ROBUST MULTI-MODAL MODELS BY EXPLICITLY OPTIMIZING UNIFORMITY AND ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

We define the robustness of a multi-modal embedding using the margin of one nearest neighbor retrieval and derive the form of the optimal robust embedding (ORE) for various finite dataset sizes (N) and dimensions (d). Specifically, we show that for $N \leq d+1$ the most robust embedding is one where the modalities are aligned and the embeddings form the vertices of a regular simplex. Previous work has shown that the ORE is also the global optimum of the CLIP loss yet gradient-based optimization of the CLIP loss does not converge to the ORE, and instead learns embeddings with a modality gap. Empirically we show that the robustness of CLIP embeddings is far from that of the ORE and is below that of simple baselines. To alleviate this failure, we analyze an alternative loss, Explicit Multi-Modal Uniformity and Alignment (EMMUA) and show that its global optimum is also the ORE. We show both theoretically and experimentally that gradient-based optimization on the EMMUA loss does not learn a modality gap and finds significantly more robust embeddings than those obtained with the CLIP loss.

1 INTRODUCTION

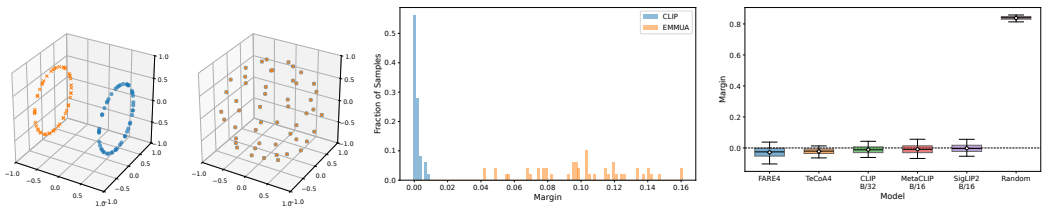
Multi-modal models (Radford et al., 2021; Zhai et al., 2023) trained with the contrastive loss (van den Oord et al., 2019; Sohn, 2016; Zhang et al., 2022) have shown great utility in combining various modalities into useful representations, often integrated into state-of-the-art VLMs (Liu et al., 2023) and image generation models (Nichol et al., 2022; Ramesh et al., 2022; Rombach et al., 2022). This usefulness relies on the assumption that these models learn a representation of the data which aligns the different modalities, as expressed by their impressive zero-shot capabilities (Radford et al., 2021).

However, despite their impressive performance, multi-modal models have been shown to suffer from brittle representations (Tu et al., 2023; Schlarmann et al., 2024; Mao et al., 2023), partially due to the *modality gap* (Liang et al., 2022) - a setting where multi-modal models embed different modalities in linearly separable regions of the hypersphere. Figure 1 illustrates this effect in three dimensions: when a bimodal embedding is found by minimizing the CLIP loss of the embeddings (Shi et al., 2023), the embedding that emerges places the two modalities in distinct rings (1a), rather than spreading the embeddings of both modalities uniformly over the sphere (1b).

In this paper we follow Chowers et al. (2026) in investigating the effect of the modality gap on robustness. In Figure 1c we plot a histogram of the cosine margin $\boldsymbol{x}_i^T \boldsymbol{y}_i - \max_{j \neq i} \boldsymbol{x}_i^T \boldsymbol{y}_j$, i.e. the difference in cosine similarities between a correct retrieval and the hardest negative example. It can be seen that for embeddings obtained by minimizing the CLIP loss, the margin is much smaller than that of the simple baseline where the two modalities are aligned and the embeddings are sampled uniformly on the unit sphere. A similar picture emerges when we compute the cosine margin for pretrained models: the embeddings obtained by minimizing the CLIP loss have margins close to zero, while the random baseline achieves margins close to 0.8.

The small margins obtained by minimizing the CLIP loss are surprising since they seem to contradict the goal of training with a contrastive loss which aims to increase the similarity of positive pairs while decreasing the similarity of negative pairs. For example, in the case of text and images, contrastive losses try to push the embedding of a given image and its corresponding text towards each other, while simultaneously pushing the embedding of an image away from those of all non-corresponding texts. Intuitively, such a loss should also increase the cosine margin of each embedding.

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



(a) Modality Gap (b) Uniform (c) Margins (d) Pretrained Margins

Figure 1: Optimizing the CLIP loss on the unit sphere in 3D leads to a modality gap (1a) with smaller cosine margins than the baseline of uniform and aligned embeddings (1b). Similarly, for pretrained models on the MS-COCO evaluation set (1d), the random baseline, which consists of uniformly sampling $\boldsymbol{x}_i$ on the hypersphere and setting $\boldsymbol{y}_i = \boldsymbol{x}_i$, consistently outperforms models trained with CLIP, even for FARE and TeCoA, which are adversarially trained.

In this paper, we show that this failure is primarily a *failure of optimization.* We show that the embedding that is optimal in terms of robustness (the optimal robust embedding, or ORE) is also the *global minimum* of the CLIP loss. Thus if CLIP could find the global minimum of its loss, it would find a robust embedding. Yet it does not, presumably due to the interaction between initialization and optimization (Zhang et al., 2024; Chowers et al., 2026). Therefore, a good candidate for a loss function for multi-modal learning would be one with the same global minima as CLIP, but without the obstacles in optimization.

Our main contribution is analyzing such a loss. We focus on a recently proposed loss (Yin et al., 2026b) which explicitly minimizes multi-modal uniformity and alignment (EMMUA). We show that the global minima of the EMMUA loss are also the ORE, maximally separating true and negative examples. Importantly, we show that the gradient descent dynamics of training with EMMUA close the modality gap by overcoming the various shortcomings of the CLIP loss (namely dimensionality collapse of the embeddings and the maintenance of the gap throughout training (Zhang et al., 2024; Yi et al., 2026)). As a result, we show that gradient based optimization of the EMMUA loss yields embeddings that are much more robust to different types of noise and learn embedding spaces with much larger margins than those of CLIP.

2 Related Works

Many works have tried to mitigate the modality gap (Liang et al., 2022) by augmentations (Oh et al., 2023; Mager et al., 2026), adjusting backbone architecture (Eslami & de Melo, 2024), adding regularization during training or finetuning (Fahim et al., 2024; Yi et al., 2026; Zhou et al., 2026), and applying post training algorithms (Zhang et al., 2023; 2024; Yamashita et al., 2025). Still, the gap remains even when accounting for data imbalance (Schrodi et al., 2025), suggesting that it is difficult to avoid the modality gap when optimizing the CLIP loss due to biases in the optimization process (Chowers et al., 2026).

Finding the globally optimal configurations in terms of the CLIP loss has been of great interest (Wang & Isola, 2020; Yin et al., 2026a;b; Yi et al., 2026). Recent works have shown that this problem is related to a larger set of problems on the hypersphere such as sphere packing, minimizing energy (Thomson, 1904) and neural collapse in classification (Papyan et al., 2020), many of which are solved for $N \leq d + 1$ by the regular simplex (Liu et al., 2021; Lu & Steinerberger, 2022; Sreenivasan et al., 2023). We continue this line of work, specifically from the perspective of the robustness of the embedding space.

In the unimodal training setting, when the number of examples goes to infinity, many losses are optimized when the embeddings are uniformly spread over the hypersphere (Chen et al., 2020; Wang & Isola, 2020; Bardes et al., 2022). This property has proven to be useful (Wang & Isola, 2020; Garrido et al., 2023), specifically in noise suppression (Wu et al., 2025). We expand on these by observing that when embeddings are equally spread on the hypersphere they are also maximally robust, even in the multi-modal case.

Much of our investigation follows the new loss proposed by Yin et al. (2026b), primarily for the case where the number of modalities is greater than 2. However, in this work we show its advantages from an optimization standpoint - learning with EMMUA circumvents the problems of training the CLIP loss with gradient descent (Chowers et al., 2026) and this leads to significantly more robust embeddings (even when there are only two modalities).

Our work follows up on the recent paper of Chowers et al. (2026), who showed that the modality gap decreases robustness and suggested improving the robustness of trained models by shifting one modality towards the other by a global shift vector. However, as we show in this paper, postprocessing pretrained models yields only a modest improvement in robustness, especially for large datasets. In contrast, we suggest an alternative loss that causes the embeddings to approach the ORE and the improvements in robustness are significantly higher.

3 THEORY

3.1 PRELIMINARIES

Given an embedding of a data point and the embedding of its augmentation $(\boldsymbol{x}_i, \tilde{\boldsymbol{x}}_i)$, the unimodal contrastive loss (or Info-NCE loss) (van den Oord et al., 2019; Chen et al., 2020) is defined as:

$$\mathcal{L}_{NCE}(\mathcal{X}) = -\sum_{i=1}^{N} \log\left(\frac{e^{\boldsymbol{x}_i^T \tilde{\boldsymbol{x}}_i/\tau}}{\sum_j e^{\boldsymbol{x}_i^T \boldsymbol{x}_j/\tau}}\right) \tag{1}$$

where τ is a learned temperature parameter. This loss has been extended to the multi-modal case, where given pairs of data embeddings $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ from two modalities, $\mathcal{X}, \mathcal{Y}$ the multi-modal contrastive loss (Zhang et al., 2022), or CLIP loss (Radford et al., 2021) is defined as:

$$\mathcal{L}_{CLIP}(\mathcal{X}, \mathcal{Y}) = -\sum_{i=1}^{N} \left(\log\left(\frac{e^{\boldsymbol{x}_i^T \boldsymbol{y}_i/\tau}}{\sum_j e^{\boldsymbol{x}_i^T \boldsymbol{y}_j/\tau}}\right) + \log\left(\frac{e^{\boldsymbol{x}_i^T \boldsymbol{y}_i/\tau}}{\sum_j e^{\boldsymbol{y}_i^T \boldsymbol{x}_j/\tau}}\right)\right) \tag{2}$$

Since the embeddings are constrained to lie on the unit sphere, the CLIP loss can also be written:

$$\mathcal{L}_{CLIP}(\mathcal{X}, \mathcal{Y}) = -\sum_{i=1}^{N} \left(\log\left(\frac{e^{-\frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{y}_i\|^2/\tau}}{\sum_j e^{-\frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{y}_j\|^2/\tau}}\right) + \log\left(\frac{e^{-\frac{1}{2}\|\boldsymbol{y}_i - \boldsymbol{x}_i\|^2/\tau}}{\sum_j e^{-\frac{1}{2}\|\boldsymbol{y}_i - \boldsymbol{x}_j\|^2/\tau}}\right)\right) \tag{3}$$

Wang & Isola (2020) have shown that the unimodal contrastive loss (van den Oord et al., 2019; Chen et al., 2020) can be viewed as the sum of two terms - the first aligning positive pairs, i.e. minimized when $\forall i : \boldsymbol{x}_i = \tilde{\boldsymbol{x}}_i$ and the second pushing them to be uniformly distributed over the hypersphere:

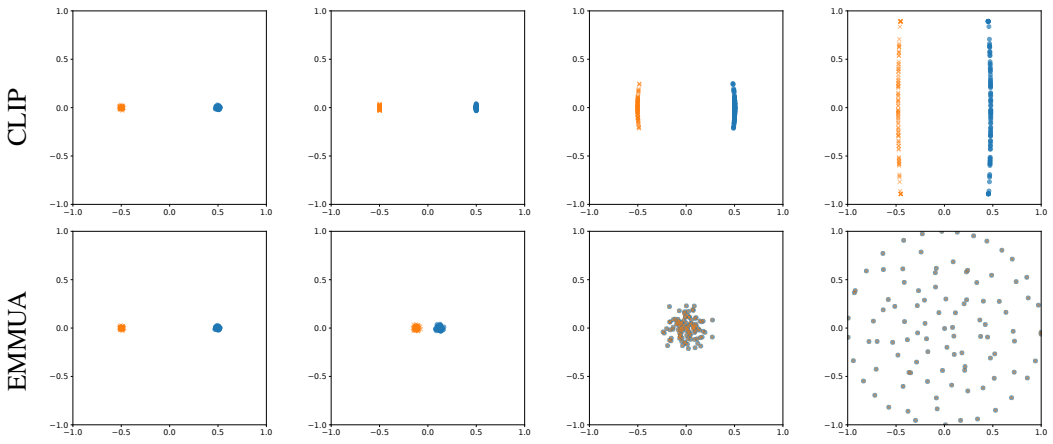
$$\mathcal{L}_{NCE}(\mathcal{X}) = -\sum_{i=1}^{N} \frac{\boldsymbol{x}_i^T \tilde{\boldsymbol{x}}_i}{\tau} + \sum_{i=1}^{N} \log \sum_{j=1}^{N} e^{\boldsymbol{x}_i^T \boldsymbol{x}_j/\tau} = \mathcal{L}_{\text{align}}(\mathcal{X}) + \mathcal{L}_{\text{uniform}}(\mathcal{X}) \tag{4}$$

In the multi-modal case the loss can also be viewed as the sum of an alignment term, minimized when $\forall i : \boldsymbol{x}_i = \boldsymbol{y}_i$, and a second term that is sometimes mistakenly also termed "uniformity" (see Section A.1).

3.2 KNOWN RESULTS ON THE CLIP LOSS

Even though the CLIP loss contains terms that reward alignment and uniformity, it has been repeatedly reported that gradient based optimization usually converges to a modality gap where each modality occupies a distinct narrow cone in the embedding space (Liang et al., 2022) - a clearly non-uniform and non-aligned solution. This gap is often characterized by the distance between the modality means being much greater than zero ($\boldsymbol{g} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i - \boldsymbol{y}_i$ with $\|\boldsymbol{g}\| > 0$) and the two modalities being linearly separable.

At the same time, it has been shown that the *global optimum* of the loss is a solution with no gap. In fact, for N pairs embedded in d dimensions on the unit sphere ($N \leq d+1$), the global minimum of the loss (for any finite temperature τ) is when $\forall i : \boldsymbol{x}_i = \boldsymbol{y}_i$ and the points form the vertices of the



(a) Step 1 (b) Step 250 (c) Step 1000 (d) Step 5000

Figure 2: We optimize embeddings using the CLIP loss (top row) and EMMUA loss (bottom) starting from the same initialization, imitating CLIP's "narrow cones" initialization (Liang et al., 2022). While CLIP converges to a modality gap, EMMUA converges to a uniform and aligned solution.

regular simplex (Sreenivasan et al., 2023), a set of N points on the unit sphere for which the distance between any two points is the same. This geometric configuration also arises as the optimal solution in the unimodal case under the same setting (Lu & Steinerberger, 2022; Liu et al., 2021).

Gradient-based optimization of the CLIP loss fails to find this optimal solution, even for small N's, due to two previously reported effects:

- **Dimensionality Collapse** (Chowers et al., 2026; Yi et al., 2026). Even if each modality is initialized with a distribution that is full-rank, the initial steps of gradient descent lead to an embedding where each modality occupies only a small number of dimensions.
- **Gap Constancy** (Zhang et al., 2024). After a certain number of iterations, gradient descent does not change the embeddings in the gap direction, and the embeddings are only optimized in directions orthogonal to the gap.

Figure 2 (top, following Chowers et al. (2026)) illustrates these known results in the case of optimizing the CLIP loss inside the unit circle in two dimensions. Even though each modality is initialized with a two dimensional distribution, the initial iterations cause each modality to collapse to a one dimensional distribution. This is true even though training used the original CLIP's low temperature ($\tau = 0.07$), which encourages alignment.

Furthermore, after a certain number of iterations, the gradient dynamics only move the embeddings in the direction orthogonal to the gap, leading to a suboptimal solution (even when $N > d + 1$ and see Section A.2). In this figure, we use the ℓ_2 based version of the CLIP loss (equation 3) since we allow the embeddings to be inside the unit circle, rather than constraining them to lie on the unit sphere. Similar results have been reported when the embeddings are constrained to be on a high dimensional sphere (Shi et al., 2023; Zhang et al., 2024; Chowers et al., 2026; Yi et al., 2026).

These two properties of the dynamics of gradient descent on the CLIP loss have also been proven theoretically under certain assumptions. Specifically, dimensionality collapse will provably occur when the initialization satisfies the **Narrow Cones Assumption** (Liang et al., 2022) (A.10) and gap constancy will provably occur when there exists an iteration for which the embedding has an **orthogonal gap** (Zhang et al., 2023) (A.2) and the logit matrices satisfy **double stochasticity** (Chowers et al., 2026) (A.13). See the appendix for a formal definition of these three assumptions.

3.3 Optimal Solution to the CLIP Loss is Maximally Robust

Our first theoretical result shows that the global minimum of the CLIP loss, the perfectly aligned configuration of the regular simplex, is also the optimally robust embedding.

Definition 3.1. For a query point $\boldsymbol{x}_i$ and a set of possible retrievals $\{\boldsymbol{y}_j\}$ on the unit sphere the **retrieval margin** of the nearest neighbor at $\boldsymbol{x}_i$ is defined as:

$$\text{margin}(\boldsymbol{x}_i, \{\boldsymbol{y}_j\}) = \langle \boldsymbol{x}_i, \boldsymbol{y}_i \rangle - \max_{j \neq i} \langle \boldsymbol{x}_i, \boldsymbol{y}_j \rangle \tag{5}$$

An **optimally robust embedding** (ORE) is a pair $(\mathcal{X}^*, \mathcal{Y}^*)$ that solves the ORE problem - the maximum average margin over all queries:

$$ORE(d, N) = R(\mathcal{X}^*, \mathcal{Y}^*) = \max_{\mathcal{X}, \mathcal{Y} \in \mathbb{S}^{N \times d}} R(\mathcal{X}, \mathcal{Y}) = \max_{\mathcal{X}, \mathcal{Y} \in \mathbb{S}^{N \times d}} \frac{1}{N} \sum_{i=1}^{N} \text{margin}(\boldsymbol{x}_i, \mathcal{Y}) \tag{6}$$

The motivation for this definition of robustness is that having a large margin guarantees the robustness of cross modality retrieval, even when an adversary is allowed to perturb the embeddings. This is formalized in the following theorem (proof in Section A.3.3):

Theorem 3.2. *If* $\text{margin}(\boldsymbol{x}_1, \{\boldsymbol{y}_i\}) = \delta$ *then for any* $\boldsymbol{v}$ *such that* $\|\boldsymbol{v}\| < \frac{\delta}{2}$, $\boldsymbol{x}_1 + \boldsymbol{v}$ *is closer to* $\boldsymbol{y}_1$ *than it is to* $\boldsymbol{y}_i$ *in terms of cosine similarity:* $\langle \boldsymbol{x}_1 + \boldsymbol{v}, \boldsymbol{y}_1 \rangle > \langle \boldsymbol{x}_1 + \boldsymbol{v}, \boldsymbol{y}_i \rangle$ *for all* $i \neq 1$.

Using this definition of robustness, we now solve the ORE problem equation 6 when the number of points N satisfies $N \leq d + 1$.

Theorem 3.3. *Assume* $N \leq d + 1$*, then the ORE configuration (3.1)* $\mathcal{X} = \{\boldsymbol{x}_i\}, \mathcal{Y} = \{\boldsymbol{y}_j\}$ *is when* $\boldsymbol{x}_i = \boldsymbol{y}_i$ *and they are vertices of the regular simplex. This is the same configuration that provides the optimal solution to the CLIP loss.*

For full proof see Section A.3.2. For larger N we do not know the exact form of the ORE, but we can obtain an upper bound on its average margin by noting that as the number of pairs grows they are forced to be packed more densely on the hypersphere, therefore decreasing the margin. This is formalized in the following theorem (proof in Section A.3.1):

Theorem 3.4. *For a fixed dimension* d*, the average margin of the optimal robustness embedding* $ORE(d, N)$ *decreases with* N.

In summary, for $N \leq d + 1$ solving either the ORE problem or minimizing the CLIP loss lead to the same solution - a perfectly aligned embedding space realized by the vertices of the regular simplex, while for larger values of N both solutions are unknown. But due to training dynamics, optimizing the CLIP loss leads to a modality gap and does not find the ORE. Can we use another loss that has the same global minima but is easier to optimize?

3.4 Explicitly Optimizing Uniformity and Alignment

Since CLIP does not converge to its global minimum, which is also the Optimal Robust Embedding, our goal is to find a new loss function with the same global minimum only with different training dynamics, thus enabling convergence to a robust solution. We argue that the simplest way is to replace the uniformity term of the multi-modal contrastive loss with that of the unimodal loss (Yeh et al., 2022; Wang & Isola, 2020), as was recently proposed by Yin et al. (2026b). We call this loss "Explicit Multi-Modal Uniformity and Alignment" (EMMUA):

$$\mathcal{L}_{EMMUA} = \frac{1}{N} \sum_{i=1}^{N} \lambda \left\| \boldsymbol{x}_i - \boldsymbol{y}_i \right\|^2 + \frac{1}{2} \log \left(\sum_{j \neq i} e^{\frac{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\tau}} \right) + \frac{1}{2} \log \left(\sum_{j \neq i} e^{\frac{-\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2}{\tau}} \right) \tag{7}$$

The above contains three terms - the first encouraging alignment between positive pairs and the other two pushing for a uniform distribution of the embeddings of each modality on the hypersphere, with λ controlling the balance between the two (Wang & Isola, 2020; Chen et al., 2021).

When the number of points goes to infinity, it is straighgtforward to see that this loss benefits from having the same global minima as the unimodal contrastive loss, i.e. when $\mathcal{X} = \mathcal{Y}$ and they are uniformly distributed on the hypersphere. In the finite case of $N \leq d + 1$, we now prove that this loss has the same minima as the CLIP loss and ORE problem:

Theorem 3.5. *For* $N \leq d + 1$*, any* λ*, and any finite temperature* τ*, the unique global minimum (up to rotation and permutation) of* $\mathcal{L}_{EMMUA}$ *(Eq. (7)) is attained when* $\forall i \in [N] : \boldsymbol{x}_i = \boldsymbol{y}_i$ *and the points form the vertices of the regular simplex. This configuration is also the ORE.*

Proof. Each uniformity term is minimized when $\{\boldsymbol{x}_i\}$ and $\{\boldsymbol{y}_i\}$ are the vertices of some regular simplex (up to rotation / permutation) (Lu & Steinerberger, 2022). Setting them to be the same simplex, with $\boldsymbol{x}_i = \boldsymbol{y}_i$ also minimizes the first term. □

3.5 Optimization of EMMUA

More importantly, the gradient descent dynamics of minimizing EMMUA do not trap the learned embeddings in a modality gap, in stark contrast to those of the CLIP loss. This is true even when the two modalities satisfy either of the three previously discussed assumptions: narrow cones, orthogonal gap, or double stochasticity (Sections A.10, A.2, A.13 respectively) - all previously reported to contribute to the formation and maintenance of the modality gap in CLIP (Liang et al., 2022; Zhang et al., 2024; Chowers et al., 2026; Yi et al., 2026). We formally prove this in the following theorem:

Theorem 3.6. *Contrary to the CLIP loss, optimizing EMMUA (7) with gradient descent shrinks the modality gap under any of the following previously reported conditions: narrow cones initialization (A.10), an orthogonal gap (A.2) or double stochasticity of the loss's logit matrix (A.13).*

Proof Sketch. The main idea of the proof follows from the difference between the gradients of EMMUA and the CLIP loss:

$$\nabla_{\boldsymbol{y}_i}\mathcal{L}_{CLIP} \propto 2(\boldsymbol{x}_i - \boldsymbol{y}_i) + \sum_{j\neq i}(\boldsymbol{Q}^{\mathcal{X}}(i,j) + \boldsymbol{Q}^{\mathcal{X}}(j,i))(\boldsymbol{y}_i - \boldsymbol{x}_j)$$
$$\nabla_{\boldsymbol{y}_i}\mathcal{L}_{EMMUA} \propto 2(\boldsymbol{x}_i - \boldsymbol{y}_i) + \sum_{j\neq i}(\boldsymbol{Q}^{\mathcal{Y}}(i,j) + \boldsymbol{Q}^{\mathcal{Y}}(j,i))(\boldsymbol{y}_i - \boldsymbol{y}_j) \tag{8}$$

where the $\boldsymbol{Q}$'s are the assignment probabilities for $\boldsymbol{y}_i$ to the modalities $\mathcal{X}, \mathcal{Y}$. While both gradients push $\boldsymbol{y}_i$ towards its pair $\boldsymbol{x}_i$, this term is countered in CLIP's gradient by the other term which pushes $\boldsymbol{y}_i$ away from the $\boldsymbol{x}_j$'s cluster, where the balance between the two is determined by $\boldsymbol{Q}^{\mathcal{X}}(j,i) + \boldsymbol{Q}^{\mathcal{X}}(j,i)$. On certain occasions, this can push $\boldsymbol{y}_i$ away from the $\mathcal{X}$ modality overall. In EMMUA's gradient there is no such counter effect - the gradient pushes away from the other $\boldsymbol{y}_j$'s and towards $\boldsymbol{x}_i$. □

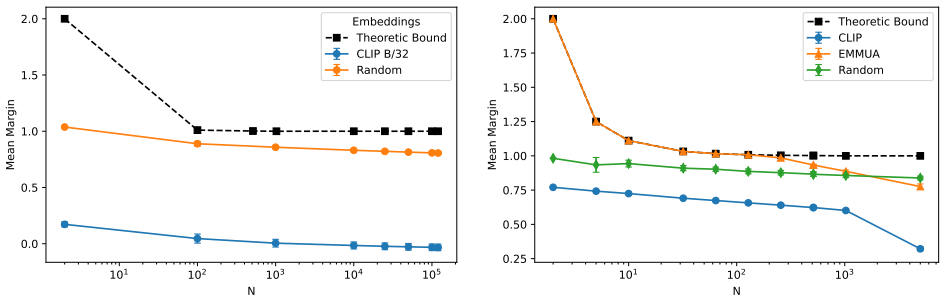
A simple example of this can be seen in the bottom of Fig. 2. We optimize embeddings with the EMMUA loss, and compare them to those trained with the CLIP loss under the same initialization and hyperparameters. While the CLIP trained embeddings are trapped in a modality gap configuration, those of the EMMUA loss learn a perfectly aligned solution.

4 Experiments

We set out to test our theory in simulation and practice. Ideally, models trained with EMMUA will converge to solutions without a modality gap, for which the margin is much larger than that of CLIP, without affecting clean performance. In all experiments, when comparing the CLIP loss (2) to EMMUA (7) we train two models with the same initialization, hyperparameters and seed, thus controlling for all factors but the loss.

4.1 Simulations

We first focus on inspecting the margins of embeddings learned by EMMUA. According to Theorem 3.3 and Theorem 3.5, we know that for N pairs in $d \geq N-1$ dimensions, EMMUA should converge to the regular simplex, thus maximizing the average margins. To inspect this property of both EMMUA and CLIP, we directly optimize embedding vectors on the unit hypersphere using both losses, for a constant dimension commonly used in CLIP ($d = 512$), and measure the learned margins. We compare these with a random baseline, consisting of sampling N vectors $\boldsymbol{x}_i$ from the uniform distribution on the hypersphere, and setting $\boldsymbol{y}_i = \boldsymbol{x}_i$. This baseline is intended to simulate a perfectly aligned and uniform embedding space. The results are presented in Fig. 3b - even when $N \leq d+1$, for which CLIP and EMMUA have the same global minima, only EMMUA reaches the theoretical upper bound. For larger N's, EMMUA still outperforms CLIP, performing similarly to the random baseline.



(a) Pretrained (b) Optimized Embeddings

Figure 3: We compare the margins of CLIP trained models to those of uniformly sampled points and embeddings trained with EMMUA. We calculate the margins for subsets of N sample pairs, both for a pretrained CLIP B/32 on MSCOCO (3a) and for embeddings optimized with either CLIP loss or EMMUA (3b). In both cases, CLIP is worse than the random baseline. When optimizing the embeddings, EMMUA is much closer to the theoretical bound. See Section B for further information.

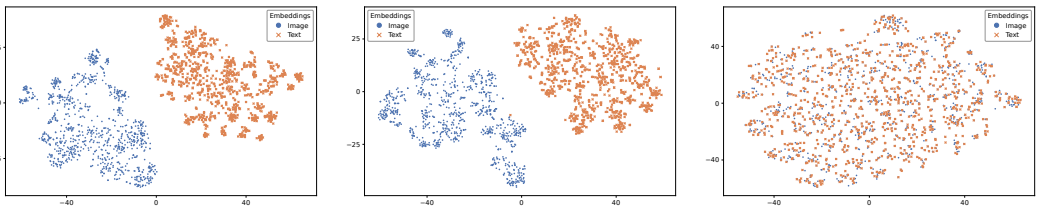
We also measure the margins for a pretrained model in Fig. 3a, using a pretrained CLIP B/32 and measuring on randomly sampled pairs from the MS-COCO training set. Again, CLIP's margins are much smaller than those of the theoretical bound and random baseline.

4.2 Geometry of Learning with EMMUA

Seeing that EMMUA can indeed learn embedding spaces that overcome the modality gap and have large margins, we turn our attention to training CLIP variants from scratch on various datasets. We follow Schrodi et al. (2025) and train on the Morpho-MNIST based (Castro et al., 2019) MAD dataset. In this setting, each augmented image is matched with a caption describing the augmentations and digit. Additionally, we inspect models trained on CIFAR100 and ImageNet, where the text captions are "a photo of a CLS". Finally, we finetune a pretrained CLIP B/32 on MS-COCO with both losses.

The results of these experiments are summarized in Table 1. Across all datasets, the two losses yield models with comparable performance. However, while CLIP learns a modality gap, EMMUA does not - the gap (distance between modality means) is much smaller, the true pairs have higher cosine similarity and the two modalities cannot be distinguished by a linear classifier. This is also apparent when visualizing the embedding space, as can be seen in Fig. 4.

Table 1 also shows that the text embeddings of EMMUA are much more uniform than those of CLIP. However, this does not generally hold for images. This is because the optimal solution for both losses is when all images of the same class converge to a single point, which reduces uniformity, since the uniformity measure also accounts for intra-class distances.



(a) Pretrained (b) Finetuned (CLIP) (c) Finetuned (EMMUA)

Figure 4: t-SNE visualizations of the embedding space of a pretrained CLIP B/32 and the model finetuned on MS-COCO with the CLIP loss or EMMUA. Finetuning with EMMUA is enough to close the gap. More visualizations can be found in Section C.2.

Dataset	Loss	Accuracy / R@1	Modality Gap Norm	Linear Classifier %	$\mathcal{U}_{\text{img}}$	$\mathcal{U}_{\text{txt}}$
MAD	CLIP	94.2%	0.642	100.0%	2.52	-4.72
	EMMUA	94.0%	0.011	50.3%	2.47	-18.54
CIFAR-100	CLIP	59.3%	0.356	94.9%	-2.78	-12.31
	EMMUA	56.3%	0.043	60.3%	-0.52	-20.71
ImageNet	CLIP	55.0%	0.572	99.0%	-8.45	-11.58
	EMMUA	57.3%	0.027	54.8%	-3.91	-14.18
MS-COCO (finetune)	CLIP	43.0%	0.642	99.9%	-5.56	-6.93
	EMMUA	40.9%	0.017	57.7%	-11.40	-10.17

Table 1: EMMUA's embedding space is both more uniform and aligned than CLIP's. We measure different aspects of performance and the modality gap for models trained on different datasets with the CLIP loss and EMMUA. Uniformity is measured as in equation 7. While clean performance is comparable, EMMUA learns an embedding space without a modality gap, for which the performance of a linear classifier is close to chance.

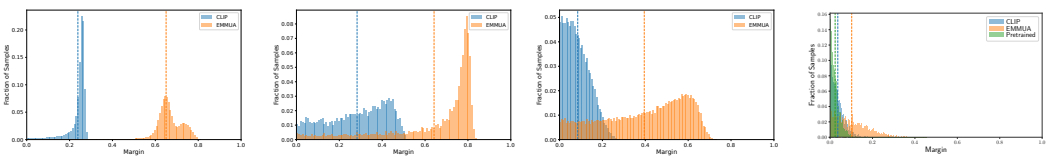
4.3 EMMUA IS MORE ROBUST THAN CLIP

After verifying that EMMUA indeed closes the gap, we now compare its robustness with that of CLIP. As expected, EMMUA indeed learns embedding spaces that have margins much larger than CLIP. Figure 5 measures this - we compute the per sample margin over each dataset's evaluation set and plot them in a histogram. It is evident that EMMUA learns an embedding space with much larger margins, while CLIP's margins stay comparable to the pretrained model's.

To confirm that these margins translate to embedding space robustness, we compare EMMUA and CLIP's embedding space robustness. We subject the embedding space to two types of noise - gaussian noise, where we add i.i.d. samples from a gaussian to either the text or image embeddings, and quantization noise generated by quantizing both modalities' embedding coordinates to one of L different values. In both cases retrieval is measured using ℓ_2 distance, as noisy embeddings are not guaranteed to lie on the hypersphere. Table 2 presents the accuracy of the models under gaussian noise added to the image embeddings and quantization applied to both modalities. As can be seen, EMMUA produces a much more robust embedding space relative to CLIP, even when using post-processing algorithms intended to increase robustness (Chowers et al., 2026).

5 LIMITATIONS AND EXTENSIONS

In this paper we have used a notion of robustness which allows an adversary to perturb the embeddings by at most δ and this directly leads to the definition of margin that we used. While this is exactly the case when quantizing embeddings, for example in large databases, perhaps a more natural definition of robustness is one in which the adversary is allowed to perturb the inputs to the model with a fixed perturbation budget. Since we have no guarantee on the Lipschitz constants of the encoders, even a tiny perturbation of the input may yield a large perturbation of the embeddings (see Section A.3.4).



(a) MAD (b) CIFAR100 (c) ImageNet (d) MS-COCO (finetuned)

Figure 5: We compare models trained with CLIP and EMMUA on MAD, CIFAR100, ImageNet and finetuning on MS-COCO. We plot the margins of the correctly classified samples from each evaluation set. Across the different datasets, EMMUA exhibits much larger margins, creating a model more robust to embedding space noise. Exact margin means can be found in Table 2.

Dataset	Model	Mean Margin (correct)	Clean Acc.	Quantization Acc.			Gaussian Acc.		
				L=2	L=4	L=16	σ=0.1	σ=0.15	σ=0.2
MAD	CLIP	0.239	**94.2**	93.1	93.6	**94.2**	93.8	89.4	73.5
	CLIP + Shift	0.238	**94.2**	93.8	93.8	**94.2**	93.8	89.3	73.5
	EMMUA	**0.650**	94.0	**94.0**	**94.0**	94.0	**93.9**	**93.8**	**90.4**
CIFAR-100	CLIP	0.283	**59.3**	**59.0**	**58.9**	**59.5**	57.5	53.6	45.5
	CLIP + Shift	0.271	59.2	58.6	58.6	59.2	57.0	52.9	44.8
	EMMUA	**0.642**	56.3	56.2	56.2	56.3	**55.9**	**55.2**	**52.9**
ImageNet	CLIP	0.085	55.0	47.0	47.0	49.4	30.9	14.9	7.4
	CLIP + Shift	0.085	54.5	48.3	48.3	51.7	31.0	15.2	7.5
	EMMUA	**0.395**	**57.3**	**57.0**	**57.0**	**57.2**	**56.0**	**53.6**	**45.5**
MS-COCO (finetune)	Pretrained	0.024	32.6	16.1	16.2	12.6	3.0	1.0	0.4
	CLIP	0.036	**43.0**	24.5	24.5	26.1	7.2	2.2	1.0
	CLIP + Shift	0.036	42.0	24.7	24.9	30.1	7.3	2.4	1.0
	EMMUA	**0.102**	40.9	**37.8**	**37.7**	**40.7**	**30.8**	**19.8**	**10.5**

Table 2: EMMUA is consistently more robust to embedding space noise. We compare the accuracy of the models under different strengths of quantization and Gaussian noise for models trained with EMMUA, CLIP and those same CLIP models with a post-processing closing of the modality gap using the algorithm from Chowers et al. (2026). For Gaussian noise, results are an average over 100 different samples of i.i.d. gaussian noise added to the image embeddings. We also present accuracy when adding noise to the text embeddings, as well as robustness measurements - i.e. the percent of samples whose retrieval under noise is consistent to the clean retrieval - in Section C.3.

Thus, when we measure robustness to perturbations of the input, training with EMMUA yields only a very modest improvement in robustness relative to training with CLIP (Section C.4). In future work, it could be interesting to combine EMMUA loss with encoders that are constrained to have low Lipschitz constants.

Our analysis only found the optimal robust embedding for the case when the number of points N is not much larger than the dimension d. In practice, the number of points may be much larger and it would be interesting to extend the analysis to these cases as well. A possible extension can use the definition of minibatch ORE, where we want each minibatch to be as robust as possible. A similar approach for the CLIP loss was taken in Sreenivasan et al. (2023).

6 Discussion

In this work, we compared the CLIP loss with the recently proposed EMMUA loss (Yin et al., 2026b) through the lens of the ORE problem: maximizing embedding space margins in order to guarantee robustness. Building on prior work, we showed that all three problems share the same global minimum for $N \leq d + 1$: a perfectly aligned embedding space realized by the vertices of the regular simplex. Since in practice CLIP converges to a modality gap, which is both suboptimal and less robust, we studied EMMUA as a replacement loss that reaches this robust global minimum and overcomes CLIP's optimization failures.

Our work highlights the importance of studying not only a loss's global minima but also whether its training dynamics can reach them. We also show that such shortcomings cannot always be overcome by post-processing algorithms (Table 2), and may instead require retraining with a different loss.

Some open questions remain. In practice, the additional hyperparameter λ, alongside the temperature τ, requires further investigation to enable training at scale. Furthermore, while EMMUA's clean accuracy is comparable to CLIP's, it is slightly lower on some datasets, and whether this disparity persists on larger datasets remains to be seen. Finally, while we present extensive empirical evidence on the robustness and margins of both EMMUA and CLIP for large N, proving that the ORE and the global minima of these losses coincide in this setting remains open.

AI USE STATEMENT

In this work, we have not used generative AI tools for tasks with required disclosure. We used generative AI tools for writing the editing the code that produces the experiments, and generating the code for creating tables and figures. We have reviewed all AI-assisted work. Additionally, we've used generative AI tools to assist in table and figure formatting in latex, and find and correct typos and grammar mistakes. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All experiment details and full proofs are presented in the appendix.

REFERENCES

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Daniel C. Castro, Jeremy Tan, Bernhard Kainz, Ender Konukoglu, and Ben Glocker. Morpho-mnist: Quantitative assessment and diagnostics for representation learning. *Journal of Machine Learning Research*, 20(178):1–29, 2019. URL http://jmlr.org/papers/v20/19-033.html.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=rYhBGWYm6AU.

Rhea Chowers, Oshri Naparstek, Udi Barzelay, and Yair Weiss. Is the modality gap a bug or a feature? a robustness perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30288–30298, June 2026.

Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Investigating approaches for improving cross-modal alignment in clip, 2024. URL https://arxiv.org/abs/2406.17639.

Abrar Fahim, Alex Murphy, and Alona Fyshe. It's not a modality gap: Characterizing and addressing the contrastive gap, 2024. URL https://arxiv.org/abs/2405.18570.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10929–10974. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/garrido23a.html.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=S7Evzt9uit3.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Weiyang Liu, Rongmei Lin, Zhen Liu, Li Xiong, Bernhard Schölkopf, and Adrian Weller. Learning with hyperspherical uniformity. In Arindam Banerjee and Kenji Fukumizu (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1180–1188. PMLR, 2021. URL http://proceedings.mlr.press/v130/liu21d.html.

Jianfeng Lu and Stefan Steinerberger. Neural collapse under cross-entropy loss. *Applied and Computational Harmonic Analysis*, 59:224–241, 2022. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2021.12.011. URL https://www.sciencedirect.com/science/article/pii/S1063520321001123.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Fabian Mager, Hiba Nassar, and Lars Kai Hansen. On the modality gap and the contrastive loss in multi-modal representation learning. *CoRR*, abs/2607.10698, 2026. doi: 10.48550/ARXIV.2607.10698. URL https://doi.org/10.48550/arXiv.2607.10698.

Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=P4bXCawRi5J.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16784–16804. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/nichol22a.html.

Changdae Oh, Junhyuk So, Hoyoon Byun, YongTaek Lim, Minchul Shin, Jong-June Jeon, and Kyungwoo Song. Geodesic multi-modal mixup for robust fine-tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=iAAXq60Bw1.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020. doi: 10.1073/pnas.2015509117. URL https://www.pnas.org/doi/abs/10.1073/pnas.2015509117.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust CLIP: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43685–43704. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/schlarmann24a.html.

Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=uAFHCZRmXk.

Peiyang Shi, Michael C. Welle, Mårten Björkman, and Danica Kragic. Towards understanding the modality gap in CLIP. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*, 2023. URL https://openreview.net/forum?id=8W3KGzw7fNI.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/6b180037abbebea991d8b1232f8a8ca9-Paper.pdf.

Kartik Sreenivasan, Keon Lee, Jeong-Gwan Lee, Anna Lee, Jaewoong Cho, Jy yong Sohn, Dimitris Papailiopoulos, and Kangwook Lee. Mini-batch optimization of contrastive loss. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023. URL https://openreview.net/forum?id=7n2dNATSHeQ.

J.J. Thomson. Xxiv. on the structure of the atom: an investigation of the stability and periods of oscillation of a number of corpuscles arranged at equal intervals around the circumference of a circle; with application of the results to the theory of atomic structure. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 7(39):237–265, 1904. doi: 10.1080/14786440409463107. URL https://doi.org/10.1080/14786440409463107.

Weijie Tu, Weijian Deng, and Tom Gedeon. A closer look at the robustness of contrastive language-image pre-training (CLIP). In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=wMNpMe0vp3.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

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

Mingqi Wu, Qiang Sun, and Archer Y. Yang. PCA++: How uniformity induces robustness to background noise in contrastive learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=at87L8EuzR.

Shuhei Yamashita, Daiki Shirafuji, and Tatsuhiko Saito. Bridging the modality gap by similarity standardization with pseudo-positive samples. In Chu-Ren Huang, Yasunari Harada, Jong-Bok Kim, Nguyen T.M. Huyen, Le Thanh Huong, Pham Hien, Emmanuele Chersoni, Le Minh Nguyen, Rachel Edita Oñate Roxas, and Sherly Dita (eds.), *Proceedings of the 39th Pacific Asia Conference on Language, Information and Computation*, pp. 126–139, Hanoi, Vietnam, December 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.paclic-1.11/.

Jingwen Ye, Ruonan Yu, Songhua Liu, and Xinchao Wang. Mutual-modality adversarial attack with semantic perturbation. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pp. 6657–6665. AAAI Press, 2024. doi: 10.1609/AAAI.V38I7.28488. URL https://doi.org/10.1609/aaai.v38i7.28488.

Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 668–684, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19809-0.

Lingjie Yi, Raphael Douady, and Chao Chen. The convergent representation of contrastive vision-language models: Geometry, modality gap and shared space alignment. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=206P2ygDYN.

Wenzhe Yin, Zehao Xiao, Pan Zhou, Shujian Yu, Jiayi Shen, Jan-jakob Sonke, and Efstratios Gavves. Distributional vision-language alignment by cauchy-schwarz divergence. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 39688–39710, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/42bc612558891859b1b8717051f2c7b0-Paper-Conference.pdf.

Wenzhe Yin, Pan Zhou, Zehao Xiao, Jie Liu, Shujian Yu, Jan-Jakob Sonke, and Stratis Gavves. Towards uniformity and alignment for multimodal representation learning. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=im2Dw94HjX.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 11941–11952. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01100. URL https://doi.org/10.1109/ICCV51070.2023.01100.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text. In Zachary Lipton, Rajesh Ranganath, Mark Sendak, Michael Sjoding, and Serena Yeung (eds.), *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pp. 2–25. PMLR, 05–06 Aug 2022. URL https://proceedings.mlr.press/v182/zhang22a.html.

Yuhui Zhang, Jeff Z. HaoChen, Shih-Cheng Huang, Kuan-Chieh Wang, James Zou, and Serena Yeung. Diagnosing and rectifying vision models using language. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=D-zfUK7BR6c.

Yuhui Zhang, Elaine Sui, and Serena Yeung. Connect, collapse, corrupt: Learning cross-modal tasks with uni-modal data. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ttXg3SKAg5.

Zihan Zhou, Yang Zhou, Ruoming Jin, Pan He, and Patrick Emami. Mitigating the modality gap in vision–language models with fractal spectral geometry. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=pGkM5BjfD1.

A Theorem Proofs

A.1 Optimal Solution to CLIP's "Uniformity" is not Uniform

The second term of the CLIP loss has previously been referred to as "uniformity", (Fahim et al., 2024; Shi et al., 2023; Oh et al., 2023; Yin et al., 2026a). However, extending the proof of Wang & Isola (2020) to the multi-modal case has yet to be proven. In fact, the global minima of the "uniformity" term in the multi-modal case is not uniform at all, but achieved when the two modalities collapse to two antipodal points, urging the embeddings distributions to be $(\boldsymbol{x}_i, \boldsymbol{y}_i) \sim \delta(\boldsymbol{x} - \boldsymbol{\mu}) \times \delta(\boldsymbol{y} + \boldsymbol{\mu})$ for some $\boldsymbol{\mu} \in \mathbb{S}^d$.

Lemma A.1. *The global minima over the unit hypersphere of the multi-modal "uniformity" term:*

$$\mathcal{L}_{uniform} = \sum_i \log\left(\sum_{j=1}^{N} e^{\boldsymbol{x}_i^T \boldsymbol{y}_j}\right) + \log\left(\sum_{j=1}^{N} e^{\boldsymbol{y}_i^T \boldsymbol{x}_j}\right) \tag{9}$$

is when the modalities collapse to two antipodal points: $\forall i, j : \boldsymbol{x}_j = \boldsymbol{x}_i = -\boldsymbol{y}_j = -\boldsymbol{y}_i$.

Proof. Since $\forall i, j : \|\boldsymbol{x}_i\| = \|\boldsymbol{y}_j\| = 1$ then $\|x_i - y_j\| \in [0, 2]$, with maximum when $\boldsymbol{x}_i = -\boldsymbol{y}_j$ therefore:

$$e^{-\|\boldsymbol{x}_i - \boldsymbol{y}_j\|^2} \in [e^{-4}, 1] \tag{10}$$

obtaining the global minima when $\forall i, j : \boldsymbol{x}_i = \boldsymbol{x}_j = -\boldsymbol{y}_i = -\boldsymbol{y}_j$.
For uniqueness, as explained above, to minimize a specific pair $e^{-\|\boldsymbol{x}_i - \boldsymbol{y}_j\|^2}$ we must have $\boldsymbol{x}_i = -\boldsymbol{y}_j$. Since the two sums include all possible pairs, then $\forall i, j : \boldsymbol{x}_i = -\boldsymbol{y}_j$ which implies $\boldsymbol{x}_i = -\boldsymbol{y}_j = -\boldsymbol{x}_j = -(-\boldsymbol{y}_i)$, as required. □

A.2 Proof of Suboptimality of the Modality Gap

We begin with a previously reported assumption of an orthogonal gap, and prove that under it a gapped solution is suboptimal.

Definition A.2. Orthogonal Gap (Zhang et al., 2023; Chowers et al., 2026): A solution $\tilde{\mathcal{X}}, \tilde{\mathcal{Y}}$ with an orthogonal and non-zero gap, for which the local gap vectors are equal to the global gap vector, i.e.:

$$\forall i \neq j : \boldsymbol{g} = \boldsymbol{x}_i - \boldsymbol{y}_i = \boldsymbol{x}_j - \boldsymbol{y}_j \wedge \langle \boldsymbol{g}, \boldsymbol{x}_i - \boldsymbol{\mu}_x \rangle = \langle \boldsymbol{g}, \boldsymbol{y}_i - \boldsymbol{\mu}_y \rangle = 0 \tag{11}$$

with $\boldsymbol{g} = \boldsymbol{\mu}_x - \boldsymbol{\mu}_y = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i - \boldsymbol{y}_i$ and $\|\boldsymbol{g}\| > 0$.

Lemma A.3. *Let $\boldsymbol{v} \in \mathbb{R}^d$ s.t. $\boldsymbol{v} \perp \mathcal{X}$ and $\boldsymbol{v} \perp \mathcal{Y}$. For any such $\boldsymbol{v}$, shifting either $\mathcal{X}$ or $\mathcal{Y}$ by $\alpha \cdot \boldsymbol{v}$ (for any $\alpha \in \mathbb{R}$) does not change the contrastive loss, i.e.:*

$$\forall \alpha \in \mathbb{R} : \mathcal{L}(\mathcal{X}, \mathcal{Y} + \alpha \cdot \boldsymbol{v}) = \mathcal{L}(\mathcal{X}, \mathcal{Y}) \tag{12}$$

Proof. Assume WLOG that $\boldsymbol{v}$ has unit norm. By the assumption that $\boldsymbol{v} \perp \mathcal{X}$:

$$\forall \boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X} : \langle \boldsymbol{v}, \boldsymbol{x}_i \rangle = \langle \boldsymbol{v}, \boldsymbol{x}_j \rangle = c_x \tag{13}$$

And similarly for $\mathcal{Y}$ it pertains that $\forall \boldsymbol{y}_j \in \mathcal{Y} : \langle \boldsymbol{v}, \boldsymbol{y}_j \rangle = c_y$.
Substituting into the contrastive loss:

$$\begin{aligned}
\mathcal{L}(\mathcal{X}, \mathcal{Y} + \alpha \cdot \boldsymbol{v}) &= -\sum_{i=1}^{N} \log\left(\frac{e^{\boldsymbol{x}_i^T(\boldsymbol{y}_i + \alpha \cdot \boldsymbol{v})}}{\sum_j e^{\boldsymbol{x}_i^T(\boldsymbol{y}_j + \alpha \cdot \boldsymbol{v})}}\right) + \log\left(\frac{e^{\boldsymbol{x}_i^T(\boldsymbol{y}_i + \alpha \cdot \boldsymbol{v})}}{\sum_j e^{(\boldsymbol{y}_i + \alpha \cdot \boldsymbol{v})^T \boldsymbol{x}_j}}\right) \\
&= -\sum_{i=1}^{N} \log\left(\frac{e^{\boldsymbol{x}_i^T \boldsymbol{y}_i + \alpha c_x}}{\sum_j e^{\boldsymbol{x}_i^T \boldsymbol{y}_j + \alpha c_x}}\right) + \log\left(\frac{e^{\boldsymbol{x}_i^T \boldsymbol{y}_i + \alpha c_x}}{\sum_j e^{\boldsymbol{y}_i^T \boldsymbol{x}_j + \alpha c_x}}\right) \\
&= -\sum_{i=1}^{N} \log\left(\frac{e^{\boldsymbol{x}_i^T \boldsymbol{y}_i}}{\sum_j e^{\boldsymbol{x}_i^T \boldsymbol{y}_j}}\right) + \log\left(\frac{e^{\boldsymbol{x}_i^T \boldsymbol{y}_i}}{\sum_j e^{\boldsymbol{y}_i^T \boldsymbol{x}_j}}\right) \\
&= \mathcal{L}(\mathcal{X}, \mathcal{Y})
\end{aligned} \tag{14}$$

Notice that the same holds for the loss when calculated using $-\|\boldsymbol{x}_i - \boldsymbol{y}_i\|^2$ instead of cosine similarity since:

$$-\|\boldsymbol{x}_i - (\boldsymbol{y}_j + \alpha \cdot \boldsymbol{v})\|^2 = -\|\boldsymbol{x}_i - \boldsymbol{y}_j\|^2 - \underbrace{\alpha^2 \cdot \|\boldsymbol{v}\|^2 - 2\alpha c_y + 2\alpha c_x}_{\text{independent of } \boldsymbol{x}_i, \boldsymbol{y}_j} \tag{15}$$

□

Lemma A.4. *For any multi-modal embeddings $\mathcal{X}, \mathcal{Y}$ that have perfect alignment (i.e. $\forall i \in [N]$: $\boldsymbol{x}_i = \boldsymbol{y}_i$), the multi-modal contrastive loss $\mathcal{L}(\mathcal{X}, \mathcal{Y})$ is equal to the info-NCE loss of any of the two modality with the positive pairs being the embeddings themselves i.e. $\mathcal{L}_{NCE}(\mathcal{X}, \mathcal{X})$.*

Proof. The multi-modal contrastive loss is:

$$\mathcal{L}(\mathcal{X}, \mathcal{Y}) = -\frac{1}{N} \cdot \frac{1}{2} \sum_i \log \left(\frac{e^{-\|\boldsymbol{x}_i - \boldsymbol{y}_i\|^2}}{\sum_j e^{-\|\boldsymbol{x}_i - \boldsymbol{y}_j\|^2}} \right) + \log \left(\frac{e^{-\|\boldsymbol{x}_i - \boldsymbol{y}_i\|^2}}{\sum_j e^{-\|\boldsymbol{y}_i - \boldsymbol{x}_j\|^2}} \right) \tag{16}$$

and under the assumption of alignment becomes:

$$\begin{aligned} \mathcal{L}(\mathcal{X}, \mathcal{Y}) &= -\frac{1}{N} \cdot \frac{1}{2} \sum_i \log \left(\frac{1}{\sum_j e^{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}} \right) + \log \left(\frac{1}{\sum_j e^{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}} \right) \\ &= -\frac{1}{N} \sum_i \log \left(\frac{1}{\sum_j e^{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}} \right) \end{aligned} \tag{17}$$

On the other hand:

$$\begin{aligned} \mathcal{L}_{NCE}(\mathcal{X}, \mathcal{X}) &= -\frac{1}{N} \sum_i \log \left(\frac{e^{-\|\boldsymbol{x}_i - \boldsymbol{x}_i\|^2}}{e^{-\|\boldsymbol{x}_i - \boldsymbol{x}_i\|^2} + \sum_{j \neq i} e^{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}} \right) \\ &= -\frac{1}{N} \sum_i \log \left(\frac{1}{\sum_j e^{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}} \right) \\ &= \mathcal{L}(\mathcal{X}, \mathcal{Y}) \end{aligned} \tag{18}$$

□

Theorem A.5. *Let $\tilde{\mathcal{X}}, \tilde{\mathcal{Y}}$ as in assumption A.2, and let $\mathcal{X}^*, \mathcal{Y}^*$ be uniformly aligned solutions. Then:*

$$\mathcal{L}(\mathcal{X}^*, \mathcal{Y}^*) < \mathcal{L}(\tilde{\mathcal{X}}, \tilde{\mathcal{Y}}) \tag{19}$$

i.e. $\tilde{\mathcal{X}}, \tilde{\mathcal{Y}}$ is not an optimal solution to the multimodal contrastive loss.

Proof. Define the intermediate solution $\tilde{\mathcal{X}}', \tilde{\mathcal{Y}}'$ s.t. $\tilde{\mathcal{X}}' = \tilde{\mathcal{X}}, \tilde{\mathcal{Y}}' = \tilde{\mathcal{Y}} + \boldsymbol{g}$. According to Theorem A.3:

$$\mathcal{L}(\tilde{\mathcal{X}}, \tilde{\mathcal{Y}}) = \mathcal{L}(\tilde{\mathcal{X}}', \tilde{\mathcal{Y}}') \tag{20}$$

According to Theorem A.4, since under assumption A.2 $\tilde{\mathcal{X}}', \tilde{\mathcal{Y}}'$ is perfectly aligned:

$$\mathcal{L}(\tilde{\mathcal{X}}', \tilde{\mathcal{Y}}') = \mathcal{L}_{NCE}(\tilde{\mathcal{X}}', \tilde{\mathcal{X}}') \tag{21}$$

By Wang & Isola (2020), the minima of $\mathcal{L}_{NCE}(\tilde{\mathcal{X}}', \tilde{\mathcal{X}}')$ is when $\tilde{\mathcal{X}}'$ is uniformly distributed on the hypersphere. Since by assumption A.2 $\boldsymbol{g} \perp \tilde{\mathcal{X}}', \mathcal{X}$ then the variance of $\tilde{\mathcal{X}}'$ in $\boldsymbol{g}$ is zero, meaning it isn't uniform on the hypersphere therefore hasn't minimal Info-NCE loss. Define $\mathcal{X}^* = \mathcal{Y}*$ which is uniformly distributed, then:

$$\mathcal{L}_{NCE}(\mathcal{X}^*, \mathcal{X}^*) < \mathcal{L}_{NCE}(\tilde{\mathcal{X}}', \tilde{\mathcal{X}}') \tag{22}$$

By construction and according to Theorem A.4:

$$\mathcal{L}_{NCE}(\mathcal{X}^*, \mathcal{X}^*) = \mathcal{L}(\mathcal{X}^*, \mathcal{Y}^*) \tag{23}$$

Giving in total:

$$\mathcal{L}(\mathcal{X}^*, \mathcal{Y}^*) < \mathcal{L}(\tilde{\mathcal{X}}', \tilde{\mathcal{Y}}') = \mathcal{L}(\tilde{\mathcal{X}}, \tilde{\mathcal{Y}}) \tag{24}$$

Meaning the solution with the gap isn't optimal in terms of multi-modal contrastive loss.

While the above is true in the distributional sense, i.e. when $|\mathcal{X}|, |\mathcal{Y}| \to \infty$, Theorem A.6 proves that $\tilde{\mathcal{X}}'$ is not an optimal solution to the NCE loss even when $|\mathcal{X}|, |\mathcal{Y}| = N$ is finite. □

Lemma A.6. *Let $\mathcal{X}$ be a set of $N \geq 2$ points in $\mathbb{S}^{d-1}$ s.t. $\exists \boldsymbol{v} \quad \forall \boldsymbol{x}_i \in \mathcal{X} : \langle \boldsymbol{v}, \boldsymbol{x}_i \rangle = c$ with $\|v\| = 1$ and $c > 0$. Then exists a solution $\mathcal{X}^*$ of N points in $\mathbb{S}^{d-1}$ that has lower Info-NCE loss than $\mathcal{X}$.*

Proof. We prove by construction. $\forall \boldsymbol{x}_i$, decompose $\boldsymbol{x}_i = a_i \cdot \boldsymbol{u}_i + c \cdot \boldsymbol{v}$ for some $\boldsymbol{u}_i \perp \boldsymbol{v}$. Now define a counterpart $\boldsymbol{x}_i^*$ in the following way:

$$\boldsymbol{x}_i^* = \sqrt{1 - (c - \epsilon)^2}\boldsymbol{u}_i + (c - \epsilon)\boldsymbol{v} \tag{25}$$

For some small $\epsilon < \frac{c}{2}$. By construction, $\|\boldsymbol{x}_i^*\| = 1$. Assuming perfect alignment, the loss consists only of the uniformity term which is controlled by the dot products $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$ for $i \neq j$. Observe:

$$\begin{aligned} \left\langle \boldsymbol{x}_i^*, \boldsymbol{x}_j^* \right\rangle &= (1 - (c - \epsilon)^2) \langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle + (c - \epsilon)^2 \\ &= \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + (2\epsilon c - \epsilon^2)(\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle - 1) \end{aligned} \tag{26}$$

Since $\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle \in [-1, 1)$ (since $\boldsymbol{u}_i \neq \boldsymbol{u}_j$) then $\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle - 1 < 0$. Since $\epsilon < \frac{c}{2} \Rightarrow (2\epsilon c - \epsilon^2) > 0$, that means that in total:

$$(2\epsilon c - \epsilon^2)(\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle - 1) < 0 \Rightarrow \left\langle \boldsymbol{x}_i^*, \boldsymbol{x}_j^* \right\rangle < \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \tag{27}$$

Since the uniformity loss is strictly monotone in the inner products, decreasing them means decreasing the loss, in total:

$$\mathcal{L}_{NCE}(\mathcal{X}^*) < \mathcal{L}_{NCE}(\mathcal{X}) \tag{28}$$

as required. □

A.3 Max Margin and Robustness

A.3.1 Proof of Theorem 3.4

Proof. Given a dataset of size N we show how to construct a dataset of size $N - 1$ whose robustness is at least that of the original dataset. We do this by removing the point i^* for which margin_i is minimal. Clearly:

$$\begin{aligned} \frac{1}{N-1} \sum_{i \neq i^*} \text{margin}(\boldsymbol{x}_i, \{\boldsymbol{y}_j\}) &\geq \frac{1}{N} \sum_i \text{margin}(\boldsymbol{x}_i, \{\boldsymbol{y}_j\}) && (29) \\ &= R(\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^N) && (30) \end{aligned}$$

Similarly for each fixed $i \neq i^*$:

$$\begin{aligned} \text{margin}(\boldsymbol{x}_i, \{\boldsymbol{y}_j\}) &= \langle \boldsymbol{x}_i, \boldsymbol{y}_i \rangle - \max_{j \neq i} \langle \boldsymbol{x}_i, \boldsymbol{y}_j \rangle && (31) \\ &\leq \langle \boldsymbol{x}_i, \boldsymbol{y}_i \rangle - \max_{j \neq i, j \neq i^*} \langle \boldsymbol{x}_i, \boldsymbol{y}_j \rangle && (32) \end{aligned}$$

So that:

$$\begin{aligned} R(\{\boldsymbol{x}_i, \boldsymbol{y}_i\})_{i \neq i^*} &= \frac{1}{N-1} \sum_{i \neq i^*} \langle \boldsymbol{x}_i, \boldsymbol{y}_i \rangle - \max_{j \neq i^*} \langle \boldsymbol{x}_i, \boldsymbol{y}_j \rangle && (33) \\ &\geq \frac{1}{N-1} \sum_{i \neq i^*} \text{margin}(\boldsymbol{x}_i, \{\boldsymbol{y}_j\}) && (34) \\ &\geq R(\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^N) && (35) \end{aligned}$$

□

A.3.2 Proof of Theorem 3.3

Lemma A.7. *Assume some $N \in \mathbb{N}$ and sets of vectors $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N \subset \mathbb{S}^d$. Then:*

$$\sum_{i,j \neq i} \boldsymbol{c}_{ii} - \boldsymbol{c}_{ij} \leq N^2 \tag{36}$$

with $\boldsymbol{c}_{ij} = \langle \boldsymbol{x}_i, \boldsymbol{y}_j \rangle$.

Proof.

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{j\neq i}(\boldsymbol{c}_{ij}-\boldsymbol{c}_{ii}) \\
&=\left(\sum_{i=1}^{N}\boldsymbol{x}_i\right)\cdot\left(\sum_{j=1}^{N}\boldsymbol{y}_j\right)-\sum_{i=1}^{N}\boldsymbol{x}_i\cdot\boldsymbol{y}_i-(N-1)\sum_{i=1}^{N}\boldsymbol{x}_i\cdot\boldsymbol{y}_i \\
&=-\left(N\sum_{i=1}^{N}\boldsymbol{x}_i\cdot\boldsymbol{y}_i-\left(\sum_{i=1}^{N}\boldsymbol{x}_i\right)\cdot\left(\sum_{j=1}^{N}\boldsymbol{y}_j\right)\right)=-\sum_{i<j}(\boldsymbol{x}_i-\boldsymbol{x}_j)\cdot(\boldsymbol{y}_i-\boldsymbol{y}_j) \\
&\underbrace{\geq}_{\text{Cauchy Schwarz}}-\left(\sum_{i<j}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2\right)^{1/2}\cdot\left(\sum_{i<j}\|\boldsymbol{y}_i-\boldsymbol{y}_j\|^2\right)^{1/2}
\end{aligned}\tag{37}$$

Since $\|\boldsymbol{x}_i\|=\|\boldsymbol{y}_j\|=1$ we get:

$$\|\boldsymbol{y}_i-\boldsymbol{y}_j\|^2=2-2\boldsymbol{y}_i\cdot\boldsymbol{y}_j \tag{38}$$

Therefore:

$$\begin{aligned}
\sum_{i<j}\|\boldsymbol{y}_i-\boldsymbol{y}_j\|^2&=\sum_{i<j}2-2\cdot\sum_{i<j}\boldsymbol{y}_i\cdot\boldsymbol{y}_j=N(N-1)-2\cdot\sum_{i<j}\boldsymbol{y}_i\cdot\boldsymbol{y}_j \\
&=N(N-1)-\left(||\sum_i\boldsymbol{y}_i||^2-N\right)\leq N^2
\end{aligned}\tag{39}$$

With exact equality if and only if the $\boldsymbol{y}_i$'s are centered around the origin (and the same holds for $\boldsymbol{x}_i$). Substituting back:

$$-\left(\sum_{i<j}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2\right)^{1/2}\cdot\left(\sum_{i<j}\|\boldsymbol{y}_i-\boldsymbol{y}_j\|^2\right)^{1/2}\geq-(N^2)^{1/2}\cdot(N^2)^{1/2}=-N^2 \tag{40}$$

In total:

$$\sum_{i=1}^{N}\sum_{j\neq i}(\boldsymbol{c}_{ij}-\boldsymbol{c}_{ii})\geq-N^2\Leftrightarrow\sum_{i=1}^{N}\sum_{j\neq i}(\boldsymbol{c}_{ii}-\boldsymbol{c}_{ij})\leq N^2 \tag{41}$$

As required. □

Theorem A.8. *Assume $N \leq d+1$, then the configuration $\{\boldsymbol{x}_i\},\{\boldsymbol{y}_j\}$ that maximizes $\sum_i \text{margin}(\boldsymbol{y}_i,\{\boldsymbol{x}_j\})$ is when $\boldsymbol{x}_i=\boldsymbol{y}_i$ and they are vertices of the regular simplex.*

Proof. We first show a bound on the sum of margins.

$$\begin{aligned}
\sum_i\text{margin}(y_i,\{x_j\})=&\quad\sum_i\langle\boldsymbol{y}_i,\boldsymbol{x}_i\rangle-\max_{j\neq i}\langle\boldsymbol{y}_i,\boldsymbol{x}_j\rangle \\
\leq&\quad\sum_i\langle\boldsymbol{y}_i,\boldsymbol{x}_i\rangle-\frac{1}{N-1}\sum_{j\neq i}\langle\boldsymbol{y}_i,\boldsymbol{x}_j\rangle \\
\leq&\quad\frac{N^2}{N-1}
\end{aligned}\tag{42}$$

Where the last step is based on lemma A.7.

and:

$$\begin{aligned}\sum_{i,j\neq i} c_{ii} - c_{ij} &= \sum_i \sum_{j\neq i} c_{ii} - \sum_i \sum_{j\neq i} c_{ij} \\ &= \sum_i (N-1)c_{ii} - \sum_i \sum_{j\neq i} c_{ij} \\ &= (N-1)\left(\sum_i c_{ii} - \frac{1}{N-1}\sum_{j\neq i} c_{ij}\right) \end{aligned} \tag{43}$$

Now for the regular simplex $c_{ii} = 1, c_{ij} = -\frac{1}{N-1}$ so that the sum of the margins is equal to $N(1 + \frac{1}{N-1}) = \frac{N^2}{N-1}$ so that it achieves the bound. □

A.3.3 Proof of Theorem 3.2

Proof. By linearity of the dot product:

$$\langle \boldsymbol{x}_1 + \boldsymbol{v}, \boldsymbol{y}_i \rangle = \langle \boldsymbol{v}, \boldsymbol{y}_i \rangle + \langle \boldsymbol{x}_1, \boldsymbol{y}_i \rangle \tag{44}$$

and since $\boldsymbol{y}_i$ is on the unit sphere and $\|\boldsymbol{v}\| < \frac{\delta}{2}$

$$|\langle \boldsymbol{y}_i, \boldsymbol{v} \rangle| \leq \|\boldsymbol{v}\| \|\boldsymbol{y}_i\| < \frac{\delta}{2} \tag{45}$$

Which means that

$$|\langle \boldsymbol{x}_1 + \boldsymbol{v}, \boldsymbol{y}_1 \rangle - \langle \boldsymbol{x}_1, \boldsymbol{y}_1 \rangle| < \frac{\delta}{2} \tag{46}$$

Applying the same logic to $\boldsymbol{y}_i$ shows that:

$$|\langle \boldsymbol{x}_1 + \boldsymbol{v}, \boldsymbol{y}_1 \rangle - \langle \boldsymbol{x}_1, \boldsymbol{y}_i \rangle| < \frac{\delta}{2} \tag{47}$$

and since $\langle \boldsymbol{x}_1, \boldsymbol{y}_1 \rangle - \langle \boldsymbol{x}_1, \boldsymbol{y}_i \rangle \geq \delta$, equations 46,47 imply that $\langle \boldsymbol{x}_1 + \boldsymbol{v}, \boldsymbol{y}_1 \rangle > \langle \boldsymbol{x}_1 + \boldsymbol{v}, \boldsymbol{y}_i \rangle$ for all $i \neq 1$. □

A.3.4 Extension to Input Noise

We prove that for a large family of functions this also induces input robustness, such as to adversarial examples (Madry et al., 2018), without any explicit adversarial training (Ye et al., 2024):

Theorem A.9. *Let $f_\theta : \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{out}}$ be some L-Lipschitz continuous embedding function s.t. $\forall \boldsymbol{x} : \|f_\theta(\boldsymbol{x})\| \leq m$. Then any $\boldsymbol{x}_i$ which maintains $\operatorname{margin}(f_\theta(\boldsymbol{x}_i), \{f_\theta(\boldsymbol{y}_j)\}) \geq \frac{2L\|\delta\|}{m}$ is robust to any input perturbation of $\|\delta\|$.*

Proof. Assume an embedding function $f_\theta : \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{out}}$ which is L-Lipschitz. By definition:

$$|f_\theta(\boldsymbol{x}) - f_\theta(\boldsymbol{x} + \delta)| \leq L \cdot |\boldsymbol{x} - (\boldsymbol{x} + \delta)| = L\,|\delta| \tag{48}$$

The change in cosine similarity is therefore:

$$\begin{aligned}
\Delta &= |\text{cossim}(f_\theta(\boldsymbol{y}), f_\theta(\boldsymbol{x})) - \text{cossim}(f_\theta(\boldsymbol{y}), f_\theta(\boldsymbol{x}+\delta))| \\
&= \left\| \frac{f_\theta(\boldsymbol{y})}{\|f_\theta(\boldsymbol{y})\|} \cdot \frac{f_\theta(\boldsymbol{x}))}{\|f_\theta(\boldsymbol{x}))\|} - \frac{f_\theta(\boldsymbol{y})}{\|f_\theta(\boldsymbol{y})\|} \cdot \frac{f_\theta(\boldsymbol{x}+\delta))}{\|f_\theta(\boldsymbol{x}+\delta))\|} \right\| \\
&= \left\| \left(\frac{f_\theta(\boldsymbol{x}))}{\|f_\theta(\boldsymbol{x}))\|} - \frac{f_\theta(\boldsymbol{x}+\delta))}{\|f_\theta(\boldsymbol{x}+\delta))\|} \right) \cdot \frac{f_\theta(\boldsymbol{y})}{\|f_\theta(\boldsymbol{y})\|} \right\| \\
&\leq \left\| \frac{f_\theta(\boldsymbol{x}))}{\|f_\theta(\boldsymbol{x}))\|} - \frac{f_\theta(\boldsymbol{x}+\delta))}{\|f_\theta(\boldsymbol{x}+\delta))\|} \right\| \cdot \left\| \frac{f_\theta(\boldsymbol{y})}{\|f_\theta(\boldsymbol{y})\|} \right\| \\
&= \left\| \frac{f_\theta(\boldsymbol{x}))}{\|f_\theta(\boldsymbol{x}))\|} - \frac{f_\theta(\boldsymbol{x}+\delta))}{\|f_\theta(\boldsymbol{x}+\delta))\|} \right\| \\
&\leq \frac{\|f_\theta(\boldsymbol{x}) - f_\theta(\boldsymbol{x}+\delta)\|}{\min} \left(\|f_\theta(\boldsymbol{x})\|, \|f_\theta(\boldsymbol{x}+\delta)\| \right) \\
&\leq \frac{L\,\|\delta\|}{\min} \left(\|f_\theta(\boldsymbol{x})\|, \|f_\theta(\boldsymbol{x}+\delta)\| \right) \qquad (49)
\end{aligned}$$

By Cauchy-Schwarz and Lipschitzness. Assuming the embedding norms are bounded by some $m > 0$, the change in cosine similarity is at most:

$$|\text{cossim}(f_\theta(\boldsymbol{y}), f_\theta(\boldsymbol{x})) - \text{cossim}(f_\theta(\boldsymbol{y}), f_\theta(\boldsymbol{x}+\delta))| \leq \frac{L\,\|\delta\|}{m} \qquad (50)$$

□

A.4 Analysis of EMMUA

We follow three conditions which were previously reported to contribute to the formation of the modality gap and its maintaining throughout training. We define each of them and prove that the gap is still closed under each of them for EMMUA.

For our analysis, we use a slightly different definition of EMMUA:

$$\begin{aligned}
\mathcal{L}_{EMMUA} &= -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{-\lambda\|\boldsymbol{x}_i - \boldsymbol{y}_i\|^2}}{e^{-\lambda\|\boldsymbol{x}_i - \boldsymbol{y}_i\|^2} + \sum_{j\neq i} e^{\frac{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\tau}}} \\
&\quad + \log \frac{e^{-\lambda\|\boldsymbol{x}_i - \boldsymbol{y}_i\|^2}}{e^{-\lambda\|\boldsymbol{x}_i - \boldsymbol{y}_i\|^2} + \sum_{j\neq i} e^{\frac{-\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2}{\tau}}} \\
&= \frac{1}{N}\sum_{i=1}^{N} \lambda\,\|\boldsymbol{x}_i - \boldsymbol{y}_i\|^2 + \frac{1}{2}\log\left(\sum_{j\neq i} e^{\frac{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\tau}}\right) + \frac{1}{2}\log\left(\sum_{j\neq i} e^{\frac{-\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2}{\tau}}\right) \qquad (51)
\end{aligned}$$

since this version of the loss has a well defined logit matrix (equation 53) and softmax of the logit matrix.

In practice, we drop the term $e^{-\lambda\|\boldsymbol{x}_i - \boldsymbol{y}_i\|^2}$ from the denominators term, similar to the decoupled contrastive loss (Yeh et al., 2022). For simplicity, we assume $\lambda = 1/\tau$, where in practice we use $\lambda = \frac{1}{1.2\cdot\tau}$.

A.4.1 Dynamics under Narrow Cones

Definition A.10. Narrow Cones (Liang et al., 2022; Zhang et al., 2024; Chowers et al., 2026): the setting where at initialization, the two modalities are embedded in two clusters which are relatively more concentrated around their means than the distance between them, i.e. $\forall i : \|\boldsymbol{y}_i - \boldsymbol{\mu}_y\| \leq \epsilon$, $\|\boldsymbol{x}_i - \boldsymbol{\mu}_x\| \leq \epsilon$ and $\|\boldsymbol{\mu}_x - \boldsymbol{\mu}_y\| \gg \epsilon$.

Theorem A.11. *Assume that assumption A.10 holds. Then the gradient of EMMUA w.r.t. the embeddings translates each modality towards the other, shrinking the gap.*

Proof. The gradient of EMMUA w.r.t. the embedding is:

$$\nabla_{\boldsymbol{y}_i}\mathcal{L}_{EMMUA} = \frac{1}{N\tau}\left(2\left(\boldsymbol{x}_i - \boldsymbol{y}_i\right) - \sum_{j\neq i}\left(\boldsymbol{Q}^y(i,j) + \boldsymbol{Q}^y(j,i)\right)\left(\boldsymbol{y}_j - \boldsymbol{y}_i\right)\right) \tag{52}$$

Where $\boldsymbol{Q}$ are the two logit matrices of EMMUAs:

$$\begin{aligned}\boldsymbol{Q}^y_{ij} &= \frac{e^{\boldsymbol{y}_i^T\boldsymbol{x}_j/\tau}}{e^{\boldsymbol{y}_i^T\boldsymbol{x}_j/\tau} + \sum_{k\neq j} e^{\boldsymbol{y}_i^T\boldsymbol{y}_k/\tau}} \\ \boldsymbol{Q}^x_{ij} &= \frac{e^{\boldsymbol{x}_i^T\boldsymbol{y}_j/\tau}}{e^{\boldsymbol{x}_i^T\boldsymbol{y}_j/\tau} + \sum_{k\neq j} e^{\boldsymbol{x}_i^T\boldsymbol{x}_k/\tau}}\end{aligned} \tag{53}$$

Under our assumption, $\|\boldsymbol{y}_i - \boldsymbol{y}_j\| = O(\epsilon)$, therefore observe the taylor expansion of the gaussian kernel around zero:

$$k(\boldsymbol{y}_i, \boldsymbol{y}_j) = e^{-\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2/\tau} = 1 - O(\epsilon^2/\tau) + O(\epsilon^4/\tau) \tag{54}$$

Substituting the linearized kernel into the definition of $\boldsymbol{Q}^y$:

$$\begin{aligned}\boldsymbol{Q}^y_{ij} = \frac{k(\boldsymbol{y}_i, \boldsymbol{x}_j)}{\sum_{r\neq j} k(\boldsymbol{y}_i, \boldsymbol{y}_r)} &\approx \frac{1 - \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2/\tau}{\sum_{r\neq j} 1 - \|\boldsymbol{y}_i - \boldsymbol{y}_r\|^2/\tau} \\ &= \frac{1 - O(\epsilon^2/\tau)}{(N-1)\left(1 - O(\epsilon^2/\tau)\right)} = \frac{1}{N-1} + O(\epsilon^2/\tau)\end{aligned} \tag{55}$$

Substituting into the gradient:

$$\begin{aligned}\nabla_{\boldsymbol{y}_i}\mathcal{L}_{EMMUA} &\approx \frac{1}{N\tau}\left(2(-\boldsymbol{g}) + O(\epsilon)\boldsymbol{r}_{ii} - N\cdot 2\cdot\left(\frac{1}{N-1} + O(\epsilon^2/\tau)\right)\cdot O(\epsilon)\boldsymbol{r}_{ji}\right) \\ &\approx \frac{1}{N\tau}\left(2(-\boldsymbol{g}) + O(\epsilon)\boldsymbol{r}_{ii} + 2\left(\frac{N}{N-1}O(\epsilon/\tau)\boldsymbol{r}_{ji} + O(N\epsilon^3/\tau)\right)\right)\end{aligned} \tag{56}$$

Where we define $\boldsymbol{r}_{ii} = \boldsymbol{x}_i - \boldsymbol{y}_i$ and $\boldsymbol{r}_{ji} = \boldsymbol{y}_j - \boldsymbol{y}_i$. Overall we see that the direction of the gradient is approximately in the direction of the gap up to $O(\epsilon)$. □

A.4.2 Closing of an Orthogonal Gap

Theorem A.12. *Assume that assumption A.2 holds, then the gradient shrinks the gap.*

Proof. Assume some $\boldsymbol{v}$ s.t. $\forall i : \boldsymbol{v}^T\boldsymbol{x}_i = c_x$ and $\boldsymbol{v}^T\boldsymbol{y}_i = c_y$. Substituting into the gradient:

$$\begin{aligned}\boldsymbol{v}^T\nabla_{\boldsymbol{y}_i}\mathcal{L}_{EMMUA} &= \frac{1}{N\tau}\left(2\boldsymbol{v}^T\left(\boldsymbol{x}_i - \boldsymbol{y}_i\right) - \sum_{j\neq i}\left(\boldsymbol{Q}^y(i,j) + \boldsymbol{Q}^y(j,i)\right)\boldsymbol{v}^T\left(\boldsymbol{y}_j - \boldsymbol{y}_i\right)\right) \\ &= \frac{1}{N\tau}2\left(c_x - c_y\right)\end{aligned} \tag{57}$$

since $\boldsymbol{v}^T(\boldsymbol{y}_j - \boldsymbol{y}_i) = c_y - c_y = 0$. From symmetry of the loss the same applies to the other modality $\boldsymbol{v}^T\nabla_{\boldsymbol{x}_i}\mathcal{L}_{EMMUA} = \frac{1}{N\tau}2\left(c_y - c_x\right)$. Therefore throughout training, the difference between gradient steps $t, t+1$ becomes:

$$\begin{aligned}\boldsymbol{v}^T(\boldsymbol{x}_i(t+1) - \boldsymbol{y}_i(t+1)) &= \boldsymbol{v}^T\left(\boldsymbol{x}_i(t) + \eta\nabla_{\boldsymbol{x}_i}\mathcal{L}_{EMMUA} - \boldsymbol{y}_i(t) - \eta\nabla_{\boldsymbol{y}_i}\mathcal{L}_{EMMUA}\right) \\ &= c_x - c_y - \eta\cdot\boldsymbol{v}^T\left(\nabla_{\boldsymbol{y}_i}\mathcal{L}_{EMMUA} - \nabla_{\boldsymbol{x}_i}\left(\mathcal{L}_{EMMUA}\right)\right) \\ &= c_x - c_y - \eta\frac{2}{N\tau}\left(2c_x - 2c_y\right) \\ &= \left(1 - \frac{4\eta}{N\cdot\tau}\right)\left(c_x - c_y\right)\end{aligned} \tag{58}$$

Meaning that as long as the learning rate $\left|\frac{4\eta}{N\cdot\tau}\right| < 1$ (realistic since N is usually large and $\eta < \tau$), the gap in $\boldsymbol{v}$ shrinks throughout training. □

A.4.3 Dynamics under Double-Stochasticity

Definition A.13. Double Stochasticity (Chowers et al., 2026): The logit matrix (equation 53) of the loss is said to be doubly stochastic if all its rows and columns sum to 1.

Theorem A.14. *Assume that assumption A.13 holds, the gradient of EMMUA shrinks the gap.*

Proof. Observe the gradient of the mean of modality $\mathcal{Y}$:

$$\frac{1}{N}\sum_i \nabla_{\boldsymbol{y}_i}\mathcal{L}_{EMMUA} = \frac{1}{N^2\tau}\left(2\sum_i(\boldsymbol{x}_i-\boldsymbol{y}_i) - \sum_i\sum_{j\neq i}\left(\boldsymbol{Q}^y(i,j)+\boldsymbol{Q}^y(j,i)\right)(\boldsymbol{y}_j-\boldsymbol{y}_i)\right) \tag{59}$$

Notice that by double stochasticity:

$$\sum_i\sum_{j\neq i}\left(\boldsymbol{Q}^y(i,j)+\boldsymbol{Q}^y(j,i)\right)(\boldsymbol{y}_j-\boldsymbol{y}_i) =$$
$$\sum_i\sum_{j\neq i}\boldsymbol{Q}^y(i,j)(\boldsymbol{y}_j-\boldsymbol{y}_i) + \sum_i\sum_{j\neq i}\boldsymbol{Q}^y(j,i)(\boldsymbol{y}_j-\boldsymbol{y}_i) = 0 \tag{60}$$

Since $\boldsymbol{Q}^y(j,i) = \boldsymbol{Q}^y(i,j)$ and we sum both over $\boldsymbol{y}_i - \boldsymbol{y}_j$ and $\boldsymbol{y}_j - \boldsymbol{y}_i$ due to the outer sum. In total:

$$\frac{1}{N}\sum_i \nabla_{\boldsymbol{y}_i}\mathcal{L}_{EMMUA} = \frac{1}{N^2\tau}\left(2\sum_i(\boldsymbol{x}_i-\boldsymbol{y}_i)\right) = \frac{2}{N\tau}(\boldsymbol{\mu}_x-\boldsymbol{\mu}_y) = \frac{2}{N\tau}\boldsymbol{g} \tag{61}$$

Meaning the gradient shrinks the gap throughout training. □

B Experiment Details

Figure 3 measures the margins for both pretrained models and for embeddings we directly optimize using the different losses. To compute the pretrained margins in Fig. 3a, we use a pretrained CLIP B/32 (open-ai) to create the embeddings for the MS-COCO training set. We sample subsets of size N and compute the margins over this subset, and compare it to the random baseline where we sample $\boldsymbol{x}_i$ uniformly on the hypersphere of the same dimension as the pretrained model ($\mathbb{R}^{512}$), set $\boldsymbol{y}_i = \boldsymbol{x}_i$ and compute the margins.

For Fig. 3b, we directly optimize N pairs of embeddings on the unit hypersphere in $\mathbb{R}^{512}$ using the different losses, starting from the same initialization and with the same hyperparameters ($\tau = 0.07$, $\eta = 0.01$, for 10^4 epochs) and full batch gradient descent. We initialize the embeddings by sampling $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{e}_1, \sigma^2 I)$ and $\boldsymbol{y}_i \sim \mathcal{N}(\boldsymbol{e}_2, \sigma^2 I)$ and then normalizing to the unit hypersphere. The vectors $\boldsymbol{e}_{1,2}$ are the first two vectors of the standard basis of $\mathbb{R}^{512}$ and we set $\sigma = 0.01$.

To create Fig. 2 we follow the same setting as above, only optimize the embeddings in the unit ball in $\mathbb{R}^2$ ($\|\boldsymbol{x}_i\| \leq 1$), measuring similarities using $-\|\cdot\|_2^2$ distance instead of cosine similarity. We initialize the two modalities from two isotropic Gaussians with means $[\pm 0.5, 0]$ with covariance $\sigma^2 I$ and $\sigma = 0.01$.

Figure 1 follows Fig. 3a - we measure the margins for pretrained models on the MS-COCO evaluation set (with $\mathcal{X}$ being the images and $\mathcal{Y}$ being the captions) and compare to uniformly sampled points on the unit hypersphere (same dimension, same size of dataset) setting $\boldsymbol{x}_i = \boldsymbol{y}_i$.

For experiments on real datasets, we train models using either the CLIP loss or EMMUA, starting with the same initialization. Hyperparameters are presented in Table 3. To compute quantization robustness as in Table 2, we quantize each coordinate of the embedding vectors to L equally spaced levels between the minimum and maximum values of that coordinate. We repeat this per model and dataset.

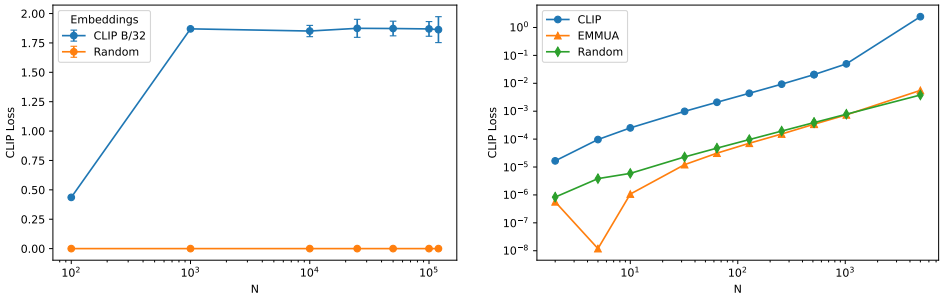
Parameter	MAD	CIFAR100	ImageNet	MS-COCO (finetune)
Variant	MAD/7	S/4	S/8	B/32
Epochs	30	50	50	10
Batch Size	512	512	1024	256
Learning Rate	5e-4	5e-4	5e-4	1e-5
Weight Decay	0.2	0.2	0.2	0.1
Image Size	28	32	64	224

Table 3: All training runs use a cosine scheduler with 5% of the training steps for warmup, and use AdamW optimizer with $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-6}$. All are initialized with $\tau = 0.07$ and $\lambda = \frac{1}{\tau \cdot 1.2}$. For MAD, we followed Schrodi et al. (2025) but trained for less epochs and used all attributes (except color) resulting in 240 different captions. For augmentations, we use random center crop and random horizontal flip for CIFAR100 and ImageNet, and use only random center crop for MS-COCO.

C Further Results

C.1 CLIP Loss of EMMUA

Apparently, EMMUA not only learns a more robust embedding space but can have a lower CLIP loss as well. We measure this in two settings - the first when optimizing the embeddings directly as in the setting of Fig. 3, and the second when comparing a pretrained CLIP to the random baseline. Results are presented in Fig. 6 - in both cases the CLIP trained model actually has a higher CLIP loss.



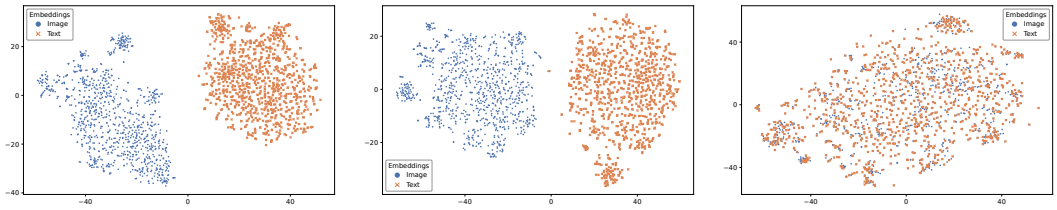
(a) Pretrained

(b) Optimized Embeddings

Figure 6: We calculate the CLIP loss (in batches of 10^3) over subsets of the MS-COCO training set of varying size N, for a pretrained CLIP B/32 and random uniform points on the hypersphere. If CLIP indeed had learned to embed the points in an aligned and uniform manner, it would have both larger margins and lower loss. Plot on the right is in symmetrical log scale. Notice EMMUA isn't monotone due to numerical precision.

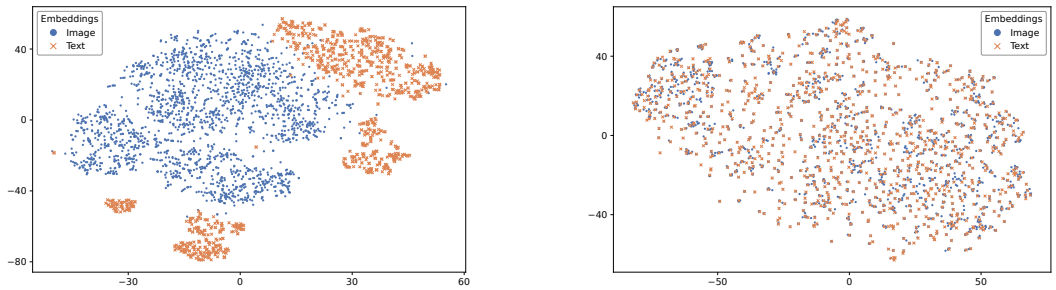
C.2 t-SNE Visualizations

For completeness, we present additional visualizations of the trained models. In all cases, the t-SNE visualizations show that the embeddings learned by EMMUA are completely mixed, while in those of CLIP the two modalities can be separated by a non-linear, or sometimes linear classifier.



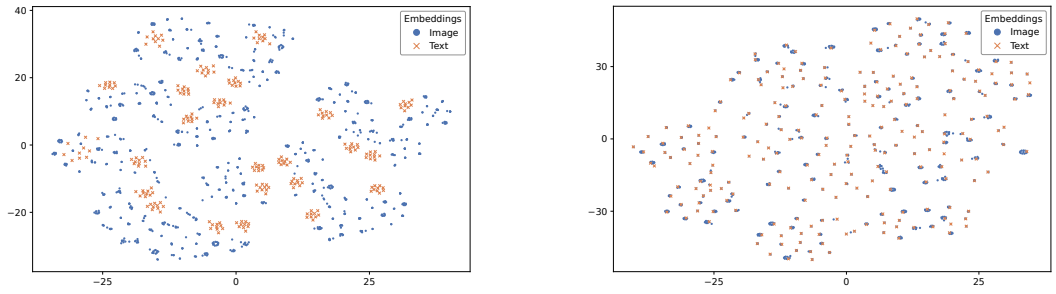
(a) Pretrained (b) Finetuned (CLIP) (c) Finetuned (EMMUA)

Figure 7: t-SNE visualizations of the embeddings of the Flickr30k evaluation set of a pretrained CLIP B/32 and the model finetuned on MS-COCO with the CLIP loss or EMMUA. Finetuning with EMMUA is enough to close the gap.



(a) CLIP (b) EMMUA

Figure 8: t-SNE visualizations of the embeddings of the ImageNet evaluation set of a model trained with the CLIP loss or EMMUA.



(a) CLIP (b) EMMUA

Figure 9: t-SNE visualizations of the embeddings of the MAD evaluation set of a model trained with the CLIP loss or EMMUA.

C.3 Robustness Results

For completeness of Section 4.3, we also present accuracy when adding noise to the text space, as well as robustness when adding noise to both image and text embeddings. We calculate robustness with the definition of Chowers et al. (2026) - as the percent of retrievals (whether correct or not) that stay the same after adding noise.

		Accuracy (Gaussian)				Robustness (Gaussian)		
Dataset	Model	Clean	σ=0.1	σ=0.15	σ=0.2	σ=0.1	σ=0.15	σ=0.2
MAD	CLIP	94.2	41.0	12.4	4.7	42.4	12.7	4.8
	CLIP + Shift	94.2	45.7	12.2	4.1	47.4	12.5	4.2
	EMMUA	94.0	90.9	31.6	7.5	95.6	33.0	7.8
CIFAR-100	CLIP	59.3	39.5	14.4	5.5	50.8	17.9	6.6
	CLIP + Shift	59.2	40.3	13.5	5.6	52.5	16.8	6.9
	EMMUA	56.3	53.2	22.6	9.6	80.4	31.7	13.1
ImageNet	CLIP	55.0	5.5	1.1	0.4	6.5	1.3	0.5
	CLIP + Shift	54.5	6.4	1.2	0.4	7.7	1.3	0.5
	EMMUA	57.3	45.0	8.4	2.1	61.4	10.9	2.8
MS-COCO (finetune)	Pretrained	32.6	0.2	0.1	0.0	0.2	0.1	0.0
	CLIP	43.0	0.8	0.2	0.1	1.0	0.2	0.1
	CLIP + Shift	42.0	0.9	0.2	0.1	1.1	0.2	0.1
	EMMUA	40.9	10.6	1.3	0.3	15.0	1.7	0.4

Table 4: We measure robustness and accuracy when adding gaussian noise to the text embeddings following the protocol of Table 2.

		Margin	Robustness (Quantization)			Robustness (Gaussian)		
Dataset	Model	Mean margin (correct)	L=2	L=4	L=16	σ=0.1	σ=0.15	σ=0.2
MAD	CLIP	0.239	97.8	98.4	99.8	98.3	93.1	76.3
	CLIP + Shift	0.238	98.4	98.5	99.8	98.3	93.1	76.3
	EMMUA	0.650	99.8	99.8	100.0	99.4	98.9	95.0
CIFAR-100	CLIP	0.283	91.1	91.0	97.9	83.3	73.0	59.2
	CLIP + Shift	0.271	90.6	90.6	97.5	83.3	73.0	59.4
	EMMUA	0.642	95.6	95.5	99.1	91.3	86.6	79.7
ImageNet	CLIP	0.085	65.2	65.3	70.7	39.8	18.8	9.2
	CLIP + Shift	0.085	66.8	66.7	79.1	39.8	18.9	9.2
	EMMUA	0.395	93.6	93.7	98.4	86.8	78.2	62.8
MS-COCO (finetune)	Pretrained	0.024	23.9	24.2	18.6	5.6	1.7	0.8
	CLIP	0.036	34.0	33.9	33.6	10.8	3.3	1.4
	CLIP + Shift	0.036	32.8	33.1	44.0	10.9	3.3	1.4
	EMMUA	0.102	69.0	68.9	88.6	50.1	29.3	14.9

Table 5: Quantization and Gaussian robustness when adding the Gaussian noise to the images and quantizing both modalities, following the protocol of Table 2.

C.4 Robustness to Augmentations

Tables 6 and 7 present the robustness of the models examined in Section 4 to different image augmentations of different strengths. While not as single-sided as the results in Table 2, EMMUA still exhibits slight gains over CLIP. All evaluations are on the entire dataset's evaluation sets.

Dataset	Model	Clean Acc.	Gaussian blur σ=0.5	Gaussian blur σ=1.5	Gaussian blur σ=3	Rotation 10°	Rotation 30°	Rotation 60°
CIFAR-100	CLIP	**59.3**	**61.5** (+2.2)	26.4 (-32.9)	14.3 (-45.0)	**51.1** (**-8.2**)	25.7 (-33.6)	17.5 (-41.8)
	CLIP + Shift	59.2	61.0 (+1.8)	25.5 (-33.7)	13.8 (-45.4)	50.2 (-9.0)	24.9 (-34.3)	17.1 (-42.1)
	EMMUA	56.3	58.8 (**+2.5**)	25.6 (**-30.7**)	13.7 (**-42.6**)	46.4 (-9.9)	25.3 (**-31.0**)	17.3 (**-39.0**)
ImageNet	CLIP	**55.0**	47.7 (-7.3)	10.3 (-44.7)	**3.4** (-51.6)	36.5 (-18.5)	19.1 (**-35.9**)	11.9 (**-43.1**)
	CLIP + Shift	54.5	47.2 (-7.3)	10.2 (**-44.3**)	**3.4** (**-51.1**)	35.8 (-18.7)	18.6 (**-35.9**)	11.4 (**-43.1**)
	EMMUA	**57.3**	**50.6** (**-6.7**)	**12.1** (-45.2)	3.2 (-54.1)	**40.0** (**-17.3**)	**20.7** (-36.6)	**12.6** (-44.7)
MS-COCO (finetune)	CLIP	**43.0**	**42.3** (-0.7)	36.8 (-6.2)	28.9 (-14.1)	37.7 (-5.3)	31.1 (-11.9)	21.5 (-21.5)
	CLIP + Shift	42.0	41.7 (**-0.3**)	35.9 (-6.1)	28.9 (-13.1)	36.8 (-5.2)	30.3 (-11.7)	20.9 (-21.1)
	EMMUA	40.9	40.4 (-0.5)	**36.9** (**-4.0**)	**31.0** (**-9.9**)	**37.9** (**-3.0**)	**31.8** (**-9.1**)	**22.0** (**-18.9**)

Table 6: Accuracy under Gaussian blur and rotation of the input images. The difference in clean accuracy is shown in parentheses, and the highest accuracy and smallest change relative to clean accuracy are in bold. For 12 out of 18 combinations (3 datasets, 6 augmentations), EMMUA has lower drop in accuracy relative to CLIP.

Dataset	Model	Clean Acc.	Color jitter ×1.2	Color jitter ×1.6	Color jitter ×2	Perspective d=0.2	Perspective d=0.4	Perspective d=0.6
CIFAR-100	CLIP	**59.3**	**49.3** (**-10.0**)	**20.7** (-38.6)	**11.1** (-48.2)	**33.3** (-26.0)	**22.6** (-36.7)	**13.3** (-46.0)
	CLIP + Shift	59.2	49.1 (-10.1)	20.2 (-39.0)	10.6 (-48.6)	32.6 (-26.6)	21.8 (-37.4)	12.7 (-46.5)
	EMMUA	56.3	46.1 (-10.2)	19.3 (**-37.0**)	10.2 (**-46.1**)	31.3 (**-25.0**)	21.2 (**-35.1**)	12.1 (**-44.2**)
ImageNet	CLIP	55.0	49.7 (-5.3)	27.1 (-27.9)	12.0 (-43.0)	38.6 (-16.4)	31.2 (-23.8)	16.5 (-38.5)
	CLIP + Shift	54.5	49.0 (-5.5)	26.8 (-27.7)	11.9 (**-42.6**)	37.9 (-16.6)	30.8 (-23.7)	16.3 (**-38.2**)
	EMMUA	**57.3**	**52.4** (**-4.9**)	**30.1** (**-27.2**)	**14.1** (-43.2)	**42.5** (**-14.8**)	**33.9** (**-23.4**)	**17.2** (-40.1)
MS-COCO (finetune)	CLIP	**43.0**	**41.6** (-1.4)	**36.3** (-6.7)	**30.5** (-12.5)	**38.7** (-4.3)	**34.3** (-8.7)	27.0 (-16.0)
	CLIP + Shift	42.0	41.2 (**-0.8**)	35.8 (-6.2)	29.7 (-12.3)	37.9 (-4.1)	32.8 (-9.2)	25.5 (-16.5)
	EMMUA	40.9	39.9 (-1.0)	35.2 (**-5.7**)	29.6 (**-11.3**)	37.4 (**-3.5**)	34.0 (**-6.9**)	**28.3** (**-12.6**)

Table 7: Accuracy under color jitter and perspective augmentations of the input images. The difference in clean accuracy is shown in parentheses, and the highest accuracy and smallest change relative to clean accuracy are in bold. For 14 out of 18 combinations (3 datasets, 6 augmentations), EMMUA has lower drop in accuracy relative to CLIP.