The Convergent Representation of Contrastive Vision-Language Models: Geometry, Modality Gap and Shared Space Alignment

Lingjie Yi, Raphael Douady, Chao Chen

Stony Brook University

ICML International Conference On Machine Learning

Fundamental Questions

- **Modality Gap:** Image embeddings (X) and text embeddings (Y) of a contrastive VLM are completely separated in the embeddings space.
- **Q1:** What causes modality gap?
- **Q2:** How modality gap affects downstream performance?

Preliminaries

Definition 1 (MCL Loss). *Let $(X, Y) = (x_i, y_i)_{i=1}^N$ be N paired samples in $\mathbb{S}^{h-1}$. $\forall \tau > 0$, the multimodal contrastive loss $\mathcal{L}_{\text{MCL}}$ is defined as $\mathcal{L}_{\text{MCL}}(X, Y) = \frac{1}{N}\sum_{i=1}^N \mathcal{L}^i_{\text{MCL}}(X, Y)$*

$$\text{where } \mathcal{L}^i_{\text{MCL}}(X,Y) = -\log \frac{\exp(x_i \cdot y_i/\tau)}{\sum_{j=1}^N \exp(x_i \cdot y_j/\tau)} - \log \frac{\exp(x_i \cdot y_i/\tau)}{\sum_{j=1}^N \exp(x_j \cdot y_i/\tau)}.$$

Definition 2 (Modality Gap). *Let $c_x = \frac{\mu_x}{\|\mu_x\|}$ where $\mu_x = \frac{1}{N}\sum_{i=1}^N x_i$ be the center vector of X, $c_y = \frac{\mu_y}{\|\mu_y\|}$ where $\mu_y = \frac{1}{N}\sum_{i=1}^N y_i$ be the center vector of Y. The modality gap between X and Y is then measured by the angular difference between two center vectors: $\Delta_\theta = \cos^{-1}(c_x \cdot c_y)$*

Definition 3 (Pair Alignment and Pair Matching). *Given $(x_i, y_i)_{i=1}^N$, we say (x_i, y_i) **align** if and only if $x_i = y_i$. And we say (x_i, y_i) are **well-matched** if and only if $\forall j \neq i$, $\|x_i - y_i\| < \|x_i - y_j\|$ and $\|x_i - y_i\| < \|x_j - y_i\|$, otherwise (x_i, y_i) are **mis-matched**.*

Definition 4 (vMF Distribution). *$\forall c \in \mathbb{S}^{h-1}$ and $\kappa \geq 0$, let $\nu = h/2 - 1$, the probability density of a random h-dimensional unit vector $z \sim \text{vMF}(c, \kappa)$ is $f_h(z; c, \kappa) = \frac{\kappa^\nu}{(2\pi)^{\nu+1} I_\nu(\kappa)} e^{\kappa c^\top z}$, where I_ν is the modified Bessel function of the first kind.*

Evidence of Representation Constraints

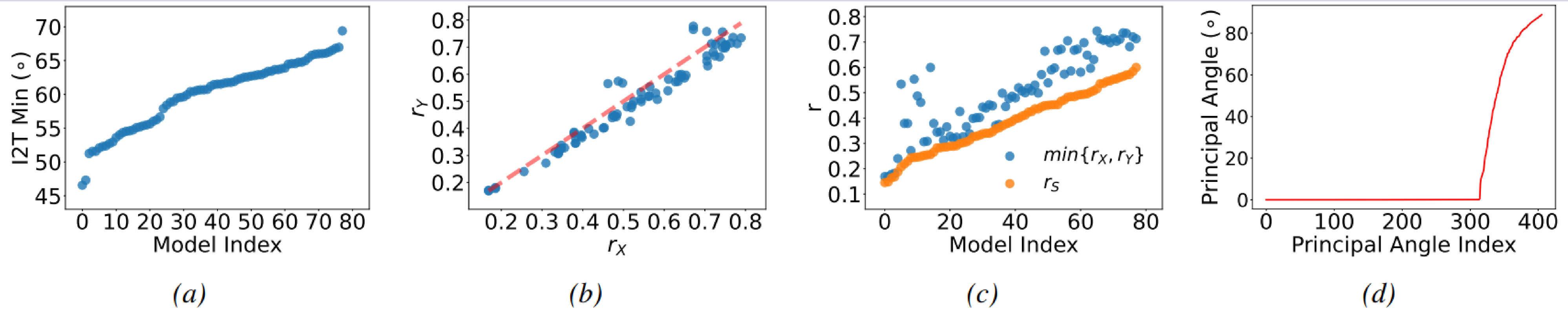

Figure 2. Model Constraints. Results on 78 pretrained VLMs. **(a)**: Sorted minimal angles of all image-text pairs (I2T-min). **(b)**: Effective rank ratio r_X vs. r_Y. Here, $r_{X/Y} = \frac{d_{X/Y}}{h}$ and $d_{X/Y}$ are dimensions of the subspace of X and Y. The dashed line denotes $r_X = r_Y$. **(c)**: Sorted overlapping rank ratio r_S vs. $\min\{r_X, r_Y\}$. Here, $r_S = \frac{d_S}{h}$ and d_S is dimensions of the shared space between X and Y. **(d)**: Principal angles between two collapsed subspaces of OpenCLIP ViT-B/32.

- **Cone Constraint:** I2T_min (> 0) in Fig 2a prove that X and Y lie in **two separated cone**.
- **Dimension Collapse:** $r_{X/Y}$ (< 1) in Fig 2b prove that X and Y **collapse into subspaces**.
- **Subspace Constraint:** $0 < r_S < \min\{r_X, r_Y\}$ in Fig 2c prove that two collapsed subspaces of X and Y are **partially overlapped**. Fig 2d shows principal angles between two collapsed subspaces of OpenCLIP ViT-B/32, further confirming this.

What Causes Modality Gap?

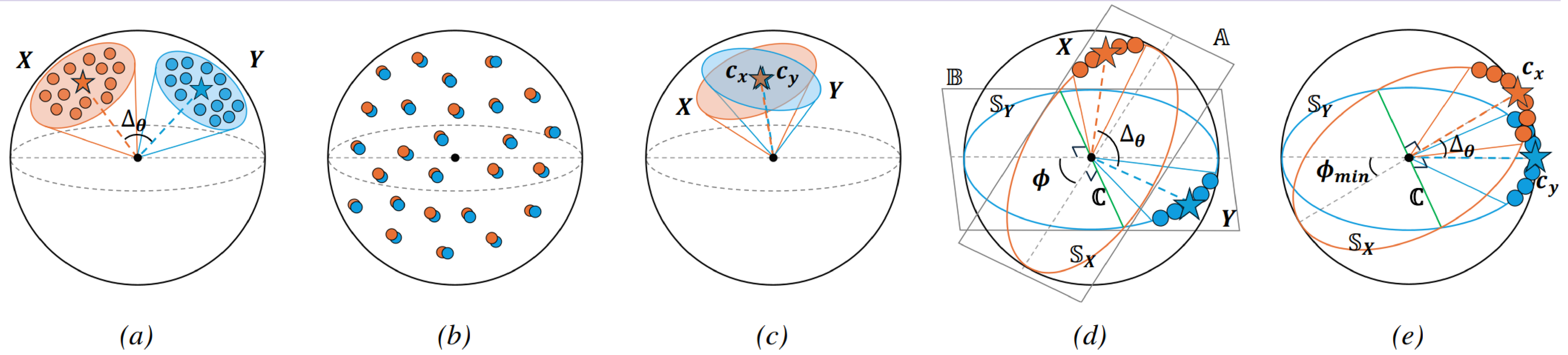

Figure 1. The COR of MCL. Orange and blue dots are X and Y. Starts are centers of X and Y (i.e., c_x, c_y). Δ_θ is the modality gap.

- **Optimal Convergent Representation (COR) :** refers to the optimal configuration of the learned representations of (X, Y) at the global minimum of $\mathcal{L}_{\text{MCL}}$.
- **Model Initialization (Imagine):** X and Y lie in two separated cone (Fig 1a).
- **No Constraint:** (X, Y) follow a paired uniform distribution and $\Delta_\theta = 0$ (Fig 1b, Thm 1).
- **Cone Constraint:** Modality gap Δ_θ still converges to 0 (Fig 1c, Thm 2).
- **Model Initialization (Real):** X and Y lie in two partially overlapping subspace (Fig 1d).
- **With Subspace Constraint:** Modality gap converges to the angel of two hyperplane (or the **weighted average** of principal angles of two subspaces) (Fig 1e, Thm 3).

Conclusion 1: Contrastive objective closes, rather than preserves, modality gap. Subspace geometry, instead of cone geometry, is the structural source of modality gap.

Modality Gap and Downstream Performance

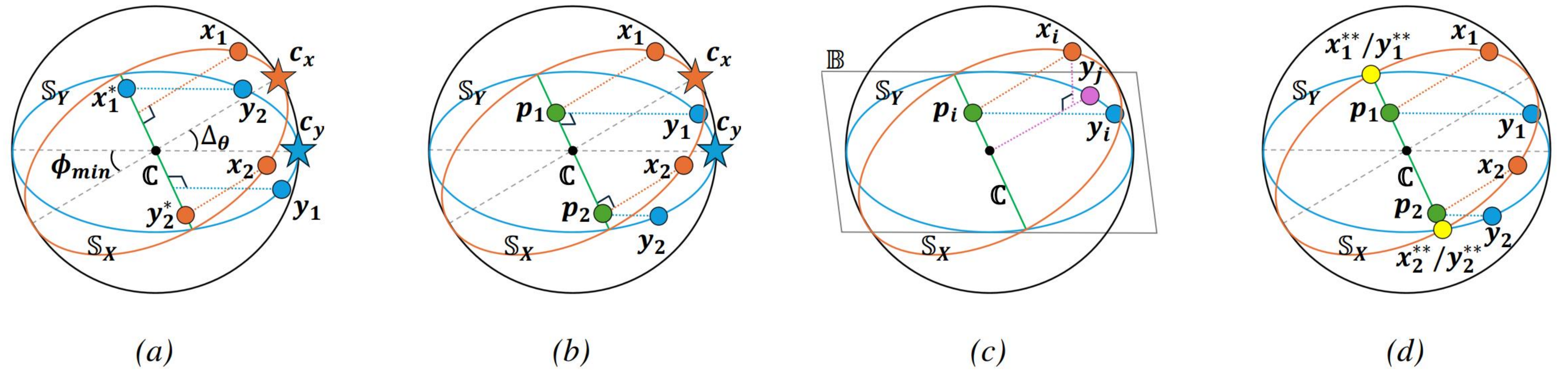

Figure 3. Pair matching in MCL. **(a)**: $c_x, c_y \perp \mathbb{C}$ when $\mathcal{L}^c_{\text{MCL}}$ is minimized. **(b)**: The projections of $(x_i, y_i)_{i \neq c}$ onto $\mathbb{C}$ align to p_i (green points), i.e., $x_i^* = y_i^* = p_i$, when $\mathcal{L}^{i \neq c}_{\text{MCL}}$ is minimized. **(c)**: When $x_i^* = y_i^*$, $P_B x_i \nparallel y_i$. If $\exists y_j$ in an evaluation set, i.e., $y_j = P_B x_i / \|P_B x_i\|$ (purple dot), then $x_i \cdot y_j > x_i \cdot y_i$ or $\|x_i - y_j\| < \|x_i - y_i\|$ and therefore, $(x_i, y_i)_{i \neq c}$ are mis-matched. **(d)**: (x_i^{**}, y_i^{**}) (yellow dots) align.

- **COR of** $(x_i, y_i)_{i \neq c}$: Their projections onto the shared space C align (Fig 3b, Thm4).
- **Pair Matching of** $(x_i, y_i)_{i \neq c}$: $(x_i, y_i)_{i \neq c}$ do not align (Fig 2c) and therefore can be mis-matched (Fig 2c, Coro 1), hence worsen the performance. However, their normalized projections onto C align (Fig 2d, Coro 2).

Empirical Proof of COR

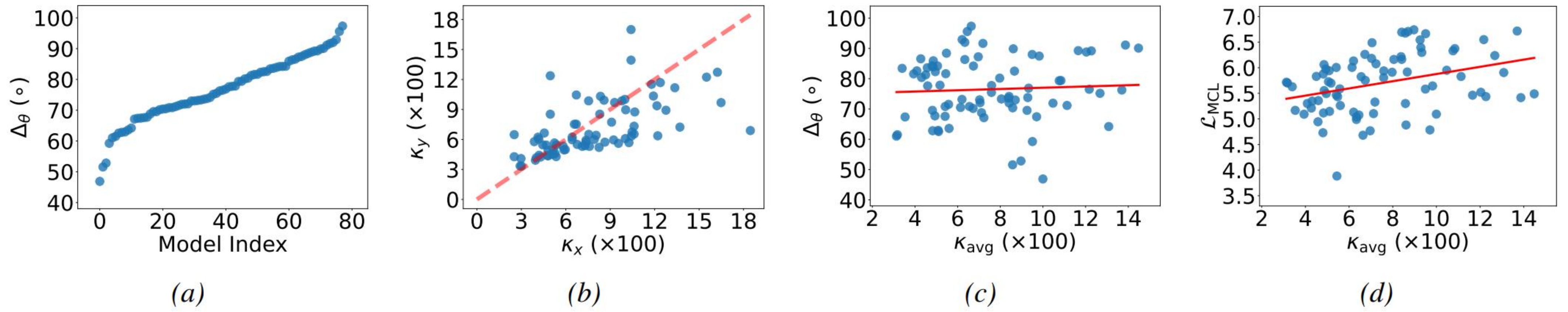

Figure 4. Results of 78 pretrained VLMs on MSCOCO. **(a)**: Sorted values of the modality gap (Δ_θ). **(b)**: Estimated values of κ_x and κ_y. The dashed line denotes $\kappa_x = \kappa_y$. **(c)**: κ_{avg} vs. Δ_θ. **(d)**: κ_{avg} vs. $\mathcal{L}_{\text{MCL}}$. The solid line denotes regression fit.

- **No Constraint:** Δ_θ (> 0) in Fig 4a and $\kappa_{x/y}$ (> 0) in Fig 4b show the unconstrained optimum is **NOT** achieved (**disprove Thm 1**). There exists representation constraints.
- **Cone Constraint: No** significant relationship between Δ_θ and $\kappa_{x/y}$. But $\mathcal{L}_{\text{MCL}}$ is positively correlated with $\kappa_{x/y}$ (Fig 4c, d and Tab 1). It shows **cone effect alone** does not cause modality gap and contrastive objective **promotes intra-modal uniformity**.
- **Subspace Constraint:** Δ_θ is positively correlated with the **weighted average** of principal angles γ_{wavg} . And $\mathcal{L}_{\text{MCL}}$ is negatively correlated to shared space alignment SAlign. (Fig 5a,b and Tab 1). These suggest contrastive objective **promotes shared space alignment.**

What Decides Downstream Performance

Dataset	κ_{avg} vs. Δ_θ	κ_{avg} vs. $\mathcal{L}_{\text{MCL}}$	γ_{wavg} vs. Δ_θ	SAlign vs. $\mathcal{L}_{\text{MCL}}$	SAlign vs. Perf.
MSCOCO	0.047(✗)	0.247(✓)	0.178(✓)	−0.465(✓)	0.733 (✓)
ImageNet	0.140(✗)	0.164(✓)	0.241(✓)	−0.496(✓)	0.689 (✓)

Table 1. Results of 78 pretrained VLMs. Kendall's τ rank correlation between two metrics. ✓ denotes statistical significance in permutation test ($p < 0.05$).

- **Performance:** is positively correlated to SAlign (Fig 5c and Tab 1). On the other hand, additional results show that modality gap has **limited impact** on the performance.

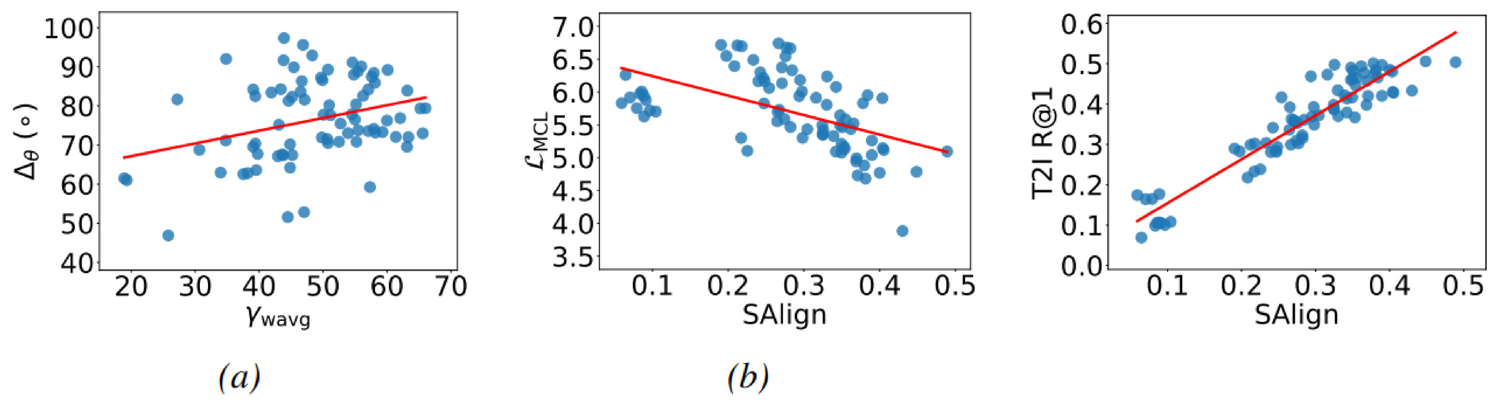

Figure 5. Results of 78 pretrained VLMs on MSCOCO. **(a)**: γ_{wavg} vs. Δ_θ. **(b)**: SAlign vs. $\mathcal{L}_{\text{MCL}}$. **(c)**: SAlign vs. I2T R@1. The solid line denotes regression fit.

Conclusion 2: Theoretically, pair matching is decided by shared space alignment and modality gap. **Empirically**, shared space alignment plays a dominant role in downstream performance.