{t=w}(w))}{I_{\nu}(\tilde{M}_{t=w}(w))} \right)$$

$$3203 \quad < 0.$$

3204

3205 So we can conclude that $\hat{\mathcal{J}}_{t=w}(\cdot)$ is a strictly decreasing function on $[0, 1]$.

3206

□

3207

3208 **Lemma 12.** Let $h \geq 3$ and $\mathbb{A}, \mathbb{B} \in \mathbb{R}^h$ be two distinct $(h-1)$ -dimensional linear subspaces, with
3209 n_A, n_B being normal vectors and P_A, P_B being the orthogonal projectors on \mathbb{A} and \mathbb{B} , respectively.
3210 Denote $\phi = \cos^{-1} \left(\frac{n_A \cdot n_B}{\|n_A\| \cdot \|n_B\|} \right) \in (0, \frac{\pi}{2})$ as the angle between \mathbb{A} and \mathbb{B} . Let $\mathbb{C} = \mathbb{A} \cap \mathbb{B}$ be an
3211 $(h-2)$ -dimensional linear subspaces. For each fixed $x \in \mathbb{S}_X = \mathbb{A} \cap \mathbb{S}^{h-1}$, $\forall y \in \mathbb{S}_Y = \mathbb{B} \cap \mathbb{S}^{h-1}$,
3212 set $w = x \cdot y$, it holds that:

3213

3214

3215

3216

3217

3218

3219

$$3220 \quad -\|P_B \cdot x\| \leq w \leq \|P_B \cdot x\|, \quad (267)$$

3221

3222

3223

3224

3225 and equalities (extreme values) are attained if and only if the following conditions hold:

3226

3227

3228

3229

3230

3231

3232

3233

3234

3235

$$(C1) \quad w = \|P_B \cdot x\| \Leftrightarrow y = \frac{P_B \cdot x}{\|P_B \cdot x\|}.$$

$$(C2) \quad w = -\|P_B \cdot x\| \Leftrightarrow y = -\frac{P_B \cdot x}{\|P_B \cdot x\|}.$$

3236

3237

3238

3239

3235 **Proof. Step 1:** First, let us decompose the embedding space. Define two vectors e_A and e_B such
3236 that:

3237

3238

3239

$$3236 \quad e_A \in \mathbb{S}_X, \text{ and } e_A \perp \mathbb{C}, \quad (268)$$

$$3237 \quad e_B \in \mathbb{S}_Y, \text{ and } e_B \perp \mathbb{C}.$$

3238 Let \mathbb{C}^\perp be the 2-dimensional orthogonal complement of \mathbb{C} , and \mathbb{C}^\perp satisfies:

3240

$$\begin{aligned} \mathbb{C}^\perp &= \text{span}\{e_A\} \oplus \text{span}\{e_B\}, \\ \mathbb{R}^h &= \mathbb{C} \oplus \mathbb{C}^\perp. \end{aligned} \tag{269}$$

3244

3245 Since $n_A, n_B \in \mathbb{C}^\perp$, $n_A \perp e_A$ and $n_B \perp e_B$, we have:

3246

$$\langle e_A, e_B \rangle = \pm \langle n_A, n_B \rangle, \tag{270}$$

3248

3249 and we choose a pair of e_A and e_B such that:

3250

$$\langle e_A, e_B \rangle = \langle n_A, n_B \rangle = \cos(\phi) \in (0, 1). \tag{271}$$

3252

3253 Therefore, $\forall x \in \mathbb{S}_X = \mathbb{A} \cap \mathbb{S}^{h-1}$ and $\forall y \in \mathbb{S}_Y = \mathbb{B} \cap \mathbb{S}^{h-1}$, $\exists u_A, u_B \in \mathbb{C} \cap \mathbb{S}^{h-1}$, such that
3254 $\cos(\theta_A) = x \cdot e_A$ and $\cos(\theta_B) = y \cdot e_B$. And then x and y can be represented as:

3255

$$\begin{aligned} x &= \cos(\theta_A)e_A + \sin(\theta_A)u_A, \\ y &= \cos(\theta_B)e_B + \sin(\theta_B)u_B. \end{aligned} \tag{272}$$

3258

3259 Using orthogonality, we have:

3260

$$\begin{aligned} P_B \cdot e_A &= \langle e_A, e_B \rangle e_B = \cos(\phi) e_B, \\ P_B \cdot u_A &= u_A, \end{aligned} \tag{273}$$

3264

3265 and

3266

$$\begin{aligned} P_A \cdot e_B &= \langle e_A, e_B \rangle e_A = \cos(\phi) e_A, \\ P_A \cdot u_B &= u_B. \end{aligned} \tag{274}$$

3268

3269 Then the projections of (x_i, y_i) are:

3271

$$\begin{aligned} P_B \cdot x &= \cos(\theta_A) \cos(\phi) e_B + \sin(\theta_A) u_A, \\ P_A \cdot y &= \cos(\theta_B) \cos(\phi) e_A + \sin(\theta_B) u_B. \end{aligned} \tag{275}$$

3274

3275 **Step 2:** Next, we can investigate the range of w .

3276

$$\begin{aligned} w &= x \cdot y \\ &= \cos(\theta_A) \cos(\theta_B) e_A e_B + \sin(\theta_A) \sin(\theta_B) u_A u_B \\ &= \cos(\theta_A) \cos(\theta_B) \cos(\phi) + \sin(\theta_A) \sin(\theta_B) u_A u_B. \end{aligned} \tag{276}$$

3281

3282 Since $u_A, u_B \in \mathbb{C}$ and $\|u_A\| = \|u_B\| = 1$, then $\|u_A \cdot u_B\| \leq 1$. Denote $f(\cdot)_{\pm}$ as:

3283

$$f_{\pm}(\theta_B) = \cos(\theta_A) \cos(\theta_B) \cos(\phi) \pm \sin(\theta_A) \sin(\theta_B), \tag{277}$$

3285

3286 then :

3287

$$f_-(\theta_B) \leq w \leq f_+(\theta_B). \tag{278}$$

3289

3290 Now, let us check the extreme values of $f_{\pm}(w)$. First, we find the derivative of $f_{\pm}(w)$:

3291

$$f'_{\pm}(\theta_B) = -\cos(\theta_A) \sin(\theta_B) \cos(\phi) \pm \sin(\theta_A) \cos(\theta_B), \tag{279}$$

3293

then:

3294

3295

3296

3297

3298 and

3299

3300

3301

3302

3303

3304

$$f'_{\pm}(\theta_B) = 0 \Rightarrow \tan(\theta_B) = \pm \frac{\sin(\theta_A)}{\cos(\theta_A) \cos(\phi)}, \quad (280)$$

3305 Denote:

3306

3307

3308

$$w \geq f_{-} \left(\arctan \left(-\frac{\sin(\theta_A)}{\cos(\theta_A) \cos(\phi)} \right) \right) = -\sqrt{\sin^2(\theta_A) + \cos^2(\theta_A) \cos^2(\phi)}, \quad (281)$$

3309

3310

3311

3312

3313

3314

3315

3316

3317

3318

3319

3320

3321

3322

3323

3324

3325

3326

3327

3328

3329

3330

3331

3332

3333

3334

3335

3336

3337

3338

3339

3340

3341

3342

3343

3344

3345

3346

3347

$$\begin{aligned} \cos(\theta_B) &= \frac{\cos(\theta_A) \cos(\phi)}{r}, \\ \sin(\theta_B) &= \frac{\sin(\theta_A)}{r}. \end{aligned} \quad (284)$$

Plugging Eq. (284) into Eq. (275), we get:

$$\begin{aligned} P_B \cdot x &= \cos(\theta_A) \cos(\phi) e_B + \sin(\theta_A) u_A \\ &= r \cos(\theta_B) \cos(\phi) e_B + r \sin(\theta_B) u_B \\ &= ry, \end{aligned} \quad (285)$$

and

$$\|P_B \cdot x\| = \|ry\| = r. \quad (286)$$

Therefore, w reaches its maximum if and only if the following condition holds:

$$(C1) \quad y = \frac{P_B \cdot x}{\|P_B \cdot x\|}.$$

When $\theta_B = \arctan \left(-\frac{\sin(\theta_A)}{\cos(\theta_A) \cos(\phi)} \right)$ and $u_A = -u_B$, w reaches its minimum. At this time:

$$\begin{aligned} \cos(\theta_B) &= -\frac{\cos(\theta_A) \cos(\phi)}{r}, \\ \sin(\theta_B) &= \frac{\sin(\theta_A)}{r}. \end{aligned} \quad (287)$$

Plugging Eq. (287) into Eq. (275), we get:

$$\begin{aligned} P_B \cdot x &= \cos(\theta_A) \cos(\phi) e_B + \sin(\theta_A) u_A \\ &= -r \cos(\theta_B) \cos(\phi) e_B - r \sin(\theta_B) u_B \\ &= -ry. \end{aligned} \quad (288)$$

3348 and

3350
$$\|P_B \cdot x\| = \| - ry \| = r. \quad (289)$$
 3351

3352 Therefore, w reaches its minimum if and only if the following condition holds:

3354 (C2)
$$y = -\frac{P_B \cdot x}{\|P_B \cdot x\|}.$$
 3355

3356 \square 33573358 **Lemma 13.** Let $h \geq 3$ and $\mathbb{A}, \mathbb{B} \in \mathbb{R}^h$ be two distinct $(h-1)$ -dimensional linear subspaces, with
3359 n_A, n_B being normal vectors and P_A, P_B being the orthogonal projectors on \mathbb{A} and \mathbb{B} , respectively.
3360 Denote $\phi = \cos^{-1} \left(\frac{n_A \cdot n_B}{\|n_A\| \cdot \|n_B\|} \right) \in (0, \frac{\pi}{2})$ as the angle between \mathbb{A} and \mathbb{B} . Let $\mathbb{C} = \mathbb{A} \cap \mathbb{B}$ be an
3361 $(h-2)$ -dimensional linear subspaces. For $x \in \mathbb{S}_X = \mathbb{A} \cap \mathbb{S}^{h-1}$, $y \in \mathbb{S}_Y = \mathbb{B} \cap \mathbb{S}^{h-1}$, the projections
3362 of x and y are collinear with the other vector:
33633364 (i) The orthogonal projection of x on \mathbb{B} is a scalar multiple of y :
3365

3366
$$P_B x = \lambda_x y, \quad \lambda_x \neq 0,$$
 3367

3368 (ii) The orthogonal projection of y on \mathbb{A} is a scalar multiple of x :

3369
$$P_A y = \lambda_y x, \quad \lambda_y \neq 0,$$
 3370

3371 if and only if the following conditions holds:
33723373 (C3) Either $x \perp \mathbb{C}$ and $y \perp \mathbb{C}$, or $x = \pm y \in \mathbb{C}$.
33743375 Moreover, in the first case ($x \perp \mathbb{C}$, $y \perp \mathbb{C}$), it holds that:
3376

3377
$$\langle x, y \rangle = \cos(\phi), \quad P_B x = (\cos(\phi)) y, \quad P_A y = (\cos(\phi)) x,$$
 3378

3379 while in the second case ($x = \pm y \in \mathbb{C}$), it holds that:
3380

3381
$$P_B x = x = (\pm 1) y, \quad P_A y = y = (\pm 1) x.$$
 3382

3383 *Proof.* **Step 1:** First, we need to decompose the embedding space. This step is the same with **Step 1**
3384 of Sec. E.3.3. For convenience in reading, we repeat this step here.
33853386 Define two vectors e_A and e_B such that:
3387

3388
$$\begin{aligned} e_A &\in \mathbb{S}_X, \text{ and } e_A \perp \mathbb{C}, \\ e_B &\in \mathbb{S}_Y, \text{ and } e_B \perp \mathbb{C}. \end{aligned} \quad (290)$$
 3389

3390 Let \mathbb{C}^\perp be the 2-dimensional orthogonal complement of \mathbb{C} , and \mathbb{C}^\perp satisfies:
3391

3392
$$\begin{aligned} \mathbb{C}^\perp &= \text{span} \{e_A\} \oplus \text{span} \{e_B\}, \\ \mathbb{R}^h &= \mathbb{C} \oplus \mathbb{C}^\perp. \end{aligned} \quad (291)$$
 3393

3394 Since $n_A, n_B \in \mathbb{C}^\perp$, $n_A \perp e_A$ and $n_B \perp e_B$, we have:
3395

3396
$$\langle e_A, e_B \rangle = \pm \langle n_A, n_B \rangle, \quad (292)$$
 3397

3398 and we choose a pair of e_A and e_B such that:
3399

3400
$$\langle e_A, e_B \rangle = \langle n_A, n_B \rangle = \cos(\phi) \in (0, 1). \quad (293)$$
 3401

3402 Therefore, $\forall x \in \mathbb{S}_X = \mathbb{A} \cap \mathbb{S}^{h-1}$ and $\forall y \in \mathbb{S}_Y = \mathbb{B} \cap \mathbb{S}^{h-1}$, $\exists u_A, u_B \in \mathbb{C} \cap \mathbb{S}^{h-1}$, such that
 3403 $\cos(\theta_A) = x \cdot e_A$ and $\cos(\theta_B) = y \cdot e_B$. And then x and y can be represented as:
 3404

$$\begin{aligned} 3405 \quad x &= \cos(\theta_A)e_A + \sin(\theta_A)u_A, \\ 3406 \quad y &= \cos(\theta_B)e_B + \sin(\theta_B)u_B. \end{aligned} \quad (294)$$

3408 Using orthogonality, we have:
 3409

$$\begin{aligned} 3410 \quad P_B \cdot e_A &= \langle e_A, e_B \rangle e_B = \cos(\phi) e_B, \\ 3411 \quad P_B \cdot u_A &= u_A, \end{aligned} \quad (295)$$

3413 and
 3414

$$\begin{aligned} 3415 \quad P_A \cdot e_B &= \langle e_A, e_B \rangle e_A = \cos(\phi) e_A, \\ 3416 \quad P_A \cdot u_B &= u_B. \end{aligned} \quad (296)$$

3418 Then the projections of (x_i, y_i) are:
 3419

$$\begin{aligned} 3420 \quad P_B \cdot x &= \cos(\theta_A) \cos(\phi) e_B + \sin(\theta_A)u_A, \\ 3421 \quad P_A \cdot y &= \cos(\theta_B) \cos(\phi) e_A + \sin(\theta_B)u_B. \end{aligned} \quad (297)$$

3423 **Step 2:** \Rightarrow Next, we prove the sufficiency. If conditions (i) and (ii) hold, then:
 3424

$$\begin{aligned} 3425 \quad \cos(\theta_A) \cos(\phi) e_B + \sin(\theta_A)u_A &= \lambda_x \cos(\theta_B)e_B + \lambda_x \sin(\theta_B)u_B, \\ 3426 \quad \cos(\theta_B) \cos(\phi) e_A + \sin(\theta_B)u_B &= \lambda_y \cos(\theta_A)e_A + \lambda_y \sin(\theta_A)u_A. \end{aligned} \quad (298)$$

3428 Decompose both equations into \mathbb{C} and \mathbb{C}^\perp . In \mathbb{C} , we get:
 3429

$$\begin{aligned} 3430 \quad \sin(\theta_A)u_A &= \lambda_x \sin(\theta_B)u_B, \\ 3431 \quad \sin(\theta_B)u_B &= \lambda_y \sin(\theta_A)u_A. \end{aligned} \quad (299)$$

3433 and in \mathbb{C}^\perp we get:
 3434

$$\begin{aligned} 3436 \quad \cos(\theta_A) \cos(\phi) e_B &= \lambda_x \cos(\theta_B)e_B, \\ 3437 \quad \cos(\theta_B) \cos(\phi) e_A &= \lambda_y \cos(\theta_A)e_A. \end{aligned} \quad (300)$$

3438 Then it can be concluded from Eq. (299) that:
 3439

$$\begin{aligned} 3441 \quad \sin(\theta_A)u_A &= \lambda_x \lambda_y \sin(\theta_A)u_A, \\ 3442 \quad \sin(\theta_B)u_B &= \lambda_x \lambda_y \sin(\theta_B)u_B. \end{aligned} \quad (301)$$

3444 Eq. (301) leads to two scenarios:
 3445

$$(S1) \quad \lambda_x \lambda_y = 1.$$

$$(S2) \quad \sin(\theta_A) = \sin(\theta_B) = 0.$$

3449 When (S1) holds, multiply two equations in Eq. (300) and we get:
 3450

$$\cos(\theta_A) \cos(\theta_B) \cos^2(\phi) = \cos(\theta_A) \cos(\theta_B). \quad (302)$$

3453 And since:
 3454

$$0 < \cos^2(\phi) < 1, \quad (303)$$

3456 we can conclude that:
 3457

$$\begin{aligned} \cos(\theta_A) &= \cos(\theta_B) = 0, \\ \sin(\theta_A) &= \sin(\theta_B) = \pm 1. \end{aligned} \quad (304)$$

3461 Plugging Eq. (304) into Eq. (299), we get:
 3462

$$\begin{aligned} u_A &= \lambda_x u_B, \\ u_B &= \lambda_y u_A. \end{aligned} \quad (305)$$

3467 Since $\|u_A\| = \|u_B\| = 1$, Eq. (305) $\Rightarrow \lambda_x = \lambda_y = \pm 1 \Rightarrow u_A = \pm u_B$. And according to Eq. (294)
 3468 and we have:
 3469

$$x = \pm y \in \mathbb{C}. \quad (306)$$

3472 We conclude that (S1) $\Rightarrow x = \pm y \in \mathbb{C}$.
 3473

3474 When (S2) holds, we have:
 3475

$$\begin{aligned} \sin(\theta_A) &= \sin(\theta_B) = 0, \\ \cos(\theta_A) &= \cos(\theta_B) = \pm 1. \end{aligned} \quad (307)$$

3479 Plugging Eq. (307) into Eq. (294), we have:
 3480

$$\begin{aligned} x &= \pm e_A \perp \mathbb{C}, \\ y &= \pm e_B \perp \mathbb{C}. \end{aligned} \quad (308)$$

3484 We conclude that (S2) $\Rightarrow x \perp \mathbb{C}$ and $y \perp \mathbb{C}$.
 3485

3486 So the sufficiency is confirmed.
 3487

Step 3: \Leftarrow Last, we prove the necessity. If $x = \pm y \in \mathbb{C}$, then
 3488

$$\begin{aligned} \cos(\theta_A) &= \cos(\theta_B) = 0, \\ \sin(\theta_A) &= \sin(\theta_B) = \pm 1. \end{aligned} \quad (309)$$

3492 and
 3493

$$\begin{aligned} x &= u_A, \\ y &= u_B, \end{aligned} \quad (310)$$

3498 According to Eq. (297) and Eq. (310), we have:
 3499

$$\begin{aligned} P_B \cdot x &= u_A = x = \pm y, \\ P_A \cdot y &= u_B = y = \pm x. \end{aligned} \quad (311)$$

3503 Let $\lambda_x = \lambda_y = \pm 1$, conditions (i) and (ii) hold.
 3504

3505 If $x \perp \mathbb{C}$ and $y \perp \mathbb{C}$, then:
 3506

$$\begin{aligned} \sin(\theta_A) &= \sin(\theta_B) = 0, \\ \cos(\theta_A) &= \cos(\theta_B) = \pm 1. \end{aligned} \quad (312)$$

3509 and
 3510

3510
 3511 $x = \pm e_A,$
 3512 $y = \pm e_B.$ (313)
 3513

3514 According to Eq. (297) and Eq. (313), we have:

3515
 3516 $P_B \cdot x = \pm \cos(\phi) e_B = \pm \cos(\phi) y,$
 3517 $P_A \cdot y = \pm \cos(\phi) e_A = \pm \cos(\phi) x.$ (314)
 3518

3519 Let $\lambda_x = \lambda_y = \pm \cos(\phi)$, conditions (i) and (ii) hold.

3520 Therefore, the necessity is confirmed. □

3521
 3522
 3523
 3524
 3525
 3526
 3527
 3528
 3529
 3530
 3531
 3532
 3533
 3534
 3535
 3536
 3537
 3538
 3539
 3540
 3541
 3542
 3543
 3544
 3545
 3546
 3547
 3548
 3549
 3550
 3551
 3552
 3553
 3554
 3555
 3556
 3557
 3558
 3559
 3560
 3561
 3562
 3563

3564 E.4 DETAILS OF THEOREM 4
3565

3566 In this section, we provide proofs of Theorem 4 that is proposed in Sec. 4.2. We also provide
3567 details and proofs of the auxiliary theorems (Theorem S7 and Theorem S8) and the technical lemmas
3568 (Lemma 14 and Lemma 15) that support the proof Theorem 4. For convenience in reading, let us
3569 recall some related notions and definitions.

- 3570 • $h, N \in \mathbb{N}$.
- 3571 • $\mathbb{S}^{h-1} = \{z \in \mathbb{R}^h : \|z\| = 1\}$.
- 3572 • $\mathbb{A} = \{x \in \mathbb{R}^h : n_A \cdot x = 0\}$ where n_A is the normal vector of \mathbb{A} .
- 3573 • $\mathbb{B} = \{y \in \mathbb{R}^h : n_B \cdot y = 0\}$ where n_B is the normal vector of \mathbb{B} .
- 3574 • $\phi = \cos^{-1} \left(\frac{n_x \cdot n_y}{\|n_x\| \cdot \|n_y\|} \right)$ and $0 < \phi_{\min} \leq \phi < \frac{\pi}{2}$.
- 3575 • $\mathbb{S}_X = \mathbb{S}^{h-1} \cap \mathbb{A} = \{x \in \mathbb{R}^h : \|x\| = 1, n_A \cdot x = 0\} \cong S^{h-2} \in \mathbb{S}^{h-1}$.
- 3576 • $\mathbb{S}_Y = \mathbb{S}^{h-1} \cap \mathbb{B} = \{y \in \mathbb{R}^h : \|y\| = 1, n_B \cdot y = 0\} \cong S^{h-2} \in \mathbb{S}^{h-1}$.
- 3577 • $\mathbb{C} = \mathbb{A} \cap \mathbb{B}$.
- 3578 • $h_X = h_Y = h - 1$.
- 3579 • $h_C = h - 2$.
- 3580 • P_A : the projection matrix of \mathbb{A} .
- 3581 • P_B : the projection matrix of \mathbb{B} .
- 3582 • P_C : the projection matrix of \mathbb{C} .
- 3583 • $e_A = \{z \in \mathbb{S}_X : z \perp \mathbb{C}\}$.
- 3584 • $e_B = \{z \in \mathbb{S}_Y : z \perp \mathbb{C}\}$.
- 3585 • $\mathbb{C}^\perp = \text{span } \{e_A\} \oplus \text{span } \{e_B\}$
- 3586 • $\mathbb{R}^h = \mathbb{C} \oplus \mathbb{C}^\perp$.
- 3587 • $X = (x_1, \dots, x_N) \in (\mathbb{S}_X)^N$.
- 3588 • $Y = (y_1, \dots, y_N) \in (\mathbb{S}_Y)^N$.
- 3589 • $\mu_x = \frac{1}{N} \sum_{i=1}^N x_i$.
- 3590 • $\mu_y = \frac{1}{N} \sum_{i=1}^N y_i$.
- 3591 • $c_x = \frac{\mu_x}{\|\mu_x\|}$.
- 3592 • $c_y = \frac{\mu_y}{\|\mu_y\|}$.

3601 **Definition** (Multimodal Contrastive Loss (MCL Loss)). Let (X, Y) be an N -pair configuration,
3602 where $X = (x_1, \dots, x_N) \in (\mathbb{S}^{h-1})^N$ and $Y = (y_1, \dots, y_N) \in (\mathbb{S}^{h-1})^N$. $\forall \tau > 0$, the multimodal
3603 contrastive loss $\mathcal{L}_{\text{MCL}}(\cdot, \cdot) : (\mathbb{S}^{h-1})^N \times (\mathbb{S}^{h-1})^N \rightarrow \mathbb{R}$ is defined as:
3604

$$3605 \mathcal{L}_{\text{MCL}} = \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{\text{MCL}}^i, \text{ where } \mathcal{L}_{\text{MCL}}^i = \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) + \mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}(y_i; X).$$

3606 Here, $\mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}$ is the \mathcal{X} -to- \mathcal{Y} alignment and $\mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}$ is the \mathcal{Y} -to- \mathcal{X} alignment, which are defined
3607 respectively as:
3608

$$3609 \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) = -\log \frac{\exp(x_i \cdot y_i / \tau)}{\sum_{j=1}^N \exp(x_i \cdot y_j / \tau)}, \quad \mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}(y_i; X) = -\log \frac{\exp(x_i \cdot y_i / \tau)}{\sum_{j=1}^N \exp(x_j \cdot y_i / \tau)}.$$

3610 **Definition** (Modality Gap) Let (X, Y) be an N -pair configuration, where $X = (x_1, \dots, x_N) \in$
3611 $(\mathbb{S}^{h-1})^N$ and $Y = (y_1, \dots, y_N) \in (\mathbb{S}^{h-1})^N$. The modality gap between X and Y can be expressed
3612 as the angle between the center representations:
3613

3618
3619
3620

$$\Delta_\theta = \cos^{-1}(c_x \cdot c_y).$$

3621 **Definition** (vMF Distribution). $\forall c \in \mathbb{S}^{h-1}$ and $\kappa \geq 0$, the probability density of a random h -
3622 dimensional unit vector $z \sim \text{vMF}(c, \kappa)$ is given by:
3623

$$3624 \quad f_h(z; c, \kappa) = D_h(\kappa) e^{\kappa c^\top z}, \text{ where } D_h(\kappa) = \frac{\kappa^\nu}{(2\pi)^{\nu+1} I_\nu(\kappa)}.$$

3627 Here, $\nu = h/2 - 1$, and $I_\nu(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ is the modified Bessel function of the first kind of order ν ,
3628 which is defined as:
3629

$$3630 \quad I_\nu(x) = \sum_{k=0}^{\infty} \frac{1}{k! \Gamma(\nu + k + 1)} \left(\frac{x}{2}\right)^{2k+\nu}.$$

3633 **Definition** (Function \tilde{M}). $\forall \kappa, \tau > 0$, a function $\tilde{M}_\kappa(\cdot, \cdot) : [-1, 1] \times [0, 1] \rightarrow \mathbb{R}_0^+$ is defined as:
3634

$$3636 \quad \tilde{M}_\kappa(w, t) = \sqrt{\kappa^2 + \frac{2\kappa w}{\tau} + \frac{t^2}{\tau^2}}.$$

3638 **Definition** (Function $\tilde{\mathcal{J}}$). $\forall \kappa, \nu, \tau > 0$, $\tilde{\mathcal{J}}(\cdot, \cdot, \cdot; \kappa, \nu) : [-1, 1] \times [-1, 1] \times [0, 1] \rightarrow \mathbb{R}$ is defined as:
3639

$$3641 \quad \tilde{\mathcal{J}}(w_1, w_2, t; \kappa, \nu) = -\frac{w_1}{\tau} + \log \left(\frac{I_\nu(\tilde{M}_\kappa(w_2, t))}{\tilde{M}_\kappa(w_2, t)^\nu} \right) - \log \left(\frac{I_\nu(\kappa)}{\kappa^\nu} \right).$$

3645 **Definition** (Function M). $\forall \kappa, \tau > 0$, a function $M_\kappa(\cdot) : [-1, 1] \rightarrow \mathbb{R}_0^+$ is defined as:
3646

$$3647 \quad M_\kappa(w) = \sqrt{\kappa^2 + \frac{2\kappa w}{\tau} + \frac{1}{\tau^2}}$$

$$3649 \quad = \tilde{M}_\kappa(w, 1).$$

3651 **Definition** (Function \mathcal{J}). $\forall \kappa, \nu, \tau > 0$, a function $\mathcal{J}(\cdot; \kappa, \nu) : [-1, 1] \rightarrow \mathbb{R}$ is defined as:
3652

$$3654 \quad \mathcal{J}(w; \kappa, \nu) = -\frac{w}{\tau} + \log \left(\frac{I_\nu(M_\kappa(w))}{M_\kappa(w)^\nu} \right) - \log \left(\frac{I_\nu(\kappa)}{\kappa^\nu} \right)$$

$$3656 \quad = \tilde{\mathcal{J}}(w, w, 1; \kappa, \nu).$$

3658 **Definition** (Function \tilde{M}). $\forall \kappa, \tau > 0$, a function $\tilde{M}_\kappa(\cdot, \cdot) : [-1, 1] \times [0, 1] \rightarrow \mathbb{R}_0^+$ is defined as:
3659

$$3660 \quad \tilde{M}_\kappa(w, t) = \tilde{M}_\kappa(w, t).$$

3662 **Definition** (Function $\hat{\mathcal{J}}$). $\forall \kappa, \nu, \tau > 0$, a function $\hat{\mathcal{J}}(\cdot, \cdot; \kappa, \nu) : [-1, 1] \times [0, 1] \rightarrow \mathbb{R}$ is defined as:
3663

$$3665 \quad \hat{\mathcal{J}}(w, t; \kappa, \nu) = -\frac{w}{\tau} + \log \left(\frac{I_\nu(\tilde{M}_\kappa(w, t))}{\tilde{M}_\kappa(w, t)^\nu} \right) - \log \left(\frac{I_\nu(\kappa)}{\kappa^\nu} \right)$$

$$3668 \quad = \tilde{\mathcal{J}}(w, w, t; \kappa, \nu).$$

3669
3670
3671

3672 E.4.1 PROOF OF THEOREM 4
36733674 In this subsection, we provide the proof of Theorem 4. For convenience in reading, we first restate
3675 Theorem 4 here.3676 **Theorem 4.** [Restate] Let (X, Y) be an N -pair configuration, where $X = (x_1, \dots, x_N) \in (\mathbb{S}_X \setminus \mathbb{C})^N$
3677 are iid samples from $\mu_x = \text{vMF}(c_x, \kappa_x)$, and $Y = (y_1, \dots, y_N) \in (\mathbb{S}_Y \setminus \mathbb{C})^N$ are iid samples
3678 from $\mu_y = \text{vMF}(c_y, \kappa_y)$. Let $\tilde{\nu} = (h-1)/2 - 1$. Denote $\Delta_\theta = \cos^{-1}(c_x \cdot c_y)$ and assume
3679 $c_x, c_y \perp \mathbb{C}$ with $c_x \cdot c_y > 0$. Suppose (X, Y) achieves Intra-Modal Isometry. Then $\forall i \in [N]$, denote
3680 $\theta_i^c = \cos^{-1}(x_i \cdot c_x) = \cos^{-1}(y_i \cdot c_y)$, and $\kappa = \kappa_x = \kappa_y$. Let $\theta_i^c \in (0, \frac{\pi}{2})$ and $\kappa > 0$, it holds that:
3681

3682
3683
$$\lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}}^{i \neq c} - 2 \log(N)$$

3684
3685
$$= \tilde{\mathcal{J}}(\cos(\Delta_\theta), \cos(\theta_i^c), \|P_B x_i\|; \kappa, \tilde{\nu}) + \tilde{\mathcal{J}}(\cos(\Delta_\theta), \cos(\theta_i^c), \|P_A y_i\|; \kappa, \tilde{\nu})$$

3686
3687
$$\geq 2\tilde{\mathcal{J}}\left(\cos^2(\theta_i^c) \cos(\phi_{\min}) + \sin^2(\theta_i^c), \cos(\theta_i^c), \sqrt{\cos^2(\theta_i^c) \cos^2(\phi_{\min}) + \sin^2(\theta_i^c)}; \kappa, \tilde{\nu}\right),$$

3688

3689 where equality is attained if and only if there exists a configuration of (X, Y) such that:
3690

3691 (A8) $P_C x_i = P_C y_i \neq \vec{0}$.

3692 (A9) $\Delta_\theta = \cos^{-1}(c_x \cdot c_y) = \phi_{\min}$.

3693 *Proof.* According to Theorem S7, the convergent function of $\lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}}^{i \neq c} - 2 \log(N)$ is:
3694

3695
3696
$$\begin{aligned} \lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}}^{i \neq c} - 2 \log(N) &= \lim_{N \rightarrow \infty} (\mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_{i \neq c}; Y) - \log(N) + \mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}(y_{i \neq c}; X) - \log(N)) \\ &= \tilde{\mathcal{J}}(w_i, w_i^c, \|P_B x_i\|; \kappa_y, \tilde{\nu}) + \tilde{\mathcal{J}}(w_i, w_i^c, \|P_A y_i\|; \kappa_x, \tilde{\nu}) \\ &= 2\tilde{\mathcal{J}}(w_i, w_i^c, t; \kappa, \tilde{\nu}), \end{aligned} \tag{315}$$

3697 where
3698

3699
3700
$$\begin{aligned} w_i &= \cos^2(\theta_i^c) \cos(\Delta_\theta) + (\theta_i^c)(P_C \cdot x_i) \cdot (P_C \cdot y_i), \\ w_i^c &= \cos(\theta_i^c), \\ t &= \sqrt{\cos^2(\theta_i^c) \cos^2(\Delta_\theta) + \sin^2(\theta_i^c)}. \end{aligned} \tag{316}$$

3711 And Theorem S8 shows the lower bound of the convergent function is:
3712

3713
$$2\tilde{\mathcal{J}}(w_i, w_i^c, t; \kappa, \tilde{\nu}) \geq 2\tilde{\mathcal{J}}(w_{i,\min}, w_i^c, t_{\min}; \kappa, \tilde{\nu}), \tag{317}$$

3714 where
3715

3716
3717
$$\begin{aligned} w_{i,\min} &= \cos^2(\theta_i^c) \cos(\phi_{\min}) + \sin^2(\theta_i^c), \\ t_{\min} &= \sqrt{\cos^2(\theta_i^c) \cos^2(\phi_{\min}) + \sin^2(\theta_i^c)}, \end{aligned} \tag{318}$$

3718 and equality is attained if and only if there exists a configuration of (X, Y) such that:
3719

3720 (i) $P_C \cdot x_i = P_C \cdot y_i$.

3721 (ii) $\Delta_\theta = \phi_{\min}$.
3722

Combining Eq. (315) and Eq. (318), we conclude that:

$$\begin{aligned}
 & \lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}}^{i \neq c} - 2 \log(N) \\
 &= \tilde{\mathcal{J}}(\cos(\Delta_\theta), \cos(\theta_i^c), \|P_B x_i\|; \kappa, \tilde{\nu}) + \tilde{\mathcal{J}}(\cos(\Delta_\theta), \cos(\theta_i^c), \|P_A y_i\|; \kappa, \tilde{\nu}) \\
 &\geq 2\tilde{\mathcal{J}}\left(\cos^2(\theta_i^c) \cos(\phi_{\min}) + \sin^2(\theta_i^c), \cos(\theta_i^c), \sqrt{\cos^2(\theta_i^c) \cos^2(\phi_{\min}) + \sin^2(\theta_i^c)}; \kappa, \tilde{\nu}\right),
 \end{aligned} \tag{319}$$

where equality is attained if and only if there exists a configuration of (X, Y) such that:

$$(A8) \quad P_C x_i = P_C y_i \neq \vec{0}.$$

$$(A9) \quad \Delta_\theta = \cos^{-1}(c_x \cdot c_y) = \phi_{\min}.$$

□

E.4.2 AUXILIARY THEOREMS PART 4

In this subsection, we provide details and proofs of the auxiliary theorems (Theorem S5 and Theorem S7) that support the proof of Theorem 4.

Theorem S7. Let (X, Y) be an N -pair configuration, where $X = (x_1, \dots, x_N) \in (\mathbb{S}_X \setminus \mathbb{C})^N$ are iid samples from $\mu_x = \text{vMF}(c_x, \kappa_x)$, and $Y = (y_1, \dots, y_N) \in (\mathbb{S}_Y \setminus \mathbb{C})^N$ are iid samples from $\mu_y = \text{vMF}(c_y, \kappa_y)$. Let $\tilde{\nu} = (h-1)/2 - 1$. Denote $\Delta_\theta = \cos^{-1}(c_x \cdot c_y)$ and assume $c_x, c_y \perp \mathbb{C}$ with $c_x \cdot c_y > 0$. Suppose (X, Y) achieves Intra-Modal Isometry. Then $\forall i \in [N]$, denote $\theta_i^c = \cos^{-1}(x_i \cdot c_x) = \cos^{-1}(y_i \cdot c_y)$, and $\kappa = \kappa_x = \kappa_y$. Let $\kappa > 0$, it holds that:

$$\begin{aligned}
 \lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}}^{i \neq c} - 2 \log(N) &= \lim_{N \rightarrow \infty} (\mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_{i \neq c}; Y) - \log(N) + \mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}(y_{i \neq c}; X) - \log(N)) \\
 &= \tilde{\mathcal{J}}(w_i, w_i^c, \|P_B x_i\|; \kappa_y, \tilde{\nu}) + \tilde{\mathcal{J}}(w_i, w_i^c, \|P_A y_i\|; \kappa_x, \tilde{\nu}) \\
 &= 2\tilde{\mathcal{J}}(w_i, w_i^c, t; \kappa, \tilde{\nu}),
 \end{aligned} \tag{320}$$

where

$$\begin{aligned}
 w_i &= \cos^2(\theta_i^c) \cos(\Delta_\theta) + (\theta_i^c)(P_C \cdot x_i) \cdot (P_C \cdot y_i), \\
 w_i^c &= \cos(\theta_i^c), \\
 t &= \sqrt{\cos^2(\theta_i^c) \cos^2(\Delta_\theta) + \sin^2(\theta_i^c)}.
 \end{aligned} \tag{321}$$

Proof. **Step 1:** We first decompose $\lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}}^{i \neq c} - 2 \log(N)$ into two parts:

$$\begin{aligned}
 \lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}}^{i \neq c} - 2 \log(N) &= \lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_{i \neq c}; Y) - \log(N) \\
 &\quad + \lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}(y_{i \neq c}; X) - \log(N).
 \end{aligned} \tag{322}$$

The convergent function of $\mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_{i \neq c}; Y)$ as $N \rightarrow \infty$. $\forall i \in [N], i \neq c, x_i \in X$, denote $w_i = x_i \cdot y_i$, $w_{x_i, c_y} = x_i \cdot c_y$ and $w_{y_i, c_x} = y_i \cdot c_x$. $\forall \kappa_y > 0$, as prove in Theorem S5:

3780
 3781
 3782 $\lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) - \log(N) = \lim_{N \rightarrow \infty} -\log \frac{\exp(x_i \cdot y_i / \tau)}{\sum_{j=1}^N \exp(x_i \cdot y_j / \tau)} - \log(N)$
 3783
 3784 $= -\frac{w_i}{\tau} + \log \left(\frac{I_{\tilde{\nu}} \left(\tilde{M}_{\kappa_y} (w_{x_i, c_y}, \|P_B x_i\|) \right)}{\tilde{M}_{\kappa_y} (w_{x_i, c_y}, \|P_B x_i\|)^{\tilde{\nu}}} \right) - \log \left(\frac{I_{\tilde{\nu}} (\kappa_y)}{\kappa_y^{\tilde{\nu}}} \right) \quad (323)$
 3785
 3786
 3787
 3788 $= \tilde{\mathcal{J}} (w_i, w_{x_i, c_y}, \|P_B x_i\|; \kappa_y, \tilde{\nu}),$
 3789

3790 where $\forall \kappa, \tau > 0$, $\tilde{\mathcal{J}}(\cdot, \cdot, \cdot; \kappa, \tilde{\nu})$ is a function on $[-1, 1] \times [-1, 1] \times [0, 1]$ and $\tilde{M}_{\kappa}(\cdot, \cdot) : [-1, 1] \times [0, 1] \rightarrow \mathbb{R}_0^+$ is defined as:

3793
 3794 $\tilde{M}_{\kappa} (w, t) = \sqrt{\kappa^2 + \frac{2\kappa w}{\tau} + \frac{t^2}{\tau^2}}, \quad (324)$
 3795

3796 and I_{ν} is the modified Bessel function of the first kind of order ν , which is defined as:

3797
 3798
 3799 $I_{\nu} (m) = \sum_{k=0}^{\infty} \frac{1}{k! \Gamma(\nu + k + 1)} \left(\frac{m}{2} \right)^{2k+\nu}. \quad (325)$
 3800
 3801

3802 When (X, Y) achieves Intra-Modal Isometry, we have $w_{x_i, c_y} = x_i \cdot c_x = y_i \cdot c_x = w_{y_i, c_x}$. Denote
 3803 $w_i^c = w_{x_i, c_y} = w_{y_i, c_x} = \cos(\theta_i^c)$. This implies $\kappa_x = \kappa_y = \kappa$.

3804 Then, Eq. (323) can be re-written as:

3805
 3806
 3807
 3808 $\lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_i; Y) - \log(N) = -\frac{w_i}{\tau} + \log \left(\frac{I_{\tilde{\nu}} \left(\tilde{M}_{\kappa} (w_i^c, \|P_B x_i\|) \right)}{\tilde{M}_{\kappa} (w_i^c, \|P_B x_i\|)^{\tilde{\nu}}} \right) - \log \left(\frac{I_{\tilde{\nu}} (\kappa)}{\kappa^{\tilde{\nu}}} \right)$
 3809
 3810
 3811 $= \tilde{\mathcal{J}} (w_i, w_i^c, \|P_B x_i\|; \kappa, \tilde{\nu}). \quad (326)$
 3812

3813 Similarly, the convergent function of $\mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}(y_i; X)$ as $N \rightarrow \infty$ can be written as:

3814
 3815
 3816
 3817 $\lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}(y_i; X) - \log(N) = \lim_{N \rightarrow \infty} -\log \frac{\exp(x_i \cdot y_i / \tau)}{\sum_{j=1}^N \exp(x_i \cdot y_j / \tau)} - \log(N)$
 3818
 3819 $= -\frac{w_i}{\tau} + \log \left(\frac{I_{\tilde{\nu}} \left(\tilde{M}_{\kappa_x} (w_{y_i, c_x}, \|P_A y_i\|) \right)}{\tilde{M}_{\kappa_x} (w_{y_i, c_x}, \|P_A y_i\|)^{\tilde{\nu}}} \right) - \log \left(\frac{I_{\tilde{\nu}} (\kappa_x)}{\kappa_x^{\tilde{\nu}}} \right) \quad (327)$
 3820
 3821
 3822
 3823 $= -\frac{w_i}{\tau} + \log \left(\frac{I_{\tilde{\nu}} \left(\tilde{M}_{\kappa} (w_i^c, \|P_A y_i\|) \right)}{\tilde{M}_{\kappa} (w_i^c, \|P_A y_i\|)^{\tilde{\nu}}} \right) - \log \left(\frac{I_{\tilde{\nu}} (\kappa)}{\kappa^{\tilde{\nu}}} \right)$
 3824
 3825
 3826 $= \tilde{\mathcal{J}} (w_i, w_i^c, \|P_A y_i\|; \kappa, \tilde{\nu}).$
 3827

3828 **Step 2** Now, let us decompose the embedding space. Define two vectors e_A and e_B such that:

3829
 3830 $e_A \in \mathbb{S}_X, \text{ and } e_A \perp \mathbb{C},$
 3831 $e_B \in \mathbb{S}_Y, \text{ and } e_B \perp \mathbb{C}. \quad (328)$
 3832

3833 Let \mathbb{C}^{\perp} be the 2-dimensional orthogonal complement of C , and \mathbb{C}^{\perp} satisfies:

3834

$$\begin{aligned} \mathbb{C}^\perp &= \text{span}\{e_A\} \oplus \text{span}\{e_B\}, \\ \mathbb{R}^h &= \mathbb{C} \oplus \mathbb{C}^\perp. \end{aligned} \tag{329}$$

3835

Since $n_A, n_B \in \mathbb{C}^\perp$, $n_A \perp e_A$ and $n_B \perp e_B$, we have:

3836

3837

3838

3839

3840

3841

3842

3843

3844

3845

3846

3847

3848

3849

3850

3851

3852

3853

3854

3855

3856

3857

3858

3859

3860

3861

3862

3863

3864

3865

and we choose a pair of e_A and e_B such that:

$$\langle e_A, e_B \rangle = \pm \langle n_A, n_B \rangle, \tag{330}$$

Denote $\theta_i = \cos^{-1}(w_i)$. When $c_x, c_y \perp \mathbb{C}$, $\Delta_\theta = \phi$. And without loss of generality, we can set the coordinate as:

$$\begin{aligned} n_A &= \left(\sin\left(\frac{\Delta_\theta}{2}\right), -\cos\left(\frac{\Delta_\theta}{2}\right), 0, 0, \dots, 0 \right), \\ n_B &= \left(-\sin\left(\frac{\Delta_\theta}{2}\right), -\cos\left(\frac{\Delta_\theta}{2}\right), 0, 0, \dots, 0 \right), \\ c_x = e_A &= \left(\cos\left(\frac{\Delta_\theta}{2}\right), \sin\left(\frac{\Delta_\theta}{2}\right), 0, 0, \dots, 0 \right), \\ c_y = e_B &= \left(\cos\left(\frac{\Delta_\theta}{2}\right), -\sin\left(\frac{\Delta_\theta}{2}\right), 0, 0, \dots, 0 \right), \\ \mathbb{C} &= \text{span}\{e_3\} \oplus \text{span}\{e_3\} \oplus \dots \oplus \text{span}\{e_h\}. \end{aligned} \tag{332}$$

Therefore, $\forall x_i \in \mathbb{S}_X = \mathbb{A} \cap \mathbb{S}^{h-1}$ and $\forall y_i \in \mathbb{S}_Y = \mathbb{B} \cap \mathbb{S}^{h-1}$, $\exists u_i^x, u_i^y \in \mathbb{C} \cap \mathbb{S}^{h-1}$, such that:

$$\begin{aligned} x_i &= \cos(\theta_i^c) e_A + \sin(\theta_i^c) u_i^x = \cos(\theta_i^c) c_x + \sin(\theta_i^c) u_i^x, \\ y_i &= \cos(\theta_i^c) e_B + \sin(\theta_i^c) u_i^y = \cos(\theta_i^c) c_y + \sin(\theta_i^c) u_i^y. \end{aligned} \tag{333}$$

Using orthogonality, we have:

$$\begin{aligned} P_B \cdot e_A &= \langle e_A, e_B \rangle e_B = \cos(\Delta_\theta) e_B, \\ P_B \cdot u_i^x &= u_i^x, \end{aligned} \tag{334}$$

and

$$\begin{aligned} P_A \cdot e_B &= \langle e_A, e_B \rangle e_A = \cos(\Delta_\theta) e_A, \\ P_A \cdot u_i^y &= u_i^y, \end{aligned} \tag{335}$$

and

$$\begin{aligned} P_C \cdot e_A &= P_C \cdot e_B = 0, \\ P_C \cdot u_i^x &= u_i^x, \\ P_C \cdot u_i^y &= u_i^y. \end{aligned} \tag{336}$$

Then the projections of (x_i, y_i) are:

$$\begin{aligned} P_B \cdot x_i &= \cos(\theta_i^c) \cos(\Delta_\theta) e_B + \sin(\theta_i^c) u_i^x = \cos(\theta_i^c) \cos(\Delta_\theta) c_y + \sin(\theta_i^c) u_i^x, \\ P_A \cdot y_i &= \cos(\theta_i^c) \cos(\Delta_\theta) e_A + \sin(\theta_i^c) u_i^y = \cos(\theta_i^c) \cos(\Delta_\theta) c_x + \sin(\theta_i^c) u_i^y, \end{aligned} \tag{337}$$

and

3888
3889
3890
3891

$$\begin{aligned} P_C \cdot x_i &= \sin(\theta_i^c) u_i^x, \\ P_C \cdot y_i &= \sin(\theta_i^c) u_i^y. \end{aligned} \quad (338)$$

3892 Therefore, we get:

3893
3894
3895
3896
3897
3898
3899

$$\begin{aligned} w_i &= x_i \cdot y_i = \cos^2(\theta_i^c) c_x \cdot c_y + \sin^2(\theta_i^c) u_i^x \cdot u_i^y \\ &= \cos^2(\theta_i^c) \cos(\Delta_\theta) + \sin^2(\theta_i^c) u_i^x \cdot u_i^y \\ &= \cos^2(\theta_i^c) \cos(\Delta_\theta) + (P_C \cdot x_i) \cdot (P_C \cdot y_i), \end{aligned} \quad (339)$$

3900
3901
3902
3903
3904
3905
3906
3907
3908
3909
3910
3911
3912
3913
3914
3915
3916
3917
3918
3919
3920
3921
3922
3923
3924
3925
3926
3927
3928
3929
3930
3931
3932
3933
3934
3935
3936
3937
3938
3939
3940
3941

$$\begin{aligned} \|P_B x_i\| &= \sqrt{\cos^2(\theta_i^c) \cos^2(\Delta_\theta) c_y \cdot c_y + 2 \cos(\theta_i^c) \cos(\Delta_\theta) \sin(\theta_i^c) c_y \cdot u_i^x + \sin^2(\theta_i^c) u_i^x \cdot u_i^x} \\ &= \sqrt{\cos^2(\theta_i^c) \cos^2(\Delta_\theta) + 2 \cos(\theta_i^c) \cos(\Delta_\theta) \sin(\theta_i^c) c_y \cdot u_i^x + \sin^2(\theta_i^c)} \\ &= \sqrt{\cos^2(\theta_i^c) \cos^2(\Delta_\theta) + \sin^2(\theta_i^c)}, \end{aligned} \quad (340)$$

and

$$\begin{aligned} \|P_A y_i\| &= \sqrt{\cos^2(\theta_i^c) \cos^2(\Delta_\theta) c_x \cdot c_x + 2 \cos(\theta_i^c) \cos(\Delta_\theta) \sin(\theta_i^c) c_x \cdot u_i^y + \sin^2(\theta_i^c) u_i^y \cdot u_i^y} \\ &= \sqrt{\cos^2(\theta_i^c) \cos^2(\Delta_\theta) + 2 \cos(\theta_i^c) \cos(\Delta_\theta) \sin(\theta_i^c) c_x \cdot u_i^y + \sin^2(\theta_i^c)} \\ &= \sqrt{\cos^2(\theta_i^c) \cos^2(\Delta_\theta) + \sin^2(\theta_i^c)}, \\ &= \|P_B x_i\|. \end{aligned} \quad (341)$$

Let $t = \|P_B x_i\| = \|P_A y_i\|$. Plugging Eq. (339), Eq. (340) and Eq. (341) into Eq. (322), Eq. (326) and Eq. (327), we conclude that:

$$\begin{aligned} \lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}}^{i \neq c} - 2 \log(N) &= \lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_{i \neq c}; Y) - \log(N) \\ &\quad + \lim_{N \rightarrow \infty} \mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}(y_{i \neq c}; X) - \log(N) \\ &= \tilde{\mathcal{J}}(w_i, w_i^c, \|P_B x_i\|; \kappa, \tilde{\nu}) + \tilde{\mathcal{J}}(w_i, w_i^c, \|P_A y_i\|; \kappa, \tilde{\nu}) \\ &= 2\tilde{\mathcal{J}}(w_i, w_i^c, t; \kappa, \tilde{\nu}), \end{aligned} \quad (342)$$

where

$$\begin{aligned} w_i &= x_i \cdot y_i = \cos^2(\theta_i^c) \cos(\Delta_\theta) + (P_C \cdot x_i) \cdot (P_C \cdot y_i), \\ w_i^c &= x_i \cdot c_y = y_i \cdot c_x = \cos(\theta_i^c), \\ t &= \sqrt{\cos^2(\theta_i^c) \cos^2(\Delta_\theta) + \sin^2(\theta_i^c)}. \end{aligned} \quad (343)$$

□

Theorem S8. Let (X, Y) be an N -pair configuration, where $X = (x_1, \dots, x_N) \in (\mathbb{S}_X \setminus \mathbb{C})^N$ are iid samples from $\mu_x = \text{vMF}(c_x, \kappa_x)$, and $Y = (y_1, \dots, y_N) \in (\mathbb{S}_Y \setminus \mathbb{C})^N$ are iid samples from $\mu_y = \text{vMF}(c_y, \kappa_y)$. Let $\tilde{\nu} = (h-1)/2 - 1$. Denote $\Delta_\theta = \cos^{-1}(c_x \cdot c_y)$ and assume $c_x, c_y \perp \mathbb{C}$ with $c_x \cdot c_y > 0$. $\forall i \in [N]$, suppose $\theta_i^c = \cos^{-1}(x_i \cdot c_x) = \cos^{-1}(y_i \cdot c_y) \in (0, \frac{\pi}{2})$ and $\kappa > 0$, it holds that:

$$\tilde{\mathcal{J}}(w_i, w_i^c, t; \kappa, \tilde{\nu}) \geq \tilde{\mathcal{J}}(w_{i,\min}, w_i^c, t_{\min}; \kappa, \tilde{\nu}), \quad (344)$$

3942 where

3943

3944

$$w_i = \cos^2(\theta_i^c) \cos(\Delta_\theta) + (P_C \cdot x_i) \cdot (P_C \cdot y_i),$$

3945

$$w_i^c = \cos(\theta_i^c),$$

3946

$$t = \sqrt{\cos^2(\theta_i^c) \cos^2(\Delta_\theta) + \sin^2(\theta_i^c)},$$

3947

3948

$$w_{i,\min} = \cos^2(\theta_i^c) \cos(\phi_{\min}) + \sin^2(\theta_i^c),$$

3949

3950

$$t_{\min} = \sqrt{\cos^2(\theta_i^c) \cos^2(\phi_{\min}) + \sin^2(\theta_i^c)},$$

3951

3952 and equality is attained if and only if there exists a configuration of (X, Y) such that:

3953

$$(B6) \quad P_C \cdot x_i = P_C \cdot y_i.$$

3954

$$(B7) \quad \Delta_\theta = \phi_{\min}.$$

3955

3956 *Proof. Step 1:* Similarly to the proof of Theorem S6 in Sec. E.3.2, we start the proof by finding the convergent function of $\lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}}^{i \neq c} - 2 \log(N)$ as $N \rightarrow \infty$. Let $w_i =$

3957 As proven in Theorem S7:

3958

3959

$$\begin{aligned} \lim_{N \rightarrow \infty} \mathcal{L}_{\text{MCL}}^{i \neq c} - 2 \log(N) &= \lim_{N \rightarrow \infty} (\mathcal{L}_{\mathcal{X} \rightarrow \mathcal{Y}}(x_{i \neq c}; Y) - \log(N) + \mathcal{L}_{\mathcal{Y} \rightarrow \mathcal{X}}(y_{i \neq c}; X) - \log(N)) \\ &= \tilde{\mathcal{J}}(w_i, w_i^c, \|P_B x_i\|; \kappa, \tilde{\nu}) + \tilde{\mathcal{J}}(w_i, w_i^c, \|P_A y_i\|; \kappa, \tilde{\nu}) \\ &= 2\tilde{\mathcal{J}}(w_i, w_i^c, t; \kappa, \tilde{\nu}). \end{aligned} \quad (346)$$

3960 $\forall \kappa, \nu, \tau > 0$, $\tilde{\mathcal{J}}(\cdot, \cdot, \cdot; \kappa, \nu) : [-1, 1] \times [-1, 1] \times [0, 1] \rightarrow \mathbb{R}$ is defined as:

3961

3962

$$\tilde{\mathcal{J}}(w_1, w_2, t; \kappa, \nu) = -\frac{w_1}{\tau} + \log \left(\frac{I_\nu \left(\tilde{M}_\kappa(w_2, t) \right)}{\tilde{M}_\kappa(w_2, t)^\nu} \right) - \log \left(\frac{I_\nu(\kappa)}{\kappa^\nu} \right), \quad (347)$$

3963

3964 and $\tilde{M}_\kappa(\cdot, \cdot) : [-1, 1] \times [0, 1] \rightarrow \mathbb{R}_0^+$ is defined as:

3965

3966

$$\tilde{M}_\kappa(w, t) = \sqrt{\kappa^2 + \frac{2\kappa w}{\tau} + \frac{t^2}{\tau^2}}. \quad (348)$$

3967

3968 and I_ν is the modified Bessel function of the first kind of order ν , which is defined as:

3969

3970

$$I_\nu(m) = \sum_{k=0}^{\infty} \frac{1}{k! \Gamma(\nu + k + 1)} \left(\frac{m}{2} \right)^{2k+\nu}, \quad (349)$$

3971

3972 and

3973

3974

3975

3976

3977

3978

3979

3980

3981

3982

3983

3984

3985

3986

3987

3988

3989

3990

3991

$$w_i = \cos^2(\theta_i^c) \cos(\Delta_\theta) + \sin^2(\theta_i^c) (P_C \cdot x_i) \cdot (P_C \cdot y_i),$$

3992

3993

3994

3995

$$w_i^c = \cos(\theta_i^c),$$

3996

3997

3998

3999

4000

4001

$$t = \sqrt{\cos^2(\theta_i^c) \cos^2(\Delta_\theta) + \sin^2(\theta_i^c)}.$$

4002 **Step 2:**

4003 According to the Cauchy-Schwarz inequality and Eq. (338):

4004

4005

4006

4007

4008

4009

4010

$$(P_C \cdot x_i) \cdot (P_C \cdot y_i) \leq \sin^2(\theta_i^c), \quad (351)$$

3996 where equality is attained if and only if there exists a configuration of (X, Y) such that:
 3997

3998 (B6) $P_C \cdot x_i = P_C \cdot y_i$.
 3999

4000 And therefore:
 4001

4002 $w_i = \cos^2(\theta_i^c) \cos(\Delta_\theta) + (P_C \cdot x_i) \cdot (P_C \cdot y_i)$
 4003 $\leq \cos^2(\theta_i^c) \cos(\Delta_\theta) + \sin^2(\theta_i^c)$,
 4004

4005 and then $\tilde{\mathcal{J}}(w_i, w_i^c, t; \kappa, \tilde{\nu})$ in Eq. (346) can be bounded below by:
 4006

4008 $\tilde{\mathcal{J}}(w_i, w_i^c, t; \kappa, \tilde{\nu}) \geq \tilde{\mathcal{J}}(\cos^2(\theta_i^c) \cos(\Delta_\theta) + \sin^2(\theta_i^c),$
 4009 $\cos(\theta_i^c),$
 4010 $\sqrt{\cos^2(\theta_i^c) \cos^2(\Delta_\theta) + \sin^2(\theta_i^c)}; \kappa, \tilde{\nu})$.
 4011

4013 Here, for any given non-center pair $(x_i, y_i)_{i \neq c}$, θ_i^c is fixed, then the RHS of Eq. (353) becomes a
 4014 function of $\cos(\Delta_\theta)$.

4015 Denote:
 4016

4017 $f_1(\cos(\Delta_\theta)) := \cos^2(\theta_i^c) \cos(\Delta_\theta) + \sin^2(\theta_i^c),$
 4018
 4019 $f_2(\cos(\Delta_\theta)) := \sqrt{\cos^2(\theta_i^c) \cos^2(\Delta_\theta) + \sin^2(\theta_i^c)},$
 4020

4021 then the Eq. (346) can be re-written as:
 4022

4023 $\tilde{\mathcal{J}}(w_i, w_i^c, t; \kappa, \tilde{\nu}) \geq \tilde{\mathcal{J}}(f_1(\cos(\Delta_\theta)), \cos(\theta_i^c), f_2(\cos(\Delta_\theta)); \kappa, \tilde{\nu})$.
 4024

4025 According to Lemma 14, $\tilde{\mathcal{J}}(f_1(\cos(\Delta_\theta)), \cos(\theta_i^c), f_2(\cos(\Delta_\theta)); \kappa, \tilde{\nu})$ is a decreasing function
 4026 of $\cos(\Delta_\theta)$ when $\theta_i^c \in [0, \frac{\pi}{2}]$, we have:
 4027

4029 $\tilde{\mathcal{J}}(f_1(\cos(\Delta_\theta)), \cos(\theta_i^c), f_2(\cos(\Delta_\theta))) \geq \tilde{\mathcal{J}}(f_1(\cos(\phi_{\min})), \cos(\theta_i^c), f_2(\cos(\phi_{\min})))$.
 4030

4031 where equality is attained if and only if there exists a configuration of (X, Y) such that:
 4032

4033 (B7) $\Delta_\theta = \phi_{\min}$.
 4034

4035 Combining Eq. (351) and Eq. (356), we conclude that:
 4036

4037 $\tilde{\mathcal{J}}(w_i, w_i^c, t; \kappa, \tilde{\nu}) \geq \tilde{\mathcal{J}}(w_{i,\min}, w_i^c, t_{\min}; \kappa, \tilde{\nu})$,
 4038

4039 where
 4040

4041 $w_i = \cos^2(\theta_i^c) \cos(\Delta_\theta) + (P_C \cdot x_i) \cdot (P_C \cdot y_i),$
 4042
 4043 $w_i^c = \cos(\theta_i^c),$
 4044 $t = \sqrt{\cos^2(\theta_i^c) \cos^2(\Delta_\theta) + \sin^2(\theta_i^c)}$,
 4045
 4046 $w_{i,\min} = \cos^2(\theta_i^c) \cos(\phi_{\min}) + \sin^2(\theta_i^c),$
 4047
 4048 $t_{\min} = \sqrt{\cos^2(\theta_i^c) \cos^2(\phi_{\min}) + \sin^2(\theta_i^c)}$.
 4049

and equality is attained if and only if there exists a configuration of (X, Y) such that:

4050 (B6) $P_C \cdot x_i = P_C \cdot y_i$.

4051
4052 (B7) $\Delta_\theta = \phi_{\min}$.

4053

4054

4055

4056 E.4.3 PROOFS COROLLARY 2,3,4

4057

4058 In this subsection, we provide the proofs of Corollary 2, Corollary 3 and Corollary 4. Note that these
4059 corollaries all follow the conditions described in Theorem 3 and Theorem 4. For convenience in
4060 reading, we restate Corollary 2,3 4 before the proofs.4061 **Corollary 2.** $\forall i \in [N], i \neq c$, if $c_x, c_y \perp \mathbb{C}$ and $P_C x_i = P_C y_i \neq \vec{0}$ and $\phi > 0$, then the following
4062 holds:

4063

4064 (A10) $(x_i, y_i)_{i \neq c}$ are not perfectly aligned.

4065

4066 *Proof.* $\forall (x_i, y_i)_{i \neq c}$, denote $w_i = x_i \cdot y_i$. (x_i, y_i) are perfectly aligned when w_i reach its maximum.4067 According to Lemma 12, when x_i is fixed w_i is maximized if and only if:

4068

4069 (i) $y_i = \frac{P_B \cdot x_i}{\|P_B \cdot x_i\|}$.

4070

4071 And when y_i is fixed w_i is maximized if and only if:

4072

4073 (ii) $x_i = \frac{P_A \cdot y_i}{\|P_A \cdot y_i\|}$.

4074

4075 According to Lemma 13, when $\phi > 0$, $x_i, y_i \not\perp \mathbb{C}$ and $x_i, y_i \notin \mathbb{C}$, we have:

4076

4077
$$\begin{aligned} y_i &\neq \frac{P_B \cdot x_i}{\|P_B \cdot x_i\|}, \\ 4078 x_i &\neq \frac{P_A \cdot y_i}{\|P_A \cdot y_i\|}. \end{aligned} \tag{359}$$

4079

4080 Therefore, $(x_i, y_i)_{i \neq c}$ are not perfectly aligned.

4081

4082

4083 **Corollary 3.** $\forall i \in [N], i \neq c$, if $c_x, c_y \perp \mathbb{C}$, $P_C x_i = P_C y_i$ and $(x_i, y_i)_{i \neq c} \in \mathbb{S}^{h-1} \setminus \mathbb{C}$, then
4084 $(x_i, y_i)_{i \neq c}$ are perfectly aligned if the following condition holds:

4085

4086 (A11) $\Delta_\theta = \phi = 0$.

4087

4088 *Proof.* According to Eq. (337) and Eq. (338) in the proof of Theorem S7, the projections of (x_i, y_i)
4089 are:

4090

4091
$$\begin{aligned} P_B \cdot x_i &= \cos(\theta_i^c) \cos(\Delta_\theta) e_B + \sin(\theta_i^c) u_i^x = \cos(\theta_i^c) \cos(\Delta_\theta) c_y + \sin(\theta_i^c) u_i^x, \\ 4092 P_A \cdot y_i &= \cos(\theta_i^c) \cos(\Delta_\theta) e_A + \sin(\theta_i^c) u_i^y = \cos(\theta_i^c) \cos(\Delta_\theta) c_x + \sin(\theta_i^c) u_i^y, \end{aligned} \tag{360}$$

4093

4094 and

4095

4096
$$\begin{aligned} P_C \cdot x_i &= \sin(\theta_i^c) u_i^x, \\ 4097 P_C \cdot y_i &= \sin(\theta_i^c) u_i^y. \end{aligned} \tag{361}$$

4098

4099 Then, when $\phi = \Delta_\theta = 0$

4100

4101

4102

4103

4104

4105

$$P_B \cdot x_i = P_C \cdot x_i = P_C \cdot y_i = P_A \cdot y_i, \quad (362)$$

4106

4107 and

4108

4109

4110

$$x_i = P_B \cdot x_i = P_A \cdot y_i = y_i. \quad (363)$$

4111

4112 In this case, $(x_i, y_i)_{i \neq c}$ are not perfectly aligned.

4113

4114

4115 **Corollary 4.** $\forall i \in [N], i \neq c$, if $c_x, c_y \perp \mathbb{C}$ and $P_C x_i = P_C y_i$, then the following holds:

4116

4117

4118 (A12) $(\frac{P_C x_i}{\|P_C x_i\|}, \frac{P_C y_i}{\|P_C y_i\|})_{i \neq c}$ are perfectly aligned

4119

4120 *Proof.* Denote:

4121

$$\begin{aligned} x_i^* &= \frac{P_C x_i}{\|P_C x_i\|}, \\ y_i^* &= \frac{P_C y_i}{\|P_C y_i\|}. \end{aligned} \quad (364)$$

4126

4127 Since $P_C x_i = P_C y_i$, then:

4128

$$x_i^* = y_i^*. \quad (365)$$

4130

4131 In this case, $(x_i^*, y_i^*)_{i \neq c}$ are not perfectly aligned.

4132

4133

4134 E.4.4 TECHNICAL LEMMAS PART 4

4135

4136 In this subsection, we provide details and proofs of technical lemmas (Lemma 14 and Lemma 15) that support the proof of Theorem 4, Theorem S7 and Theorem S8.

4137

4138 **Lemma 14.** $\forall \kappa, \nu, \tau > 0$, a function $\tilde{\mathcal{J}}(\cdot; \kappa, \nu) : (0, 1] \rightarrow \mathbb{R}$ is defined as:

4139

$$\tilde{\mathcal{J}}(w_c; \kappa, \nu) = \tilde{\mathcal{J}}(f_1(w_c), \cos(\theta_i^c), f_2(w_c); \kappa, \nu), \quad (366)$$

4140

4141 where $f_1(\cdot) : (0, 1] \rightarrow \mathbb{R}_0^+$ and $f_2(\cdot) : [0, 1] \rightarrow \mathbb{R}_0^+$ are defined as:

4142

4143

$$\begin{aligned} f_1(w_c) &:= \cos^2(\theta_i^c) w_c + \sin^2(\theta_i^c), \\ f_2(w_c) &:= \sqrt{\cos^2(\theta_i^c) w_c^2 + \sin^2(\theta_i^c)}. \end{aligned} \quad (367)$$

4144

4145 and $\tilde{\mathcal{J}}(\cdot, \cdot, \cdot; \kappa, \nu) : [-1, 1] \times [-1, 1] \times [0, 1] \rightarrow \mathbb{R}$ is defined as:

4146

4147

$$\tilde{\mathcal{J}}(w_1, w_2, t; \kappa, \nu) = -\frac{w_1}{\tau} + \log \left(\frac{I_\nu(\tilde{M}_\kappa(w_2, t))}{\tilde{M}_\kappa(w_2, t)^\nu} \right) - \log \left(\frac{I_\nu(\kappa)}{\kappa^\nu} \right), \quad (368)$$

4148

4149

4150 and $\tilde{M}_\kappa(\cdot, \cdot) : [-1, 1] \times [0, 1] \rightarrow \mathbb{R}_0^+$ is defined as:

4151

4152

4153

4154

4155

4156

4157

$$\tilde{M}_\kappa(w, t) = \sqrt{\kappa^2 + \frac{2\kappa w}{\tau} + \frac{t^2}{\tau^2}}, \quad (369)$$

4158 and I_ν is the modified Bessel function of the first kind of order ν , which is defined as:
4159

$$4160 \quad 4161 \quad I_\nu(m) = \sum_{k=0}^{\infty} \frac{1}{k! \Gamma(\nu + k + 1)} \left(\frac{m}{2}\right)^{2k+\nu}, \quad 4162$$

4163 It holds that, for any fixed $\theta_i^c \in [0, \frac{\pi}{2}]$, $\bar{\mathcal{J}}(\cdot)$ is a strictly decreasing function on $(0, 1]$.
4164

4165 *Proof.* Let us first decompose the function \mathcal{J} . Denote a constant and a function C_1 and $G_2(t)$ as:
4166

$$4167 \quad 4168 \quad G_1(w_c) = -\frac{\cos^2(\theta_i^c) w_c}{\tau}, \\ 4169 \quad 4170 \quad G_3(m) = \log(I_\nu(m)) - \nu \log(m), \\ 4171 \quad 4172 \quad G_2(w_c) = G_3\left(\tilde{M}_\kappa(\cos(\theta_i^c), f_2(w_c))\right) \\ 4173 \quad 4174 \quad = \log\left(I_\nu\left(\tilde{M}_\kappa(\cos(\theta_i^c), f_2(w_c))\right)\right) - \nu \log\left(\tilde{M}_\kappa(\cos(\theta_i^c), f_2(w_c))\right). \quad 4175$$

4175 Denote the function $G(w_c)$ and the constant C as:
4176

$$4177 \quad 4178 \quad G(w_c) = G_2(w_c) + G_3(w_c), \\ 4179 \quad 4180 \quad C = -\log\left(\frac{I_\nu(\kappa)}{\kappa^\nu}\right). \quad 4181$$

4181 Then the function $\bar{\mathcal{J}}$ can be written as:
4182

$$4183 \quad 4184 \quad \bar{\mathcal{J}}(w_c; \kappa, \nu) = -\frac{\cos^2(\theta_i^c) w_c}{\tau} + \log\left(\frac{I_\nu\left(\tilde{M}_\kappa(\cos(\theta_i^c), f_2(w_c))\right)}{\tilde{M}_\kappa(\cos(\theta_i^c), f_2(w_c))^\nu}\right) - \log\left(\frac{I_\nu(\kappa)}{\kappa^\nu}\right) \\ 4185 \quad 4186 \quad = G(w_c) + C. \quad 4187$$

4189 Now, we investigate derivatives of $G(w_c)$.
4190

4191 The first derivative of G_1 is:
4192

$$4193 \quad G'_1(w_c) = -\frac{\cos^2(\theta_i^c)}{\tau} < 0. \quad 4194$$

4195 According to Lemma 7, the first derivative of $G_3(m)$ is:
4196

$$4197 \quad 4198 \quad G'_3(m) = \frac{I_{\nu+1}(m)}{I_\nu(m)} \in (0, 1). \quad 4199$$

4200 The derivative of \tilde{M}_κ with respect to is $f_2^2(w_c)$:
4201

$$4202 \quad 4203 \quad \tilde{M}'_\kappa(\cos(\theta_i^c), f_2(w_c)) = \frac{\partial}{\partial f_2^2(w_c)} \tilde{M}_\kappa(\cos(\theta_i^c), f_2(w_c)) \\ 4204 \quad 4205 \quad = \frac{\partial}{\partial f_2^2(w_c)} \left(\kappa^2 + \frac{2\kappa \cos(\theta_i^c)}{\tau} + \frac{f_2^2(w_c)}{\tau^2} \right)^{1/2} \\ 4206 \quad 4207 \quad = \frac{1}{2} \left(\kappa^2 + \frac{2\kappa \cos(\theta_i^c)}{\tau} + \frac{f_2^2(w_c)}{\tau^2} \right)^{-1/2} \cdot \frac{1}{\tau^2} \\ 4208 \quad 4209 \quad = \frac{1}{2\tau^2} \frac{1}{\tilde{M}_\kappa(\cos(\theta_i^c), f_2(w_c))} \\ 4210 \quad 4211 \quad > 0. \quad 4212$$

4212 The derivative of f_2^2 is:
 4213

$$\begin{aligned} 4214 \quad f_2^{2\prime}(w_c) &= \frac{d}{dw_c} (\cos^2(\theta_i^c) w_c^2 + \sin^2(\theta_i^c)) \\ 4215 \quad &= 2 \cos^2(\theta_i^c) w_c \\ 4216 \quad &\geq 0. \\ 4217 \end{aligned} \tag{377}$$

4219 Let $m = \tilde{M}_\kappa(\cos(\theta_i^c), f_2(w_c))$. Then, the first derivative of G_2 is:
 4220

$$\begin{aligned} 4221 \quad G_2'(w_c) &= G_3'(m) \tilde{M}'_\kappa(\cos(\theta_i^c), f_2(w_c)) f_2^{2\prime}(w_c) \\ 4222 \quad &= \frac{I_{\nu+1}(m)}{I_\nu(m)} \frac{1}{2\tau^2 m} 2 \cos^2(\theta_i^c) w_c \\ 4223 \quad &= \frac{\cos^2(\theta_i^c) w_c}{\tau^2} \frac{1}{m} \frac{I_{\nu+1}(m)}{I_\nu(m)} \\ 4224 \quad &> 0. \\ 4225 \end{aligned} \tag{378}$$

4226 Combining Eq. (374) and Eq. (378), we have:
 4227

$$\begin{aligned} 4228 \quad \bar{\mathcal{J}}'(w_c; \kappa, \nu) &= G'(w_c) = G_1'(t) + G_2'(t) \\ 4229 \quad &= \frac{\cos^2(\theta_i^c)}{\tau} \left(-1 + \frac{w_c}{\tau} \frac{1}{m} \frac{I_{\nu+1}(m)}{I_\nu(m)} \right). \\ 4230 \end{aligned} \tag{379}$$

4231 Since $0 < w_c < 1$, then:
 4232

$$\begin{aligned} 4233 \quad 0 \leq w_c^2 \leq 1 &\Leftrightarrow \sin^2(\theta_i^c) \geq \sin^2(\theta_i^c) w_c^2 \\ 4234 \quad &\Leftrightarrow \sin^2(\theta_i^c) \geq w_c^2 - \cos^2(\theta_i^c) w_c^2 \\ 4235 \quad &\Leftrightarrow \cos^2(\theta_i^c) w_c^2 + \sin^2(\theta_i^c) \geq w_c^2 \\ 4236 \quad &\Leftrightarrow f_2^2(w_c) \geq w_c^2. \\ 4237 \end{aligned} \tag{380}$$

4238 Therefore, consider $\theta_i^c \in [0, \frac{\pi}{2}]$, we have:
 4239

$$\begin{aligned} 4240 \quad m^2 &= \tilde{M}_\kappa^2(\cos(\theta_i^c), f_2(w_c)) \\ 4241 \quad &= \kappa^2 + \frac{2\kappa \cos(\theta_i^c)}{\tau} + \frac{f_2^2(w_c)}{\tau^2} \\ 4242 \quad &\geq \kappa^2 + \frac{2\kappa \cos(\theta_i^c)}{\tau} + \frac{w_c^2}{\tau^2} \\ 4243 \quad &\geq \frac{w_c^2}{\tau^2} \\ 4244 \quad &\geq 0, \\ 4245 \end{aligned} \tag{381}$$

4246 which implies:
 4247

$$m \geq \frac{w_c}{\tau} \Leftrightarrow \frac{w_c}{\tau} \frac{1}{m} \leq 1. \tag{382}$$

4248 Plugging Eq. (375) and Eq. (382) into Eq. (379), we have:
 4249

$$\begin{aligned} 4250 \quad \bar{\mathcal{J}}(w_c; \kappa, \nu) &= \frac{\cos^2(\theta_i^c)}{\tau} \left(-1 + \frac{w_c}{\tau} \frac{1}{m} \frac{I_{\nu+1}(m)}{I_\nu(m)} \right) \\ 4251 \quad &< 0. \\ 4252 \end{aligned} \tag{383}$$

4253 So we can conclude that, for any fixed $\theta_i^c \in [0, \frac{\pi}{2}]$, $\bar{\mathcal{J}}(\cdot)$ is a strictly decreasing function on $(0, 1]$. \square
 4254

4266 **Lemma 15.** Let X be an N -point configuration, where $X = (x_1, \dots, x_N) \in (\mathbb{S}^{h-1})^N$ are iid
 4267 samples from $\mu = \text{vMF}(c, \kappa)$. When κ is sufficiently large, $\forall i, j \in [K], i \neq j$, it holds that:
 4268

$$4269 \quad 4270 \quad P(x_i \cdot x_j \geq 0) \approx 1. \quad (384)$$

4271 *Proof.* Let $X \sim \text{vMF}(c, \kappa)$ on \mathbb{S}^{h-1} and set $U = c^\top X = \cos \Theta \in [-1, 1]$. Then:
 4272

$$4274 \quad 4275 \quad P(X \cdot c \geq 0) = \frac{\int_0^1 e^{\kappa u} (1-u^2)^{\frac{p-3}{2}} du}{\int_{-1}^1 e^{\kappa u} (1-u^2)^{\frac{p-3}{2}} du}. \quad (385)$$

4277 Using standard integral representations of the modified Bessel and modified Struve functions,
 4278

$$4279 \quad 4280 \quad I_\nu(z) = \frac{(z/2)^\nu}{\sqrt{\pi} \Gamma(\nu + \frac{1}{2})} \int_{-1}^1 e^{zt} (1-t^2)^{\nu-\frac{1}{2}} dt, \\ 4281 \quad 4282 \quad \nu(z) = \frac{(z/2)^\nu}{\sqrt{\pi} \Gamma(\nu + \frac{1}{2})} \int_0^1 2 \sinh(zt) (1-t^2)^{\nu-\frac{1}{2}} dt, \quad (386)$$

4285 with $\nu = h/2 - 1$, the ratio simplifies to the neat closed form
 4286

$$4287 \quad 4288 \quad P(X \cdot c \geq 0) = \frac{1}{2} \left(1 + \frac{L_\nu(\kappa)}{I_\nu} \right) \quad (387)$$

4290 where L_ν the modified Struve function. And we list numerical values of this probability:
 4291

- 4292 • $h = 128$:

κ	1	5	10	20	30	50	100	200
P	0.5353	0.6710	0.8117	0.9609	0.9956	1.0000	1.0000	1.0000

- 4297 • $h = 512$:

κ	1	5	10	20	30	50	100	200
P	0.5176	0.5875	0.6708	0.8116	0.9075	0.9863	1.0000	1.0000

- 4301 • $h = 1024$:

κ	1	5	10	20	30	50	100	200
P	0.5125	0.5621	0.6227	0.7340	0.8258	0.9409	0.9991	1.0000

□