Is the Modality Gap a Bug or a Feature? A Robustness Perspective

Rhea Chowers
Hebrew University
Jerusalem, Israel
rhea.chowers@mail.huji.ac.il

Oshri Naparstek
IBM Research
Haifa, Israel
Oshri.Naparstek@ibm.com

Udi Barzelay
IBM Research
Haifa, Israel
udib@il.ibm.com

Yair Weiss
Hebrew University
Jerusalem, Israel
yair.weiss@mail.huji.ac.il

Abstract

Many modern multi-modal models (e.g. CLIP) seek an embedding space in which the two modalities are aligned. Somewhat surprisingly, almost all existing models show a strong modality gap: the distribution of images is well-separated from the distribution of texts in the shared embedding space. Despite a series of recent papers on this topic, it is still not clear why this gap exists nor whether closing the gap in post-processing will lead to better performance on downstream tasks. In this paper we show that under certain conditions, minimizing the contrastive loss will lead to a representation in which the two modalities are separated by a global gap vector that is orthogonal to the embeddings of both modalities. We also show that under these conditions the modality gap is monotonically related to robustness: decreasing the gap does not change the clean accuracy of the models but makes it less likely that a model will change its output when small, semantically inconsequential changes are made to the input. Our experiments show that for many real-world VLMs we can significantly increase robustness by a simple post-processing step that moves one modality towards the mean of the other modality, without any loss to clean accuracy.

1. Introduction

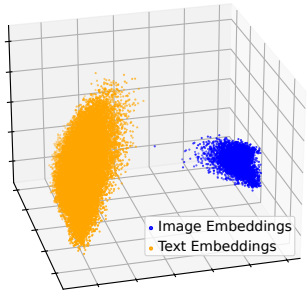
Foundation models are a common and successful approach to solving a variety of tasks - models are trained on extremely large datasets, and then adapted to various tasks either by fine tuning or zero shot application. Both these downstream uses rely on the assumption that the original models learned a meaningful embedding space that captures various semantic aspects of the data, making it useful for different tasks.

A specific class of foundation models are multi-modal models. These are trained to learn a shared embedding space for different data types such as images and texts by aligning pairs of similar images and texts via a contrastive loss [32, 34, 42]. These shared embedding spaces, specifically that of CLIP [22], are commonly used for various tasks such as zero shot classification, text-to-image retrieval, text to image generation, and more.

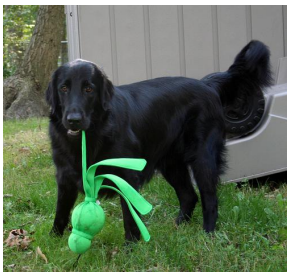
The success of multi-modal models on different tasks and their widespread integration in various models, most notably text to image generation models [24], might imply that they have successfully solved the optimization problem on which they were trained, yet some open questions persist. One of these regards the "modality gap" [13] - a phenomenon in which the two modalities are embedded into linearly separable regions of the unit sphere. Figure 1a shows projections of CLIP's embedding of the MS-COCO validation set onto its first three principal components. Clearly, the two modalities are well separated in embedding space. This phenomenon, which has been shown to be prevalent across various multi-modal models and data types, contradicts the models' training objective which pushes similar texts and images to perfectly overlap - a property termed "alignment" [36].

Although various recent works have tried to explain the cause and downstream effects of the modality gap [7, 13, 14, 23, 29], these still are not completely understood. Figure 2 presents the application of various multi-modal models on common downstream tasks. The performance of these models varies non-monotonously when adjusting the gap by moving the modality means towards each other, implying no clear relation between performance and and the size of the gap.

Another open topic regarding multi-modal models and



(a) CLIP's embedding of MS-COCO validation set



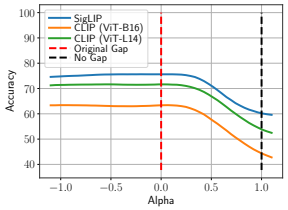
(b) "a photo of a **dog** / frog" ∨
"photograph of the dog / **frog**" ×

Figure 1. Left: Projections of CLIPs embedding of the MS-COCO validation set onto its first 3 principal components. A clear separation between images and texts is evident. Right: An image from the Imagenet [1] validation set that's misclassified by CLIP when changing the caption template. Multi-modal models can lose more than 6% of their accuracy when replacing the caption template.

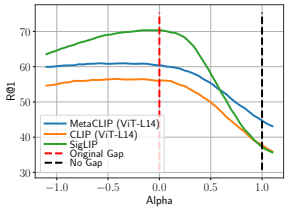
deep learning in general focuses on the robustness of models to small variations in their input. Various works have shown that despite their large training corpus and the numerous augmentations used during their training, the performance of these models on downstream tasks is sensitive to both small and natural variations in the input such as single pixel shifts [30] and caption rephrasing [11, 44]. Figure 1b demonstrates this brittleness. When trying to classify an image as a dog or a frog using CLIP, the model outputs different predictions depending on the the two texts with which the image is compared, even if the change is semantically meaningless.

In this work we establish a direct link between these two lines of work. In Sec. 3 we prove that under realistic assumptions, multi-modal models learn an embedding space containing a modality gap due to their initialization and the use of contrastive loss. We continue to prove that the size of the gap is positively correlated with the robustness of these models, i.e. their ability to perform consistently under small changes to the embedding space. Our theory along with empirical verification on real models and datasets, with both controlled (i.e. Gaussian) and real noise (i.e. text rephrasing), leads us to conclude that if robustness is a desired property of the models then the modality gap is indeed a bug caused by an interaction of the multi-modal contrastive loss and the initialization scheme.

Inspired by our theoretical findings, in Sec. 4 we present a simple, efficient algorithm to close the gap which can be applied as a post processing step when using the different models' embeddings spaces. In Sec. 5 we demonstrate a use case of this algorithm in dealing with various noise models such as quantization and other real noise settings where the theoretical assumptions are not met.



(a) Zero Shot on Imagenet



(b) I2T Retrieval on MS-COCO

Figure 2. Is the modality gap a bug or a feature? Changing the gap by moving the text embeddings by $\alpha \cdot \vec{g}$ has an inconsistent effect on downstream performance. The figure follows [13] and shows that for some datasets, models benefit from slightly enlarging the gap (Fig. 2a), some from maintaining it (Fig. 2b) .

2. Related Work

Multi-modal contrastive learning [42] is a method for learning a shared embedding space for different modalities e.g. images and texts. Models trained in such manner such as CLIP [22], SigLIP [41] and others [5] show impressive performance on many downstream tasks such as text to image retrieval and zero shot classification. This impressive performance has led to widespread use of the CLIP embedding space in various tasks, most notably in text to image generation [8, 24]. While the contrastive loss [32, 34] of multi-modal models pushes the modality pairs' representations to be aligned and uniformly spread on the unit sphere [36], in practice various multi-modal models map the different modalities to distinct areas of the shared embedding space thus creating a modality gap [13].

Several explanations as to the formation of the gap were suggested, stemming from either the data used to train these models [13, 27], the inherent differences between text and images [27] or the training procedure [29, 38]. In addition to explaining the formation of the gap, previous works have also attempted to understand its effects on downstream performance. It has previously been shown that a large gap is negatively correlated in specific settings of cross modal retrieval [14, 17], but generally recent work concludes that there is at most a weak effect of the gap on downstream performance [7, 13, 27] while other factors such as model and embedding size seem to have a stronger effect [27]. Some works even suggest maintaining the modality gap [7, 23], or show that it has no effect in specific downstream settings [43]. We build on these findings and and prove that the gap indeed has negative effect on downstream performance in terms of robustness.

Alongside the attempt to understand the gap and its implications, previous works suggested various methods for closing the gap. These were mainly motivated by failure modes in specific settings [14, 18], or by the inherent contradiction of the gap to the contrastive loss objective. Some works

suggest closing the gap by altering the training procedure of multi-modal models with specific losses and augmentations [19, 38], while others include the training of models that implicitly close the gap as part of pipelines performing complex downstream tasks such as text to image generation [20, 24]. Our work falls in line with these, but is applicable in a general setting - for any downstream task that relies on cross modal nearest neighbor retrieval, and with no additional training required.

A different line of work studies multi-modal models in the context of robustness [3, 33, 40]. Various works have shown that their performance on downstream tasks is sensitive to both textual variations [21, 44], adversarial attacks [25], and various phenomena in image distributions such as augmentations [21], spurious correlations [35], natural variations [30] and more [37]. Following these shortcoming many methods have been proposed to improve these models' robustness, mainly via finetuning and retraining with different objectives or data [2, 9, 10, 26, 31, 44]. Our theoretical work aims to enrich the understanding of these shortcomings by proving that a modality gap causes degrading in robustness.

3. Theory

3.1. Notation and Definitions

As in previous work, we assume two modalities that are embedded on the d-dimensional unit hypersphere. Our embeddings are in the form of sample matrices $\mathcal{X} \in \mathbb{R}^{N\times d}$ and $\mathcal{Y} \in \mathbb{R}^{M\times d}$ with modality means $\bar{x} = \frac{1}{N}\sum_{\vec{x}\in\mathcal{X}} \vec{x}$ and $\bar{y} = \frac{1}{M}\sum_{\vec{y}\in\mathcal{Y}} \vec{y}$. Many downstream tasks rely on cross modal retrieval - finding the nearest neighbor of sample $\vec{y} \in \mathcal{Y}$ among the samples in the other embedding $\mathcal{X}$ by computing ℓ_2 distance between the two:

$$\mathrm{NN}(\vec{y}, \mathcal{X}) := \arg\min_{\vec{x}\in\mathcal{X}} \|\vec{x} - \vec{y}\|^2 \tag{1}$$

Since most works assume the embeddings lie on the unit hypersphere, the above is usually calculated using cosine similarity.

We focus on models trained with the multi-modal contrastive loss. During training we are given a paired dataset $\{x_i, y_i\}$ (e.g. image and matching caption) and minimize the following loss:

$$\mathcal{L}(\mathcal{X}, \mathcal{Y}) = \\ -\frac{1}{2}\left(\frac{1}{N}\sum_{i=1}^{N}\log \ell(\vec{x}_i, \mathcal{Y}) + \frac{1}{N}\sum_{j=1}^{N}\log \ell(\vec{y}_j, \mathcal{X})\right) \tag{2}$$

Where ℓ contrasts a single embedding with the other modality:

$$\ell(\vec{x}_i, \mathcal{Y}) = \frac{e^{-\|\vec{x}_i - \vec{y}_i\|/\tau}}{\sum_{\vec{y}\in\mathcal{Y}} e^{-\|\vec{x}_i - \vec{y}\|/\tau}} \tag{3}$$

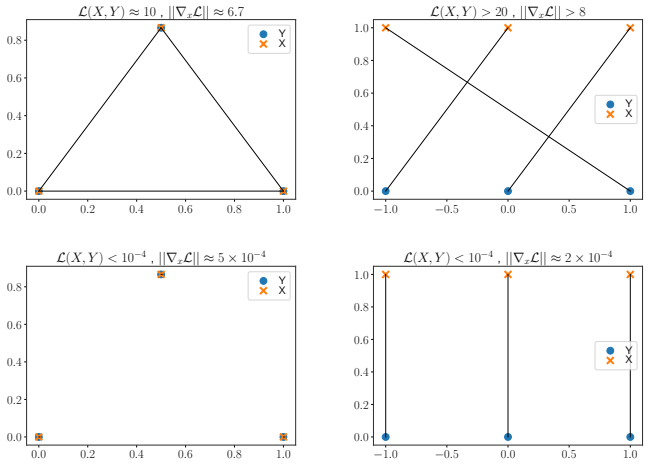


Figure 3. Three points in each modality in $\mathbb{R}^2$ and the corresponding multi-modal contrastive loss along with the average magnitude of the gradient of the loss. Lines connect between true pairs. As long as the points satisfy relative alignment - the true pair of any point is also its nearest neighbor - the loss and the gradient magnitude are close to zero, even when there exists a gap.

with $\tau > 0$ being the temperature parameter.

We follow previous works [43] and define, for a paired dataset (such as during training), the local gap as the vector between a pair from the two modalities:

$$\vec{g}_i = \vec{x}_i - \vec{y}_i \tag{4}$$

We also define the global gap [13, 29, 43] as the vector between the modality means:

$$\vec{g} = \frac{1}{M}\sum_{\vec{y}\in\mathcal{Y}} \vec{y} - \frac{1}{N}\sum_{\vec{x}\in\mathcal{X}} \vec{x} \tag{5}$$

Notice that if the two modalities are balanced ($N = M$), the global vector is the average of local vectors. Nevertheless, the global gap vector is well defined for any dataset consisting of two modalities, whether a bijective pairing exists or not.

3.2. Why Should a Global Gap Exist?

Since the contrastive loss rewards embeddings in which the representations in the two modalities are aligned [36], the presence of a global gap in embeddings learned using such a loss is highly surprising. Indeed, it is easy to show that the global minimum of the contrastive loss is obtained when the representations are maximally aligned, i.e. $\forall i : x_i = y_i$, entailing no local or global gap, i.e. $\vec{g}_i = 0$. But it is also easy to see that there exist many other embeddings that achieve almost the same loss as the globally aligned embedding. Figure 3 shows an example. The lines are the local gap vectors $\vec{g}_i$ which show the correspondence between

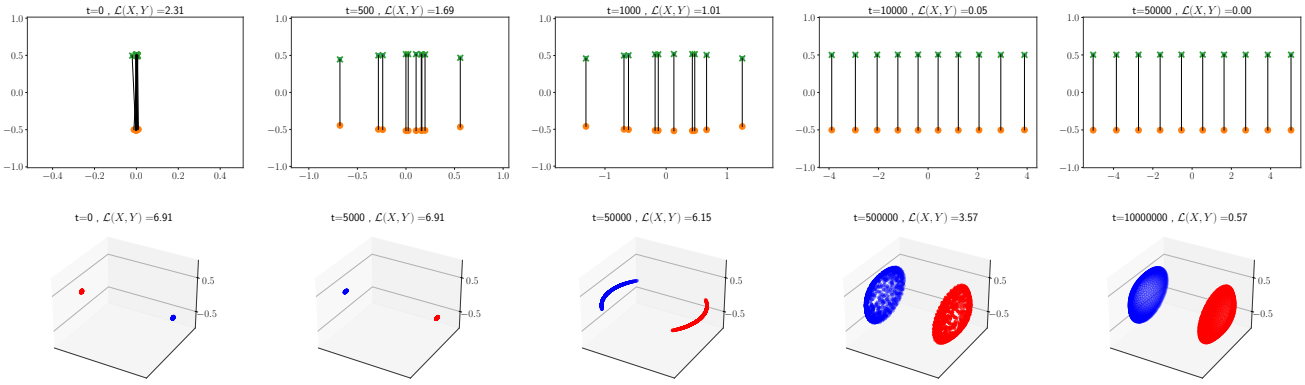


Figure 4. The evolution of embeddings using gradient descent on the contrastive loss (8) starting with two tight clusters. Both when using an unnormalized embedding space (top) and when constricting the embeddings to the sphere (bottom) they converge to a solution that has almost zero loss and for which a global gap vector exists and the gap vector is orthogonal to both modalities. We prove that minimizing the contrastive loss will lead to such a solution under certain assumptions. For full training details see the supplementary.

the two modalities: ideally we want circles and crosses that are connected to lie on top of each other in embedding space, meaning $\vec{g}_i = 0$. As can be seen in the figure, the loss and the gradient can be made arbitrarily close to zero without perfect alignment: as long as the corresponding items are closer than all other cross-modal pairs, the loss and the gradient will approach 0. In particular, the solution in the bottom right (where there is a modality gap) achieves practically the same loss and gradient magnitude as the desired solution.

Figure 4 shows dynamics of points in low dimensions, in which we minimize the contrastive loss with respect to the embeddings coordinates. In both of these cases, training converges to a solution that has almost zero loss but for which a global gap vector remains. We now prove that for an unnormalized embedding space this will happen under certain assumptions.

The assumption is defined in terms of two stochastic matrices that are the basis of the contrastive loss. Define

$$Q^x(i,j) = \frac{e^{-\|x_i - y_j\|^2/\tau}}{\sum_{j'} e^{-\|x_i - y_{j'}\|^2/\tau}} \tag{6}$$

and:

$$Q^y(i,j) = \frac{e^{-\|x_i - y_j\|^2/\tau}}{\sum_{i'} e^{-\|x_{i'} - y_j\|^2/\tau}} \tag{7}$$

Which are known as the softmax of the logit matrix [22]. Using these:

$$\mathcal{L}(\mathcal{X},\mathcal{Y}) = -\left(\frac{1}{N}\sum_{i=1}^{N} \log Q^x(i,i) + \log Q^y(i,i)\right) \tag{8}$$

By construction, both Q^x, Q^y are singly stochastic matrices, but in certain cases they will also be *doubly stochastic*. This will happen, for example, when $\tau \to 0$ and the nearest neighbor matchings between the two modalities are a perfect matching or when $\tau \to \infty$ which leads to them being constant matrices, i.e. $Q^x = Q^y = \frac{1}{N}\vec{\mathbf{1}}\vec{\mathbf{1}}^T$.

Theorem 3.1. *Assume that at a given iteration t there exists a unique direction $\vec{v}$ such that $\forall \vec{x}_i \in \mathcal{X} : \vec{v}^T\vec{x}_i = a$ and $\forall \vec{y}_i \in \mathcal{Y} : \vec{v}^T\vec{y}_i = b$. Assume that for all iterations after t the matrices Q^x, Q^y are doubly stochastic. Then gradient descent on the contrastive loss (equation 8) will converge to a solution where all the local gap vectors are the same $\vec{g}_i = \vec{g}$ and furthermore, $\vec{g}_i$ and $\vec{g}$ are orthogonal to both X and Y.*

Proof Sketch. We can write the loss $\mathcal{L}(\mathcal{X},\mathcal{Y})$ as $\mathcal{L} = L_x + L_y$. The gradients of the first loss with respect to each of the modalities are:

$$\frac{\partial L_x}{\partial y_j} = -2(x_j - y_j) + \sum_i 2Q^x(i,j)(x_i - y_j) \tag{9}$$

$$\frac{\partial L_x}{\partial x_i} = -2(x_i - y_i) + \sum_j 2Q^x(i,j)(x_i - y_j) \tag{10}$$

By the assumption that $v^T x_i = a$ for all x_i and $v^T y_i = b$ for all y_j we see that:

$$\begin{aligned} v^T\frac{\partial L_x}{\partial y_j} &= -2v^T(x_j - y_j) + 2\sum_i Q^x(i,j)v^T(x_i - y_j) \\ &= -2(a-b) + 2\sum_i Q^x(i,j)(a-b) \\ &= 0 \end{aligned} \tag{11}$$

where the last equality used the assumption of double stochasticity, From symmetry, a similar result holds for the

derivatives of the two losses with respect to both modalities: the gradient in direction $\vec{v}$ will be zero. This means that gradient descent will not change the values of either modality in direction $\vec{v}$ and hence will converge to a solution in which this direction is unchanged and in all other directions $\vec{u}$, the modalities will be perfectly aligned $\vec{u}^T x_i = \vec{u}^T y_i$. This means that for all points, the gap $\vec{g}_i$ will be in direction $\vec{v}$ and be equal to $\frac{|b-a|}{\|\vec{v}\|}\vec{v}$. Notice that by the assumption that all points in the same modality have the same projection in direction $\vec{v}$, the gap vector will be orthogonal to both modalities. □

Theorem 3.1 explains the dynamics seen in figure 4: in the top figure, the gradient in the vertical dimension remains zero throughout the learning process, and the loss continues to decrease by moving the points only in the horizontal direction until the loss reaches the value of zero with a constant, orthgonal gap vector. In the bottom figure, the loss decreases while moving in two directions, but in the third direction the gradient is almost zero in all iterations so again a global, orthogonal gap remains at the final iteration. As we show in the supplementary material, there are very similar initial conditions in two and three dimensions for which the doubly stochastic assumption *does not* hold, and in these cases the dynamics will converge to a perfectly aligned solution.

This phenomenon of an orthogonal gap has previously been noticed [43] to hold empirically for multi-modal models trained on various datatypes. We term this the *global orthogonality assumption* and formally define it as:

Assumption 3.2. The gap vector $\vec{g}$ is orthogonal to the affine subspaces defined by $\mathcal{X}$ and $\mathcal{Y}$:

$$\begin{aligned} \forall x \in \mathcal{X}, y \in \mathcal{Y} : \cos\left(x - \bar{x}, \vec{g}\right) \\ = \cos\left(y - \bar{y}, \vec{g}\right) = 0 \end{aligned} \tag{12}$$

Our theorem, while assuming an unnormalized embedding space, provides intuition to the source of the orthogonality assumption - it seems that initialization plays a key role - under certain conditions, an orthogonal gap between the two modalities at initialization would remain throughout training.

While many solutions exist to minimizing Eq. (8) (as can be seen in Fig. 3), the next section shows that these conditions on initialization are indeed the case for real multi-modal models in high dimensions, and that they converge to an orthogonal gap as well.

3.3. Empirical Evidence for Orthogonality

Although Theorem 3.1 applies to unnormalized embedding spaces, Fig. 4 (bottom) shows that an orthogonal solution with a gap between the modalities can be reached with embeddings constrained to the hypersphere as well when initializing tight clusters. Figure 5 shows that a similar case

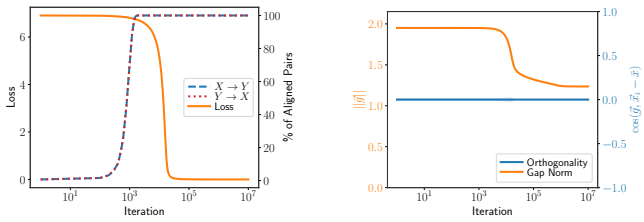


Figure 5. We follow Shi et al. [29] and learn embeddings $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{1000\times512}$ on the hypersphere using the contrastive loss (8) and gradient descent. While the loss decreases and training converges (left), a major gap remains between the embeddings and orthogonality holds (right, measured using Eq. (12)). More examples and full experimental details can be found in the supplementary.

also happens with spherical embeddings in high dimensions - when the embeddings are initialized with an orthogonal gap, training converges to a solution in which a global gap exists and it is orthogonal to both modalities.

But do real multi-modal models follow these dynamics? It has previously been reported that the different modalities of multi-modal models are initialized in tight, linearly separable clustered. This phenomenon has been termed the "cone effect" [13]. We add to these results in Fig. 6 - not only are random initializations of CLIP tightly clustered, but they also satisfy the orthogonality assumption 3.2 in high dimensions.

The final evidence comes from [43] who have empirically investigated trained multi-modal models on various datatypes. They summarize their finding by saying that the local gap vectors "can be approximated by a constant vector" which is "orthogonal to the span of image embeddings and text embeddings. For completeness, we recreate these results in Fig. 7b.

To summarize, multi-modal models are commonly initialized in non-intersecting tight clusters with an existing gap which maintains global orthogonality. Additionally, this gap remains at the end of the models' training. This, along with our simulations supports the hypothesis that the training dynamics of normalized embedding spaces behave similar to Theorem 3.1.

3.4. The Modality Gap Decreases Robustness

We define robustness ("Rob") as the probability that when adding noise sampled from some distribution $\mathcal{P}$ to the embeddings of modality $\mathcal{X}$, the nearest neighbor of some $y \in \mathcal{Y}$ will not change:

$$\text{Rob}(\mathcal{X}, \mathcal{Y}, \mathcal{P}) = \mathbb{E}_{y\sim\mathcal{Y}, \epsilon\sim\mathcal{P}}\left[\mathbb{1}_{\text{NN}(y,\mathcal{X})=\text{NN}(y,\mathcal{X}+\epsilon)}\right] \tag{13}$$

Notice that in the zero shot classification setting, this definition of robustness captures semantically meaningful changes - any change of nearest neighbor from image embedding to text, necessarily changes the classification of that image.

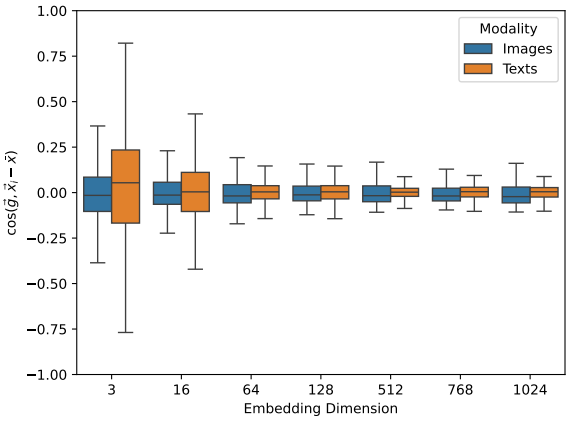
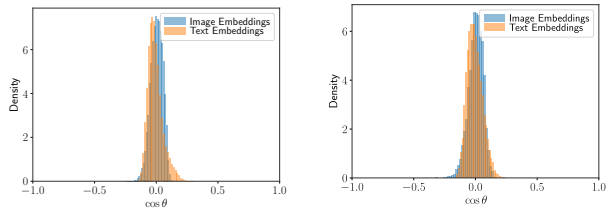


Figure 6. We measure orthogonality using Eq. (12) for 50 different initializations of CLIP (ViT-B-16) on 5000 randomly sampled images and texts from the Flickr30k dataset [12]. In high dimensions, the cosine of angles between the samples and global gap vector is close to 0.



(a) SigLIP on Imagenet (b) CLIP on MS-COCO

Figure 7. The orthogonality assumption - the cosine of angle θ between $\vec{g}$ and each embedding in each modality is nearly 0 for different models and datasets, confirming assumption 3.2.

In the case that multiple texts can match an image (such as multi-label classification, or MS-COCO) we define soft robustness that captures only semantic changes, and not any change to nearest neighbor. We define $\mathcal{X}(y)$ to be the subset of labels in $\mathcal{X}$ s.t. $NN(y, \mathcal{X}) \in \mathcal{X}(y)$.

$$\text{Soft-Rob}(\mathcal{X}, \mathcal{Y}, \mathcal{P}) = \mathbb{E}_{y\sim\mathcal{Y}, \epsilon\sim\mathcal{P}} \left[\mathbb{1}_{\mathcal{X}(y)=(\mathcal{X}+\epsilon)(y)}\right] \quad (14)$$

We now ask how does the existence of the gap relates to the robustness of the models to noise in the embedding space. Figure 8 provides intuition to such an effect. We treat cross modal retrieval as a binary classification task where an image is classified between the true caption and some wrong one. Any noise added to the text embeddings, e.g. that resulting from slight rephrasing of the text, changes the decision boundary between the two classes. In the case of a gap between the cluster of image embeddings and text embeddings, as image embeddings are further away from the texts, their sensitivity to change in the classification decision

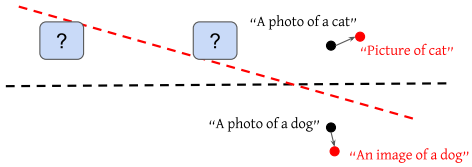


Figure 8. Illustration of the relationship between robustness and the modality gap. Two image embeddings will be classified differently with slight rephrasing depending on the "gap" - their distance from the text embeddings.

boundary grows. We prove this formally in the following theorem.

Theorem 3.3. *Let $\{\vec{y}_i\}$ be a set of points in one modality (e.g. an image embedding) and $\vec{x}_1, \vec{x}_2 \in \mathcal{X}$ two vectors in the second modality (e.g. captions) with mean $\bar{x} = \frac{\vec{x}_1+\vec{x}_2}{2}$. Assume $\vec{x}_1$ is the nearest neighbor of some $\vec{y}$ in $\mathcal{X}$. Under assumption 3.2, moving any point $\vec{y}$ towards the other modality by a global translation $\vec{g} = \bar{y} - \bar{x}$ increases robustness: the probability that the nearest neighbor of $\vec{y}$ in $\mathcal{X}$ does not change after adding noise to $\mathcal{X}$, with covariance $\sigma^2 I$ and zero empirical mean.*

Proof Sketch. The proof follows the intuition presented in Fig. 8 - as the query vector $\vec{y}$ is closer to modality $\mathcal{X}$, a larger rotation of the separating hyperplane is required to decrease robustness. Therefore the theorem holds as subtracting the global gap vector decreases the distance between $\vec{y}$ and $\mathcal{X}$. □

For full proof and generalization to more than 2 points in $\mathcal{X}$ and general noise we refer the reader to the supplementary. Theorem 3.3 proves that under the orthogonality assumption, when moving one modality towards the other by translating it with the global gap vector then the embedding spaces become more robust and less sensitive to noise. This provides us with clear motivation to close the gap.

We note that although the theorem is phrased in terms of the global gap vector, any vector that maintains orthogonality to modality $\mathcal{X}$ will suffice, as long as translating $\vec{y}$ with it will decrease the distance between $\vec{y}$ and the mean of modality $\mathcal{X}$.

3.5. Closing the Gap Without Effecting Performance

Despite this clear motivation, previous works [13, 27] have been inconclusive as to the effect of naively moving one modality towards the other by subtracting the gap vector $\vec{g}$. Figure 2 shows that there is no consistent change to performance when closing the gap - most models are invariant to at least a partial mitigation of the gap, some benefit from it (CLIP ViT-B16 in Fig. 2a), and in some cases enlarging it even improves performance (MetaCLIP in Fig. 2b).

We therefore strive to understand when does closing the gap have no effect on downstream performance, therefore solely improving robustness. Since most downstream tasks such as zero shot classification or text to image retrieval are based on cross modal nearest neighbors, we study the change in nearest neighbors when varying the gap.

Theorem 3.4. *As in Theorem 3.3, let $\vec{v}$ be some vector that is orthogonal to the affine subspace containing the modality $\mathcal{X}$. Then translating the embeddings of modality $\mathcal{X}$ by $\alpha \cdot \vec{v}$ for some $\alpha \in \mathbb{R}$, such that $\forall x_i \in \mathcal{X} : x_i \leftarrow x_i + \alpha \cdot \vec{v}$, does not change the nearest neighbor of $\vec{y}$ in $\mathcal{X}$ for any $y \in \mathcal{Y}$.*

Proof Sketch. Since the vector $\vec{v}$ is orthogonal to $\mathcal{X}$, the variance of modality $\mathcal{X}$ in $\vec{v}$ is zero, therefore moving $\mathcal{X}$ by $\alpha \cdot \vec{v}$ changes the distance from any $\vec{y} \in \mathcal{Y}$ to all $\vec{x} \in \mathcal{X}$ by the same constant, meaning cross modal nearest neighbors are preserved. For full proof see the supplementary. □

Under assumption 3.2 such a nearest neighbor preserving direction is simply the gap vector $\vec{g}$. Therefore Theorem 3.4 ensures us that narrowing the gap along this direction would not change performance of the model on downstream tasks such as zero shot classification.

4. Algorithm

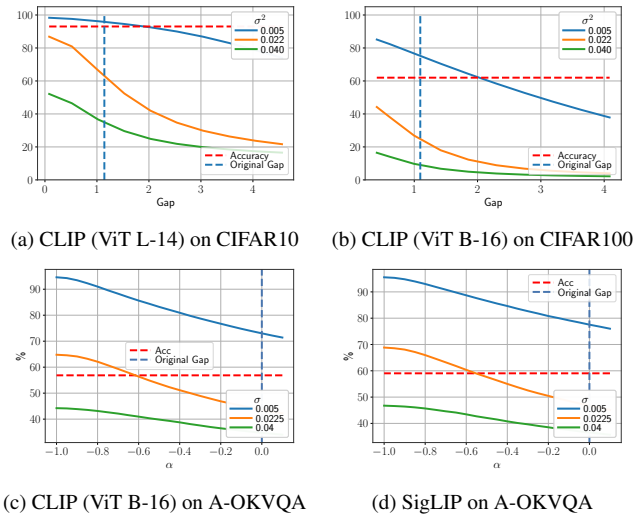
From Sec. 3 we conclude that although the contrastive loss maintains the gap created at initialization throughout the training, it is desirable to close the gap in order to improve robustness. Theorems 3.3 and 3.4 provide us with a tool to do so - move the modalities towards each other along the global gap vector, as under assumption 3.2 this provably does not affect performance on downstream tasks when there is no noise at all.

Since the gap vector is mostly but not completely orthogonal to the modality subspaces (Fig. 7), we suggest projecting the gap vector to be exactly orthogonal to the affine subspace of the modality being retrieved before closing the gap. This is done by computing the principal components V of the modality being retrieved and projecting the gap $\vec{g}$ to the orthogonal complement of the components:

$$\vec{g}' \leftarrow \vec{g} - VV^T\vec{g} \tag{15}$$

Afterwards, the modality being retrieved is moved towards the mean over the other modality by translating $\mathcal{Y} \leftarrow \mathcal{Y} - \vec{g}'$, therefore decreasing the distance of any query point to the retrieved modality, as per Theorem 3.3. Theorem 3.4 assures us that closing the gap in the direction of $\vec{g}'$ would not affect performance on any downstream tasks that relies on computing nearest neighbors. We also provide an extension to this algorithm for cases where the orthogonality assumption does not hold, and see the supplementary.

Contrary to previous attempts at closing the gap, our method requires no retraining or finetuning of the models [18, 19], or the training of new ones (such as the prior network in [20, 24]). Our method is also general and doesn't adhere to a specific setting [14] - any task using cross modal nearest neighbors could benefit from it.



(a) CLIP (ViT L-14) on CIFAR10 (b) CLIP (ViT B-16) on CIFAR100

(c) CLIP (ViT B-16) on A-OKVQA (d) SigLIP on A-OKVQA

Figure 9. The zero shot classification accuracy, multiple choice VQA accuracy and robustness under noise $\eta \sim \mathcal{N}(0, \sigma^2 I)$, on various models and datasets. As predicted by our theory, robustness monotonically increases when the gap is closed while accuracy is maintained.

5. Experiments

We turn our attention to understanding how well our theory and algorithm work in practice for various real world models and datasets. We follow [13] and experiment on standard zero shot classification and image-text retrieval tasks. We create embeddings of commonly used, pretrained, multi-modal models using using the openclip library [6].

To measure robustness, we empirically estimate Eq. (13):

$$\tilde{\text{Rob}}(\mathcal{X}, \mathcal{Y}, \mathcal{P}, \alpha) = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}} \mathbb{1}_{\text{NN}(y,\mathcal{X}+\alpha\cdot\vec{g})=\text{NN}(y,(\mathcal{X}+\epsilon_i)+\alpha\cdot\vec{g}} \tag{16}$$

Where K is the number of samples of noise, $\epsilon_i \sim \mathcal{P}$ is a sample from the noise model $\mathcal{P}$ (such as quantization or gaussian noise) and $\alpha \in \mathbb{R}$ is a scalar controlling the amount of the gap that we close. We set $K = 1$ (e.g. when there is a single deterministic quantization) unless stated otherwise.

5.1. Robustness Under Controlled Noise

Our first set of experiments attempts to verify our theory and algorithm in the presence of controlled noise added to the embedding space. We add varying degrees of independent Gaussian noise with zero mean and variance σ^2 to each text

embedding $\vec{x} \in \mathcal{X}$ and translate the text embeddings by $\alpha \cdot \vec{g}'$ according to our algorithm (Sec. 4). We do this for different values of α therefore expanding or closing the gap. For each small change in the gap, we measure both accuracy and empirical robustness (Eq. (16)) averaging over $K = 100$ samples of noise.

We test our algorithm on various models, in both zero shot classification and visual question answering (VQA) on the A-OKVQA dataset [28], following the evaluation protocol of Ghosal et al. [4] for CLIP-like models. Results are displayed in Fig. 9 - when applying our algorithm under gaussian noise robustness greatly increases when closing the gap across all models and tasks, and greatly decreases when expanding the gap. When the noise is increased, understandingly robustness is lower but still increases with closing of the gap. See the supplementary for similar results on other noise distributions.

As predicted by Theorem 3.4, the zero shot and VQA accuracy do not change when closing the gap in the direction orthogonal to the affine subspace of texts.

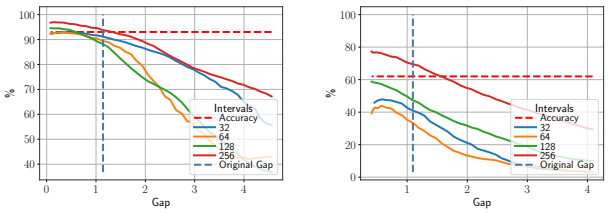
5.2. Robustness to Quantization

We continue and test our theory and algorithm on noise distributions that apply to real world applications. A setting of uncorrelated noise is that of quantization [16]. While many forms of model quantizations exist, we limit ourselves to the simple case of post-training quantizing of the embedding vectors alone, a setting relevant to tasks in which the embeddings were calculated in advance e.g. retrieval augmented generation (RAG) [39]. We close the gap to varying degrees and then quantize the text and image embedding vectors to different numbers of intervals between a constant range of $[-3, 3]$ (as all embeddings are on the unit sphere), and measure robustness.

Empirically, we find that since quantization noise does not always have zero mean, it is better to close the gap until the gap is minimized in the *quantized* embedding space rather than the original one. Figure 10 displays results for these experiments. As can be seen, robustness is greatly improved when closing the gap before quantizing the embedding vectors.

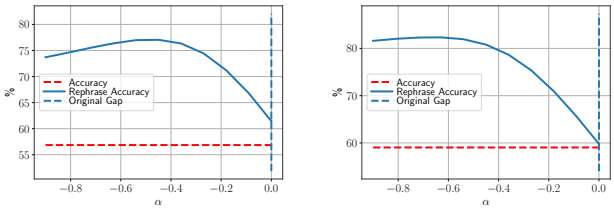
5.3. Robustness to Rephrasing

We now turn our attention to text rephrasings - noise that is added in the *input* space rather than embedding space - which we've investigated thus far. We find that in this setting our theoretical assumptions do not hold - specifically, the noise in the embedding space that results from the rephrasing has non-zero mean and tends to lie entirely within the text subspace, with little to no variance in the direction of the gap (and see the supplementary). This is not surprising as models that are trained on large datasets are expected to be invariant to various input noises, such as text rephrasing.



(a) CLIP (ViT L-14) on CIFAR10 (b) CLIP (ViT B-16) on CIFAR100

Figure 10. CLIP variants are increasingly more robust to quantization as the gap is closed. The gap norm is measured in the non-quantized space.



(a) CLIP (ViT B-16) (b) SigLIP

Figure 11. We apply rephrasing noise to textual inputs on the A-OKVQA dataset, and measure average accuracy and rephrase accuracy in the noisy setting over $K = 500$ samples of wrong and correct answers for each question. Models are increasingly more accurate once the gap is closed, without change in clean accuracy, even though this setting is far from our theoretical assumptions.

We examine if our algorithm can still be used in such cases and find the the answer is indeed yes. For each question in the A-OKVQA dataset, we simulate a sample of random noise by sampling 3 random wrong answers and a random correct answer, exact details are provided in the supplementary. Since in this setting the answers are randomly selected, we focus on measuring accuracy under noise instead of robustness, while closing the gap using our algorithm. Results are presented in Fig. 11. As can be seen, under these conditions accuracy under rephrasings greatly increases when the gap is closed using our algorithm, while clean accuracy is unaffected.

6. Discussion

In this paper we've presented both theoretical and empirical results explaining the formation of a global gap in multimodal models, and the effect of it in the context of robustness. We prove that in this context, the gap is indeed a bug - as the modality gap enlarges, the ability of models to align similar texts and images diminishes when applying subtle variations to the embedding space.

Following this theoretical insight, we presented a simple and computationally inexpensive algorithm that under our as-

sumptions improves robustness with hardly any reduction in clean accuracy. Our algorithm could easily be implemented as a post processing step when using multi-modal model embeddings.

Finally, we began examining the effect of our theory and algorithm in the context of real noise, such as quantization and caption rephrasing. As this setting covers little of possible types of variations to the input and embedding space, we view our experiments as a hint to the potential of our algorithm. We therefore encourage further investigation as to settings of noise in which our simple algorithm could improve downstream performance.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2

[2] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving CLIP training with language rewrites. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3

[3] Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (CLIP). In *Proceedings of the 39th International Conference on Machine Learning*, pages 6216–6234. PMLR, 2022. 3

[4] Deepanway Ghosal, Navonil Majumder, Roy Lee, Rada Mihalcea, and Soujanya Poria. Language guided visual question answering: Elevate your multimodal language model using knowledge-enriched prompts. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12096–12102, Singapore, 2023. Association for Computational Linguistics. 8, 3

[5] Xiaoqing Ellen Tan Po-Yao Huang Russell Howes Vasu Sharma Shang-Wen Li Gargi Ghosh Luke Zettlemoyer Hu Xu, Saining Xie and Christoph Feichtenhofer. Demystifying clip data. 2023. 2

[6] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 7

[7] Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Understanding and constructing latent modality structures in multi-modal representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 7661–7671. IEEE, 2023. 1, 2

[8] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[9] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15190–15200, 2023. 3

[10] Hyunjae Kim, Seunghyun Yoon, Trung Bui, Handong Zhao, Quan Tran, Franck Dernoncourt, and Jaewoo Kang. Finetuning CLIP text encoders with two-step paraphrasing. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2175–2184, St. Julian's, Malta, 2024. Association for Computational Linguistics. 3

[11] Hyunjae Kim, Seunghyun Yoon, Trung Bui, Handong Zhao, Quan Hung Tran, Franck Dernoncourt, and Jaewoo Kang. Fine-tuning CLIP text encoders with two-step paraphrasing. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 2175–2184. Association for Computational Linguistics, 2024. 2

[12] Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, and Yejin Choi. Collective generation of natural image descriptions. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 1: Long Papers*, pages 359–368. The Association for Computer Linguistics, 2012. 6

[13] 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 *Advances in Neural Information Processing Systems*, 2022. 1, 2, 3, 5, 6, 7

[14] Christopher Liao, Christian So, Theodoros Tsiligkaridis, and Brian Kulis. Multimodal unsupervised domain generalization by retrieving across the modality gap. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2, 7

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 6

[16] D. Marco and D.L. Neuhoff. The validity of the additive noise model for uniform scalar quantizers. *IEEE Transactions on Information Theory*, 51(5):1739–1755, 2005. 8

[17] Yifei Ming and Yixuan Li. Understanding retrieval-augmented task adaptation for vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 35719–35743. PMLR, 2024. 2

[18] Marco Mistretta, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Andrew D. Bagdanov. Cross the gap: Exposing the intra-modal misalignment in CLIP via modality inversion. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 7

[19] 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. 3, 7

[20] Maitreya Patel, Changhoon Kim, Sheng Cheng, Chitta Baral, and Yezhou Yang. Eclipse: A resource-efficient text-to-image prior for image generations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9069–9078, 2024. 3, 7

[21] Jielin Qiu, Yi Zhu, Xingjian Shi, Zhiqiang Tang, Ding Zhao, Bo Li, and Mu Li. Benchmarking robustness under distribution shift of multimodal image-text models. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022. 3

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 4

[23] Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Ajanthan Thalaiyasingam. Accept the modality gap: An exploration in the hyperbolic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27263–27272, 2024. 1, 2

[24] 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. 1, 2, 3, 7

[25] Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3677–3685, 2023. 3

[26] 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 *Proceedings of the 41st International Conference on Machine Learning*, pages 43685–43704. PMLR, 2024. 3

[27] 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. 2, 6, 4

[28] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII*, pages 146–162. Springer, 2022. 8

[29] 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. 1, 2, 3, 5, 4

[30] Ofir Shifman and Yair Weiss. Lost in translation: Modern neural networks still struggle with small realistic image transformations. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXIX*, pages 231–247. Springer, 2024. 2, 3

[31] Yang Shu, Xingzhuo Guo, Jialong Wu, Ximei Wang, Jianmin Wang, and Mingsheng Long. CLIPood: Generalizing CLIP to out-of-distributions. In *Proceedings of the 40th International Conference on Machine Learning*, pages 31716–31731. PMLR, 2023. 3

[32] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016. 1, 2

[33] 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. 3

[34] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 1, 2

[35] Qizhou Wang, Yong Lin, Yongqiang Chen, Ludwig Schmidt, Bo Han, and Tong Zhang. Do CLIP models always generalize better than imagenet models? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3

[36] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020. 1, 2, 3

[37] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7949–7961, 2022. 3

[38] Can Yaras, Siyi Chen, Peng Wang, and Qing Qu. Explaining and mitigating the modality gap in contrastive multimodal learning. In *The Second Conference on Parsimony and Learning (Proceedings Track)*, 2025. 2, 3

[39] Mert Yazan, Suzan Verberne, and Frederik Situmeang. The impact of quantization on retrieval-augmented generation: An analysis of small llms. In *Proceedings of the Workshop Information Retrieval's Role in RAG Systems (IR-RAG 2024) co-located with the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024), Washington DC, USA, 07 18, 2024*, pages 77–81. CEUR-WS.org, 2024. 8

[40] Jingwen Ye, Ruonan Yu, Songhua Liu, and Xinchao Wang. Mutual-modality adversarial attack with semantic perturbation. In *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*, pages 6657–6665. AAAI Press, 2024. 3

[41] 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*, pages 11941–11952. IEEE, 2023. 2, 6

[42] 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 *Proceedings of the 7th Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022. 1, 2

[43] 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. 2, 3, 5

[44] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

Is the Modality Gap a Bug or a Feature? A Robustness Perspective

Supplementary Material

7. Theorem Proofs

7.1. Proof of Theorem 3.1

Proof. As stated in the theorem, assume that at some iteration t of gradient descent Q^x, Q^y are doubly stochastic and that exists $\vec{v}$ in which both modalities have zero variance. Assume that Q^x, Q^y stay doubly stochastic throughout the training. We'll begin by showing that iteration $t+1$ maintains both assumptions as well. To do so, we just need to show that the gradient of the loss is zero in direction $\vec{v}$. This follows from Eq. (11):

$$\begin{aligned} v^T \frac{\partial L_x}{\partial y_j} &= -2v^T(x_j - y_j) + 2\sum_i Q^x(i,j)v^T(x_i - y_j) \\ &= -2(a-b) + 2\sum_i Q^x(i,j)(a-b) \\ &= 0 \end{aligned}$$

and from the symmetry of the loss this is true for all other derivatives. Therefore both modalities will continue having zero variance in direction $\vec{v}$.

To show that training will converge when global and local orthogonality hold, we'll change the coordinate system s.t.:

$$\vec{x_i} = (a, \tilde{x}_i)$$
$$\vec{y_i} = (b, \tilde{y}_i)$$

and minimize the loss w.r.t. $\tilde{x}_i, \tilde{y}_i$. From standard uniformity and alignment arguments, the loss will be minimal when $\tilde{x}_i = \tilde{y}_i$ and the points are uniformly distributed. Note that at such a solution all the local gap vectors will be of the form

$$\forall i : \vec{g}_i = \vec{x}_i - \vec{y}_i = (a-b, 0, 0, ...) \tag{17}$$

meaning that by definition the global gap vector will also be equal to the above (as it is the mean). Orthogonality will also hold since the gap is solely in the direction of $\vec{v}$ which is orthogonal to both modalities. □

7.2. Proof of Theorem 3.3

Lemma 7.1. *Under the orthogonality assumption, for every point $y \in \mathcal{Y}$ and $\bar{x}$ the mean of modality $\mathcal{X}$:*

$$\|\bar{x} - (y-g)\| < \|\bar{x} - y\| \tag{18}$$

with $\vec{g} = \bar{y} - \bar{x}$ the global gap vector.

Proof. Assume coordinate frame s.t. $\bar{x} = 0$. Therefore the claim reduces to:

$$\|y - g\| < \|y\| \tag{19}$$

Under the orthogonality assumption, the points $y, g, \bar{x}$ form a right angled triangle with $\measuredangle y, g, \bar{x} = \pi/2$. Therefore, from the Pythagorean theorem $\|y\|^2 = \|g\|^2 + \|y-g\|^2$. Since $\|g\|^2 > 0$ then:

$$\|y\|^2 > \|y - g\|^2 \tag{20}$$

as required. □

Now to prove the theorem:

Proof. The separating hyperplane between x_1 and x_2 is characterized by the normal to the plane $w = \frac{x_1 - x_2}{\|x_1 - x_2\|}$. Denote $\tilde{w}$ the normal to the separating hyperplane between the noisy versions X i.e $\tilde{x}_1$ and $\tilde{x}_2$.

Since the nearest neighbor of y is x_1 then $w^T y > 0$. We wish to show that the probability of $\tilde{w}^T(y+v) > 0$ is greater than the probability that $\tilde{w}^T y > 0$.

Since by our assumptions the noise only rotates the hyperplane, then $\tilde{w} = R(\theta)w$ with $R(\theta)$ some rotation matrix.

Therefore:

$$P(\tilde{w}^T y > 0) = P(w^T(R(-\theta)y) > 0) \tag{21}$$

Notice that $w^T(R(-\theta)y) > 0$ only if $\frac{\pi}{2} - \cos^{-1}(\frac{w^T y}{\|y\|}) > \theta$ where $\cos^{-1}(\frac{w^T y}{\|y\|})$ is the angle between w and y.

From Theorem 7.1, $y - g$ has smaller norm than y and due to the orthogonality assumption, $g^T w = 0$. Therefore:

$$\frac{w^T(y+g)}{\|y+g\|} = \frac{w^T y}{\|y+g\|} > \frac{w^T y}{\|y\|}$$
$$\Rightarrow \cos^{-1}(\frac{w^T(y+g)}{\|y+g\|}) < \cos^{-1}(\frac{w^T y}{\|y\|}) \tag{22}$$
$$\Rightarrow \frac{\pi}{2} - \cos^{-1}(\frac{w^T(y+g)}{\|y+g\|}) > \frac{\pi}{2} - \cos^{-1}(\frac{w^T y}{\|y\|}) \tag{23}$$

Therefore the event that $w^T y > \theta$ is a strict subset of the event that $w^T(y+g) > \theta$, meaning that:

$$P(w^T(y+g) > \theta) > P(w^T(y+g) > \theta) \tag{24}$$

□

Another way to see this is to note that we can write $\tilde{w} = w + \eta$ with η a zero mean r.v. with covariance $2\sigma^2 I$. The retrieval is robust if $\tilde{w}^T y > 0$. Substituting:

$$\tilde{w}^T y = (w+\eta)^T y = w^T y + \eta^T y \tag{25}$$

From our assumption, $w^T y > 0$, so robustness will be maintained if $w^T y > -\eta^T y$. From our assumptions:

$$\text{Var}(\eta^T y) = 2\sigma^2 \|y\|^2 \tag{26}$$

Again, according to our assumptions, replacing $y \to y + g$ maintains:

$$w^T y = w^T (y + g) \tag{27}$$

since g is orthogonal to w, and from Theorem 7.1:

$$\|y\| > \|y + g\| \tag{28}$$

Therefore:

$$\begin{aligned} \text{Var}(\eta^T (y + g)) &= 2\sigma^2 \|y + g\|^2 < 2\sigma^2 \|y\|^2 \\ &= \text{Var}(\eta^T y) \end{aligned} \tag{29}$$

meaning that:

$$P(w^T (y + g) < -\eta^T (y + g)) < P(w^T y < -\eta^T y) \tag{30}$$

Since we decreased the variance of a zero mean r.v..

7.3. Extensions to Theorem 3.3

For completeness, we provide two extensions for the proof, one in the case of perturbation with non-zero mean, and another in the case of a general covariance structure.

7.3.1. Extension to Non-Zero Mean

Suppose that the noise can change the mean of the perturbed points. Following all notations of proof 7.2, the requirement for robustness is that:

$$\tilde{w}^T y > b \tag{31}$$

where $2b = \tilde{x}_1^T \tilde{x}_1 - \tilde{x}_2^T \tilde{x}_2$, so robustness is maintained if:

$$-\eta^T y \leq w^T y - b \tag{32}$$

If we assume that $w^T y > b$, i.e. that the margin between y and the decision boundary between x_1, x_2 is at least b, then the same proof from proof 7.2 holds.

7.3.2. Extension to General Covariance

Suppose that the noise has covariance C. Then:

$$\text{Var}(\eta^T y) = y^T (2C) y \tag{33}$$

In this case we would like to chose v s.t. it is orthogonal to x_1, x_2 (maintaining $w^T y = w^T (y + v)$), but also satisfies:

$$\begin{aligned} &(y + v)^T C (y + v) \leq y^T C y \\ &\Rightarrow 2v^T C y + v^T C v \leq 0 \end{aligned} \tag{34}$$

Any direction maintaining the above will increase robustness. Notice that this direction does not necessarily point from y to the origin (which in our case is the mean of modality $\mathcal{X}$), therefore the gap is not necessarily closed.

7.4. Proof of Theorem 3.4

Proof. Observe the relative distance between some $y \in Y$ and two points in the other modality when translating modality $\mathcal{X}$ along the gap by $\alpha \cdot v$:

$$\begin{aligned} &\forall x_i, x_j \in \mathcal{X} : \\ &\|y - (x_i + \alpha \cdot v)\|^2 - \|y - (x_j + \alpha \cdot v)\|^2 = \\ &\|x_i\|^2 - 2x_i^T y - 2x_i^T v - \|x_j\|^2 + 2x_j^T y + 2x_j^T v \end{aligned} \tag{35}$$

Since x_i, x_j are embedded on the unit sphere $\|x_i\| = \|x_j\| = 1$. Since y is also on the unit sphere, $-2x^T y = \|x - y\|^2 - 2$. Additionally, under the orthogonality assumption $\forall x \in \mathcal{X} : x^T v = c$ for some constant $c \in \mathbb{R}$. Substituting into equation 7.4:

$$\begin{aligned} &\|y - (x_i + \alpha \cdot v)\|^2 - \|y - (x_j + \alpha \cdot v)\|^2 = \\ &1 + \|x_i - y\|^2 - 2c - 1 + 2c - \|x_j - y\|^2 = \\ &\|x_i - y\|^2 - \|x_j - y\|^2 \end{aligned} \tag{36}$$

Therefore:

$$\begin{aligned} &\|y - x_i\|^2 < \|y - x_j\|^2 \Rightarrow \\ &\|y - (x_i + \alpha \cdot v)\|^2 < \|y - (x_j + \alpha \cdot v)\|^2 \end{aligned} \tag{37}$$

Meaning the nearest neighbor structure is maintained after translating one modality along the gap. □

8. Robustness to Various Noise Distributions

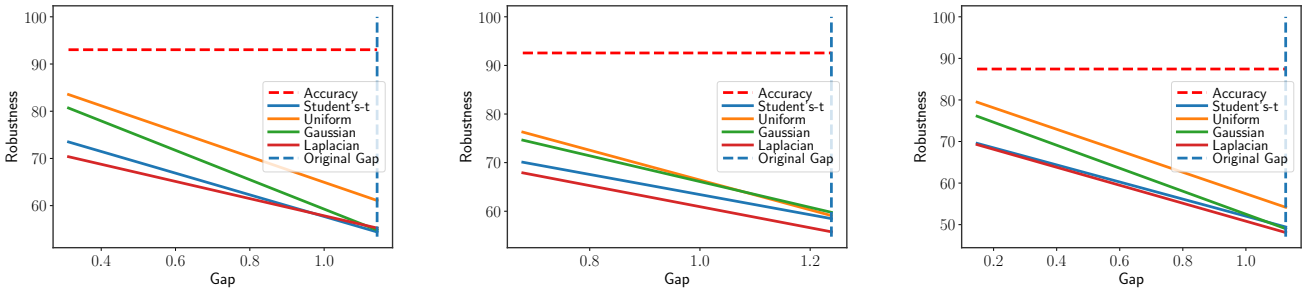
Here we expand on the results presented in Sec. 5.1. As Theorem 3.3 states, robustness should increase for any distribution of noise which has no correlation between the different dimensions and zero mean, not necessarily Gaussian noise. Figure 12 expands the experiments in Fig. 9 to non-Gaussian distributions with zero mean and where all dimensions are sampled i.i.d. . As can be seen, as long as the noise's dimensions are uncorrelated, robustness increases when the gap is closed.

9. Rephrasing Experiments - Further Details

This section provides details on the experiments presented in Sec. 5.3.

9.1. Rephrasing Noise is Correlated

Caption rephrasing, such as replacing the caption "a photo of a X" to "an image of X", tends to result in correlated noise in the embedding space. Intuitively, this is because in the input space, rephrasing can be seen as subtraction of a constant input - the string "a photo of a" followed by addition of the constant string "an image of". When these changes are applied to all inputs, it can be expected that the change in



(a) CLIP ViT-L-14 (b) SigLIP (c) CLIP ViT-B-16

Figure 12. Results for different noise distributions and models on CIFAR10. All are normalized to have variance $\sigma^2 = 0.025^2$. Robustness increases when the gap is closed.

the embedding space would either not exist (if the model is completely robust to such changes) or be consistent.

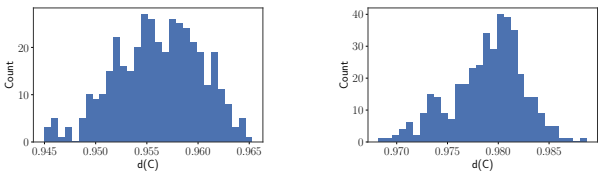
To show this, we conduct the above experiment on Imagenet. Assume all N class names (texts) are embedded with the prefix "a photo of a", resulting in the text embedding matrix X. We create a noise text embedding matrix, $\tilde{X}$ by embedding all class names with a different prefix (e.g. "an image of"). We repeat for 400 different caption templates, and calculate the empirical noise covariance. Define the noise to be $M = \tilde{X} - X$. The covariance C is then:

$$C = (M - \frac{1}{N}\vec{1}\vec{1}^T M)^T (M - \frac{1}{N}\vec{1}\vec{1}^T M) \tag{38}$$

To measure how much the noise is correlated, we simply measure the norm of all off-diagonal elements relative to the forbenius norm of the covariance matrix:

$$d(C) = \frac{\|C - \text{diag}(C)\|_F}{\|C\|_F} \tag{39}$$

This is of course equal to zero in the case that the noise is uncorrelated and $d(C) = 1$ when completely correlated. Fig. 13 measures $d(C)$ for different caption rephrasings for different models on Imagenet. As can be seen, in all cases $d(C) \approx 1$ meaning the noise is indeed extremely correlated.



(a) CLIP (ViT B-16) on Imagenet (b) SigLIP on Imagenet

Figure 13. We compute $d(C)$ according to Eq. (39). For all models we tested, with over 400 different captions, $d(C) \approx 1$ suggesting the noise is extremely correlated in the embedding space.

9.2. Rephrasing A-OKVQA

In the multiple choice VQA setting, we follow the protocol of Ghosal et al. [4]: given a question Q, for each possible answer A_i for $i \in [N_A]$ we embed the text that is the concatenation of $Q + A_i$ resulting in N_A text embeddings. We chose the estimated answer via cosine similarity with the image embedding. We evaluate on the entire A-OKVQA validation set.

In order to rephrase a correct answer, we simply swap the correct answer A_j with each of the "direct answers" options for that particular question in the dataset. Each question has up to 10 different possible correct answers, each one constitutes as a rephrasing. To replace the wrong answers we sample from the entire answer bank (not including the correct answers).

Notice that in this case our measure of robustness isn't meaningful as the "noisy" wrong answers are completely different texts, therefore we shouldn't expect the model to consistently predict the same wrong answer when adding this type of noise. Therefore, under this setting we focus on consistency w.r.t. right answers, which is similar to measuring accuracy when adding noise. If accuracy increases when closing the gap, the model is more consistent on predicting the right answer, or in other words - robust w.r.t. that answer.

10. Approximately Orthogonal Algorithm

As Theorem 3.3 proves, improvement of robustness is correlated with the amount of the gap that is closed, and is maximized when the two modality means overlap. In practice, the gap vector g can almost be non-orthogonal to the subspace of $\mathcal{X}$, meaning that closing the gap in the direction of $g' = g - VV^T g$ will produce a small increase in robustness since $\|g'\| \ll \|g\|$.

To solve this we rely on Theorem 3.3 that states that any direction which decreases the distance will increase robustness. We propose closing the gap in directions which are decreasingly orthogonal to the affine subspace of the

modality being moved. Following from Theorem 3.4, this procedure assures us that closing of the gap would result in minimal change to the clean nearest neighbor retrieval. We implement this idea by simply thresholding the number of components to which the gap vector is orthogonal, starting with those containing minimal variance. Algorithm 1 presents a simple python pseudocode implementation.

Algorithm 1: Approximately Orthogonal Gap Vector

```
// ϵ - variance threshold, g⃗ - gap vector, X - modality to move
VSV^T ← PCA(X)
inds ← {i : S_i > ϵ}
V' ← V[:, inds]
return (I − V'V'^T)g
```

This algorithm produces a tradeoff between the increment in robustness and loss of accuracy controlled by the parameter ϵ. An example of using the algorithm is displayed in Fig. 15 and the tradeoff induces can be seen in Fig. 16.

11. Simulations

11.1. Details

All simulations are done by optimizing randomly initialized embedding vectors using the Eq. (8). We use full batch gradient descent with a learning rate of 0.01.

In Fig. 4 (top) we train unnormalized embedding vectors, directly optimizing Eq. (8) without a normalization step. We initialize the embeddings sampled from two normal distributions $\mathcal{N}(\mu_{1,2}, 0.01^2 \cdot I)$ with $\mu_{1,2} = (0, \pm 0.5)$. We use a constant temperature of $\tau = 0.1$ and train for 10^7 iterations.

In Fig. 4 (bottom) and Fig. 5 we initialize 1000 embeddings per modality drawn from a normal distribution $\mathcal{N}(\pm\vec{e}_1, 0.01^2 \cdot I)$ with $\vec{e}_1$ being the first elementary basis vector. All embeddings are normalized as is done in training multi-modal contrastive models [22].

11.2. Further Experiments

We ablate both the dimension, initial distance between modality means and temperature. As stated in the main paper (and shown in [29] and [27]), there exist cases in which training converges to completely aligned modalities, without a gap such as when the modalities don't have a gap to begin with or in certain temperatures. We demonstrate this in Fig. 14.

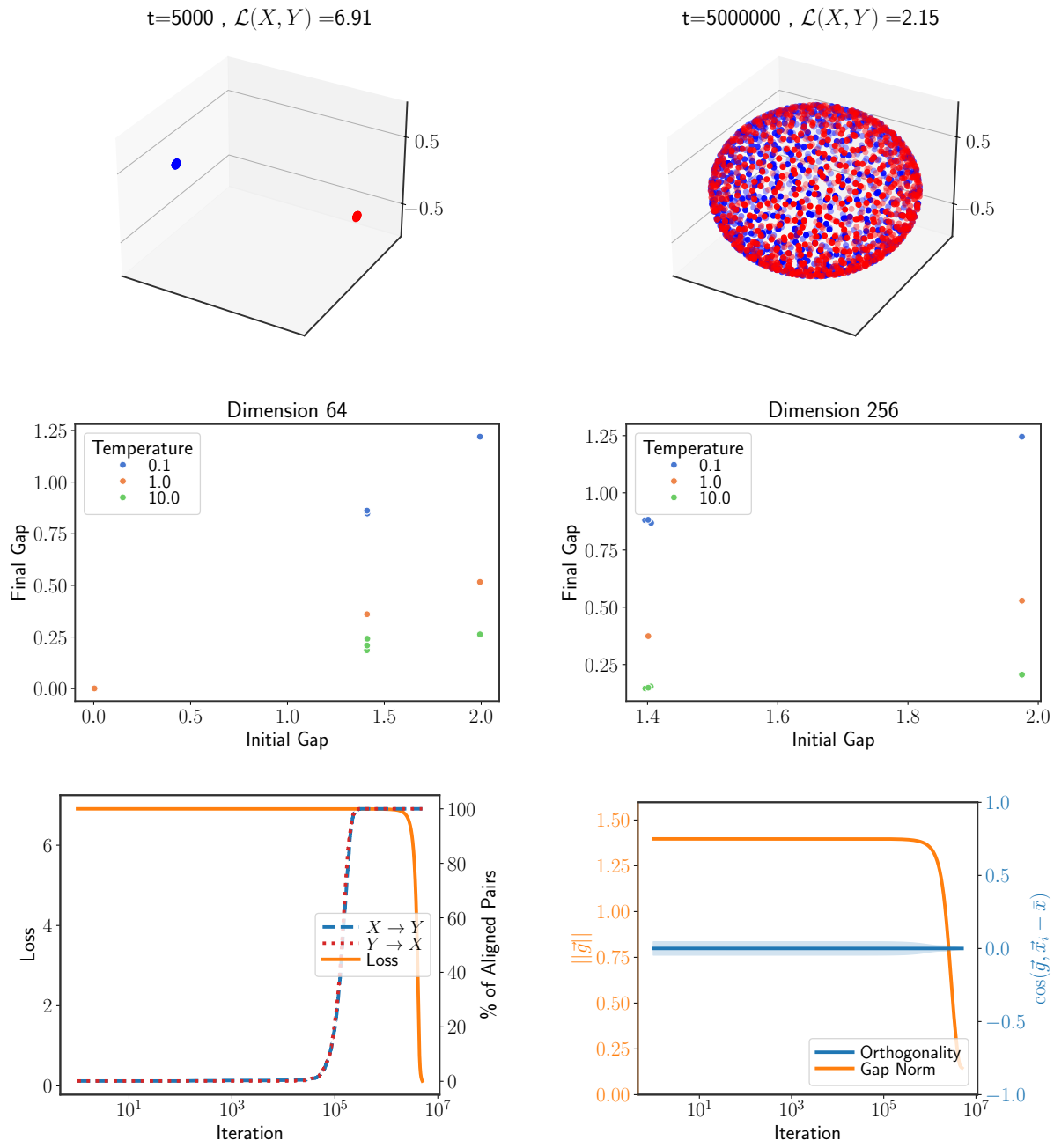
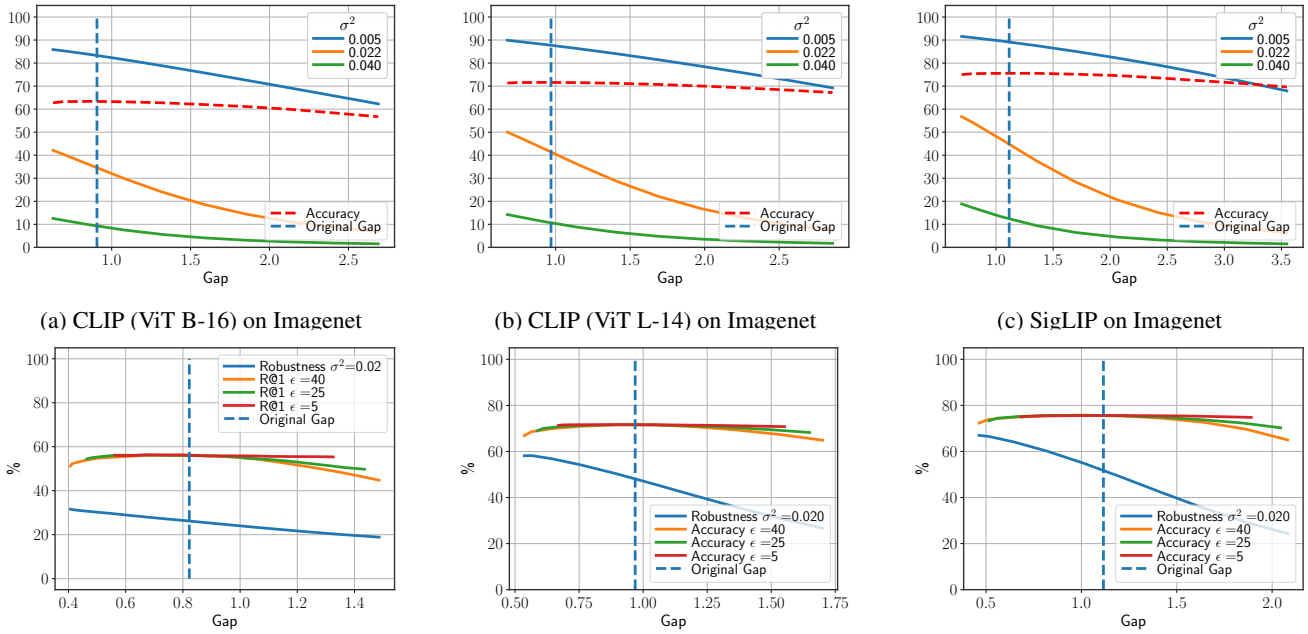


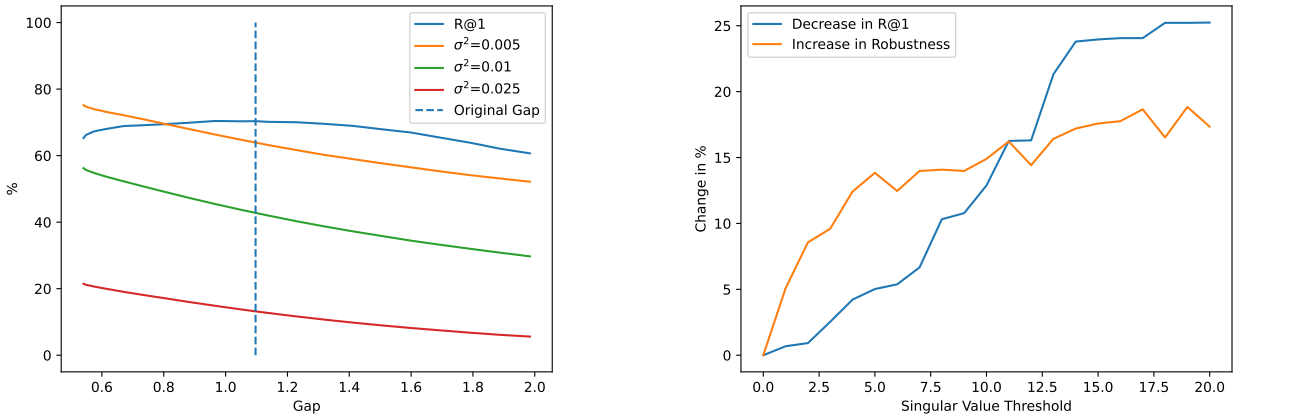
Figure 14. Top: When training with $\tau = 1$ training converges to a solution without a gap, despite existence of an initial gap and orthogonality. Middle: This is consistent for training in higher dimensions as well - different temperatures have different effects on how much of the gap is closed. When temperatures are ≥ 1, the gap closes throughout training. In higher temperatures training hardly differs from initialization. When the gap is initialized at zero it remains so for all temperatures. Bottom: Example of dynamics in $\mathbb{R}^{256}$ with $\tau = 10$. The gap closes throughout training as it converges to perfect alignment.



(a) CLIP (ViT B-16) on Imagenet (b) CLIP (ViT L-14) on Imagenet (c) SigLIP on Imagenet

(d) CLIP (ViT L-14), MS-COCO (e) CLIP (ViT L-14), ImageNet (f) SigLIP, ImageNet

Figure 15. The zero shot classification accuracy and robustness under noise $\eta \sim \mathcal{N}(0, \sigma^2 I)$, on ImageNet and MS-COCO.
Top row: When using Algorithm 1 with $\epsilon = 5\%$ of the variance, decrease in accuracy is negligible ($< 1\%$) relative to the robustness gained, which can be $\sim 10\%$.
Bottom row: As the threshold ϵ grows larger, more of the gap is closed, greatly increasing robustness at a negligible cost of accuracy.



(a) SigLIP on MS-COCO (b) Tradeoff for $\sigma^2 = 0.01$

Figure 16. Even when using Algorithm 1 the drop in R@1 for SigLIP [41] on image to text retrieval on MS-COCO dataset [15] is negligible relative to the improvement in robustness for different Gaussian noises (left). Fig. 16b shows the ranges of the singular value threshold ϵ for which the increment in robustness (for Gaussian noise with $\sigma^2 = 0.01$) is larger than the decrease in R@1.