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

Supplementary Material

A. Theorem Proofs

A.1. Proof of Theorem 3.1

Proof. Calculating the gradient of the loss using the notation in Eq. (6):

$$\nabla_{y_i}\mathcal{L} = \frac{2}{\tau}[2(x_i - y_i) - \sum_{k=1}^{N}(Q^x(k,i) + Q^y(k,i))(x_k - y_i)] \tag{13}$$

Defining $S_i = \sum_k Q^x(k,i)$ and $\tilde{x}_i$:

$$\tilde{x}_i = \frac{\sum_k (Q^x(k,i) + Q^y(k,i))x_k}{S_i + 1} \tag{14}$$

Inputting into Eq. (13):

$$\nabla_{y_i}\mathcal{L} = \frac{2}{\tau}\left[2(x_i - y_i) - (S_i + 1)(\mu_x - y_i)\right] \tag{15}$$

Assume $\forall j : \|x_j - \mu_x\| \leq \epsilon$ and $\forall i : \|y_i - \mu_y\| \leq \epsilon$. Substituting into Eq. (15):

$$\begin{aligned}\nabla_{y_i}\mathcal{L} &= \frac{2}{\tau}[2(x_i - \mu_x + \mu_x - y_i) \\ &\quad - (S_i + 1)(\tilde{x}_i - \mu_x + \mu_x - y_i)] \\ &= \frac{2}{\tau}\left[2(x_i - \mu_x) + 2(\mu_x - y_i)\right. \\ &\quad \left. - (S_i + 1)(\tilde{x}_i - \mu_x) - (S_i + 1)(\mu_x - y_i)\right] \\ &= \frac{2}{\tau}\left[(1 - S_i)(\mu_x - y_i) + O(\epsilon)\right]\end{aligned} \tag{16}$$

From our assumption $y_i = \mu_y + O(\epsilon)$ and that $\|\mu_x - \mu_y\| \gg \epsilon$:

$$\nabla_{y_i}\mathcal{L} = \frac{2}{\tau}\left[(1 - S_i)(\mu_x - \mu_y) + O(\epsilon)\right] \tag{17}$$

Meaning that the gradient of y_i is approximately along the direction of the gap $\vec{g} = \mu_x - \mu_y$.

The last thing left to prove is that the specific direction along the gap shrinks the variance in the gap. We'll show that:

$$S_i \approx 1 + \frac{2}{\tau}(\mu_x - \mu_y)\cdot(y_i - \mu_y) \tag{18}$$

Meaning that each y_i moves towards μ_y along the gap, thus shrinking the variance.

Let $k(x,y) = e^{-\|x-y\|^2/\tau}$ be the Gaussian kernel. Under our assumption, for $\epsilon \to 0$ we get that $x_i \to \mu_x$ and $y_j \to \mu_y$. Thus we can use a (linear) Taylor expansion of the kernel around the values of the means:

$$k(x_k, y_i) \approx k(x_k, \mu_y) + \nabla_y k(x_k, y)\big|_{y=\mu_y}\cdot(y_i - \mu_y) \tag{19}$$

The gradient of the kernel is $\nabla_y k(x,y) = \frac{2}{\tau}(x-y)k(x,y)$. Substituting this into the expansion:

$$k(x_k, y_i) \approx k(x_k, \mu_y)\left[1 + \frac{2}{\tau}(x_k - \mu_y)\cdot(y_i - \mu_y)\right] \tag{20}$$

Substituting the linearized kernel into the expression for $Q^x(k,i)$, the common factor $k(x_k, \mu_y)$ cancels from the numerator and denominator:

$$Q^x(k,i) \approx \frac{1 + \frac{2}{\tau}(x_k - \mu_y)\cdot(y_i - \mu_y)}{\sum_{j=1}^{N}\left[1 + \frac{2}{\tau}(x_k - \mu_y)\cdot(y_j - \mu_y)\right]} \tag{21}$$

The denominator sum simplifies significantly:

$$\begin{aligned}&\sum_{j=1}^{N}\left[1 + \frac{2}{\tau}(x_k - \mu_y)\cdot(y_j - \mu_y)\right] \\ &= N + \frac{2}{\tau}(x_k - \mu_y)\cdot\sum_{j=1}^{N}(y_j - \mu_y)\end{aligned} \tag{22}$$

Under the assumption that the cluster $\mathcal{Y}$ is centered at its mean ($\sum(y_j - \mu_y) = 0$), the denominator becomes exactly N.

Summing $Q^x(k,i)$ over all k:

$$S_i \approx \frac{1}{N}\sum_{k=1}^{N}\left[1 + \frac{2}{\tau}(x_k - \mu_y)\cdot(y_i - \mu_y)\right] \tag{23}$$

Applying the centering identity for $\mathcal{X}$ ($\sum x_k = N\mu_x$):

$$S_i \approx 1 + \frac{2}{\tau}(\mu_x - \mu_y)\cdot(y_i - \mu_y) \tag{24}$$

□

A.2. Proof of Theorem 3.2

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 the zero variance in direction $\vec{v}$: By the assumption that $v^T x_i = a$ for all x_i and $v^T y_j = b$ for all y_j we see that using Eq. (7):

$$\begin{aligned}v^T\frac{\partial\mathcal{L}}{\partial y_i} &\propto v^T(x_i - y_i) \\ &\quad - v^T\sum_{k=1}^{N}\frac{Q^x(k,i) + Q^y(k,i)}{2}(x_k - y_i) \\ &= 0\end{aligned}$$

where the last equality follows from double stochasticity.

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

$$\forall i : \vec{g}_i = \vec{x}_i - \vec{y}_i = (a - b, 0, 0, ...) \tag{25}$$

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. □

A.3. Proof of Theorem 3.4

Lemma A.1. *Under the orthogonality assumption, for every point $y \in \mathcal{Y}$ and μ_x the mean of modality $\mathcal{X}$:*

$$\|\mu_x - (y - g)\| < \|\mu_x - y\| \tag{26}$$

with $\vec{g} = \mu_y - \mu_x$ the global gap vector.

Proof. Assume coordinate frame s.t. $\mu_x = 0$. Therefore the claim reduces to:

$$\|y - g\| < \|y\| \tag{27}$$

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

$$\|y\|^2 > \|y - g\|^2 \tag{28}$$

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:

$$P(\tilde{w}^T y > 0) = P(w^T(R(-\theta)y) > 0) \tag{29}$$

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 A.1, $y - g$ has smaller norm than y and due to the orthogonality assumption, $g^T w = 0$. Therefore:

$$\frac{w^T(y+g)}{\|y+g\|} = \frac{w^T y}{\|y+g\|} > \frac{w^T y}{\|y\|}$$
$$\Rightarrow \cos^{-1}(\frac{w^T(y+g)}{\|y+g\|}) < \cos^{-1}(\frac{w^T y}{\|y\|}) \tag{30}$$
$$\Rightarrow \frac{\pi}{2} - \cos^{-1}(\frac{w^T(y+g)}{\|y+g\|}) > \frac{\pi}{2} - \cos^{-1}(\frac{w^T y}{\|y\|}) \tag{31}$$

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

$$P\left(w^T(y+g) > \theta\right) > P\left(w^T(y+g) > \theta\right) \tag{32}$$

□

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:

$$\tilde{w}^T y = (w + \eta)^T y = w^T y + \eta^T y \tag{33}$$

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

$$\text{Var}(\eta^T y) = 2\sigma^2 \|y\|^2 \tag{34}$$

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

$$w^T y = w^T(y+g) \tag{35}$$

since g is orthogonal to w, and from Theorem A.1:

$$\|y\| > \|y + g\| \tag{36}$$

Therefore:

$$\text{Var}(\eta^T(y+g)) = 2\sigma^2 \|y+g\|^2 < 2\sigma^2 \|y\|^2$$
$$= \text{Var}(\eta^T y) \tag{37}$$

meaning that:

$$P\left(w^T(y+g) < -\eta^T(y+g)\right) < P\left(w^T y < -\eta^T y\right) \tag{38}$$

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

A.4. Extensions to Theorem 3.4

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.

A.4.1. Extension to Non-Zero Mean

Suppose that the noise can change the mean of the perturbed points. Following all notations of proof A.3, the requirement for robustness is that:

$$\tilde{w}^T y > b \tag{39}$$

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

$$-\eta^T y \leq w^T y - b \tag{40}$$

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 A.3 holds.

A.4.2. Extension to General Covariance

Suppose that the noise has covariance C. Then:

$$\text{Var}(\eta^T y) = y^T (2C) y \tag{41}$$

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:

$$\begin{aligned} &(y+v)^T C (y+v) \leq y^T C y \\ &\Rightarrow 2v^T C y + v^T C v \leq 0 \end{aligned} \tag{42}$$

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.

A.5. Proof of Theorem 3.5

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 $\alpha \cdot v$:

$$\begin{aligned} &\forall x_i, x_j \in \mathcal{X}: \\ &\|y - (x_i + \alpha \cdot v)\|^2 - \|y - (x_j + \alpha \cdot v)\|^2 = \\ &\|x_i\|^2 - 2x_i^T y - 2x_i^T v - \|x_j\|^2 + 2x_j^T y + 2x_j^T v \end{aligned} \tag{43}$$

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 43:

$$\begin{aligned} &\|y - (x_i + \alpha \cdot v)\|^2 - \|y - (x_j + \alpha \cdot v)\|^2 = \\ &1 + \|x_i - y\|^2 - 2c - 1 + 2c - \|x_j - y\|^2 = \\ &\|x_i - y\|^2 - \|x_j - y\|^2 \end{aligned} \tag{44}$$

Therefore:

$$\begin{aligned} &\|y - x_i\|^2 < \|y - x_j\|^2 \Rightarrow \\ &\|y - (x_i + \alpha \cdot v)\|^2 < \|y - (x_j + \alpha \cdot v)\|^2 \end{aligned} \tag{45}$$

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

B. Does Information Imbalance Cause the Gap?

It has previously been postulated that information imbalance between texts and images, i.e. when texts are much less informative that their image counterparts, plays a role in the creation of the modality gap [26].

We observe a simplified setting in which a training set for a model trained with the contrastive loss contains only two images x_1, x_2 with the same caption y. We'll show that the minima of these dynamics isn't when a gap exists but when all points lie on top one another.

The loss is therefore:

$$\begin{aligned} \mathcal{L}(x_1, x_2, y) = &-\frac{1}{2}(\log \frac{e^{-d(x_1,y)}}{e^{-d(x_1,y)} + e^{-d(x_1,y)}} \\ &+ \log \frac{e^{-d(x_2,y)}}{e^{-d(x_2,y)} + e^{-d(x_2,y)}} \\ &+ \log \frac{e^{-d(x_1,y)}}{e^{-d(x_1,y)} + e^{-d(x_2,y)}} \\ &+ \log \frac{e^{-d(x_2,y)}}{e^{-d(x_2,y)} + e^{-d(x_1,y)}}) \\ = &-\log \frac{1}{2} + d(x_1, y) + d(x_2, y) \\ &+ \log(e^{-d(x_1,y)} + e^{-d(x_2,y)}) \end{aligned} \tag{46}$$

With $d = \ell_2$ distance. Taking the derivative:

$$\frac{d\mathcal{L}}{dx_i} = 2(y - x_i) + \frac{e^{-d(x_i,y)}}{e^{-d(x_1,y)} + e^{-d(x_2,y)}} \cdot 2(x_i - y) \tag{47}$$

Since $\frac{e^{-d(x_i,y)}}{e^{-d(x_1,y)} + e^{-d(x_2,y)}} < 1$ we get that the total gradient is $\frac{d\mathcal{L}}{dx_i} = c(y - x_i)$ for some $c > 0$. The same analysis can be done for $\frac{d\mathcal{L}}{dy}$ where we get:

$$\begin{aligned} \frac{d\mathcal{L}}{dy} &= 2(x_1 - y) + 2(x_2 - y) \\ &+ \frac{e^{-d(x_1,y)} \cdot 2(y - x_1) + e^{-d(x_2,y)} \cdot 2(y - x_2)}{e^{-d(x_1,y)} + e^{-d(x_2,y)}} \end{aligned} \tag{48}$$

Now assume by negation that the minima of these dynamics isn't when $x_1 = x_2 = y$. For x_i it is obvious from the gradient that the only minima is $x_i = y$. For y, it is enough to see that $y = \frac{x_1 + x_2}{2}$ is a minima by substituting into the gradient. Therefore contradicting that there is a different minima except $x_1 = x_2 = y$, as required.

C. Dimensionality Collapse and the Gap

As mentioned in the paper, an orthogonal gap can emerge even without dimensionality collapse, as in Fig. 4 where initialization has equal variance in all dimensions. This makes it an unnecessary condition for the formation of an orthogonal gap.

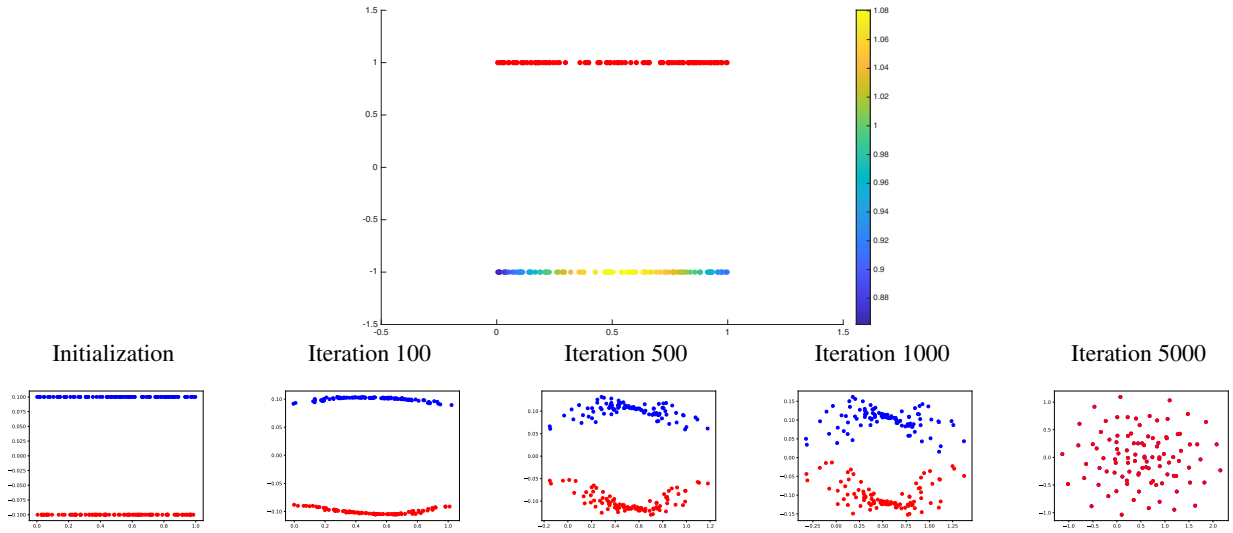


Figure 11. **(Top:** Two set of points that perfectly satisfy dimensionality collapse. The bottom points are color-coded by S_i^y. Note that $S_i > 1$ for points near the center. **(bottom:** The training dynamics. Despite no variance in the direction of the gap for both modalities, the solution converged to has no gap and is perfectly aligned. This is because in the initial iterations, points near the center are pushed away from the other modality hence destroying the original low dimensionality of each modality.

But it the condition sufficient? Figure 11 shows that the answer is no as well. We train again follow the simplified setting of experiment of Fig. 4, only initializing the two modalities with complete dimensional collapse - all variance lies in a single dimension with the other containing none. In this case, training converges to a fully aligned solution with no gap, making the condition of dimensionality collapse insufficient to explain the formation of the gap. The behavior is consistent with our analysis because in the initial iterations double stochasticity does not hold. The top of Fig. 11 color-codes the bottom points by S_i^y. Note that $S_i > 1$ for points near the center, so at the initial iterations, points near the center are pushed away from the other modality more than points at the edges hence destroying the original low dimensionality of each modality.

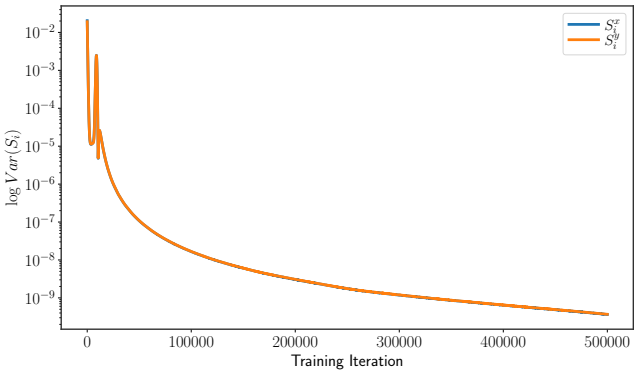


Figure 12. We calculate S_i^x, S_i^y throughout the training of $N = 500$ embedding pairs initialized from a Gaussian distribution with variance $\sigma^2 = 0.01$ and gap of $\|\vec{g}\| = 1.8$. As training progresses, the values of S_i^x, S_i^y are concentrated around their means, which by definition equal 1, making the matrices Q^x, Q^y more doubly-stochastic.

D. Empirical Evidence for Assumptions

D.1. Empirical Evidence for Apprximate Double Stochasticity

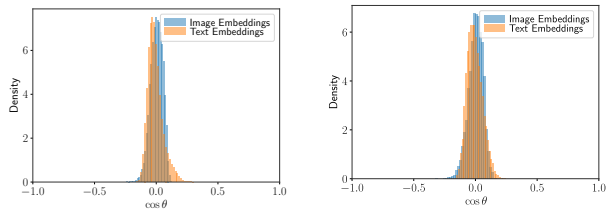
We repeat the experiment measuring S_j^x, S_j^y as in Fig. 6 only for embeddings trained on the unit hypersphere in $\mathbb{R}^{64}$. Results are shown in Fig. 12 - as training progresses, $S_j^x, S_j^y \to 1$ meaning that the matrices Q^x, Q^y become more doubly-stochastic.

D.2. Empirical Evidence for Orthogonality

For completeness we recompute results from [42] showcasing the orthogonality of the global gap vector for different models and datasets in Fig. 13.

E. Robustness to Various Noise Distributions

Here we expand on the results presented in Sec. 5.1. As Theorem 3.4 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 14 expands the experiments in Fig. 8 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.



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

Figure 13. 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 3.3.

F. Rephrasing Experiments - Further Details

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

F.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 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:

$$C = (M - \frac{1}{N}\vec{1}\vec{1}^T M)^T (M - \frac{1}{N}\vec{1}\vec{1}^T M) \tag{49}$$

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:

$$d(C) = \frac{\|C - \text{diag}(C)\|_F}{\|C\|_F} \tag{50}$$

This is of course equal to zero in the case that the noise is uncorrelated and $d(C) = 1$ when completely correlated. Fig. 15 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.

F.2. Rephrasing A-OKVQA

In the multiple choice VQA setting, we follow the protocol of Ghosal et al. [3]: 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.

G. Approximately Orthogonal Algorithm

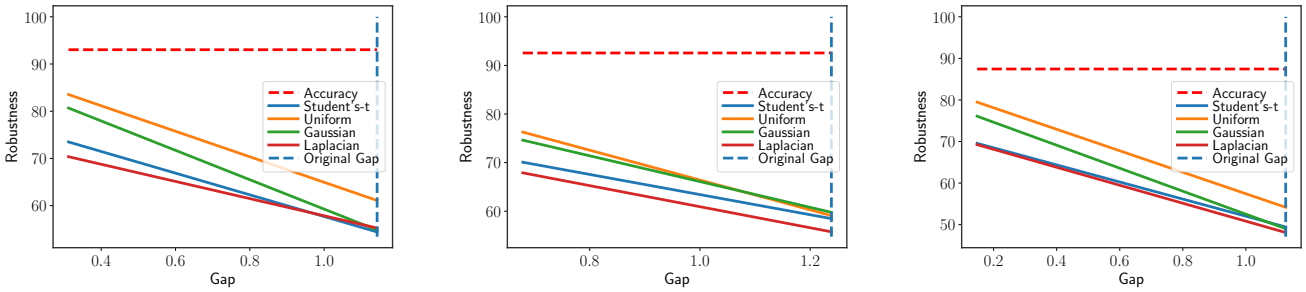
As Theorem 3.4 proves, improvement of robustness is correlated with the amount of the gap that is closed, and is maximized when the two modality means overlap. In practice, the gap vector g can almost be non-orthogonal to the subspace of $\mathcal{X}$, meaning that closing the gap in the direction of $g' = g - VV^T g$ will produce a small increase in robustness since $\|g'\| \ll \|g\|$.

To solve this we rely on Theorem 3.4 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 3.5, 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

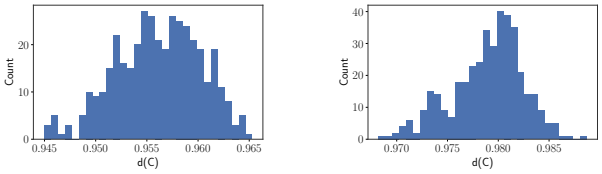
```
// ϵ - variance threshold, g⃗ - gap vector, X - modality to move
VSV^T ← PCA(X)
inds ← {i : S_i > ϵ}
V' ← V[:, inds]
return (I − V'V'^T)g
```

This algorithm produces a tradeoff between the increment



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

Figure 14. 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.



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

Figure 15. We compute $d(C)$ according to Eq. (50). For all models we tested, with over 400 different captions, $d(C) \approx 1$ suggesting the noise is extremely correlated in the embedding space.

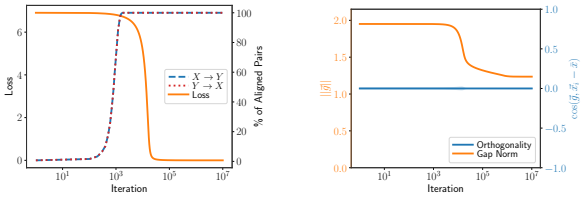


Figure 16. We follow Shi et al. [28] and learn embeddings $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{1000\times512}$ on the hypersphere using the contrastive loss (Eq. (6)) and gradient descent. While the loss decreases and training converges (left), a major gap remains between the embeddings and orthogonality holds (right, measured using Eq. (9)).

in robustness and loss of accuracy controlled by the parameter ϵ. An example of using the algorithm is displayed in Fig. 18 and the tradeoff induces can be seen in Fig. 19.

H. Simulations

H.1. Details

All simulations are done by optimizing randomly initialized embedding vectors using the Eq. (6). 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. (6) 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. 16 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 [21].

H.2. Further Experiments

We ablate both the dimension, initial distance between modality means and temperature. As stated in the main paper (and shown in [28] and [26]), there exist cases in which training converges to completely aligned modalities. This can happen when double stochasticity does not hold throughout the iterations, and double stochasticity may depend subtly on the temperature τ and the number of iterations. We demonstrate this in Fig. 17.

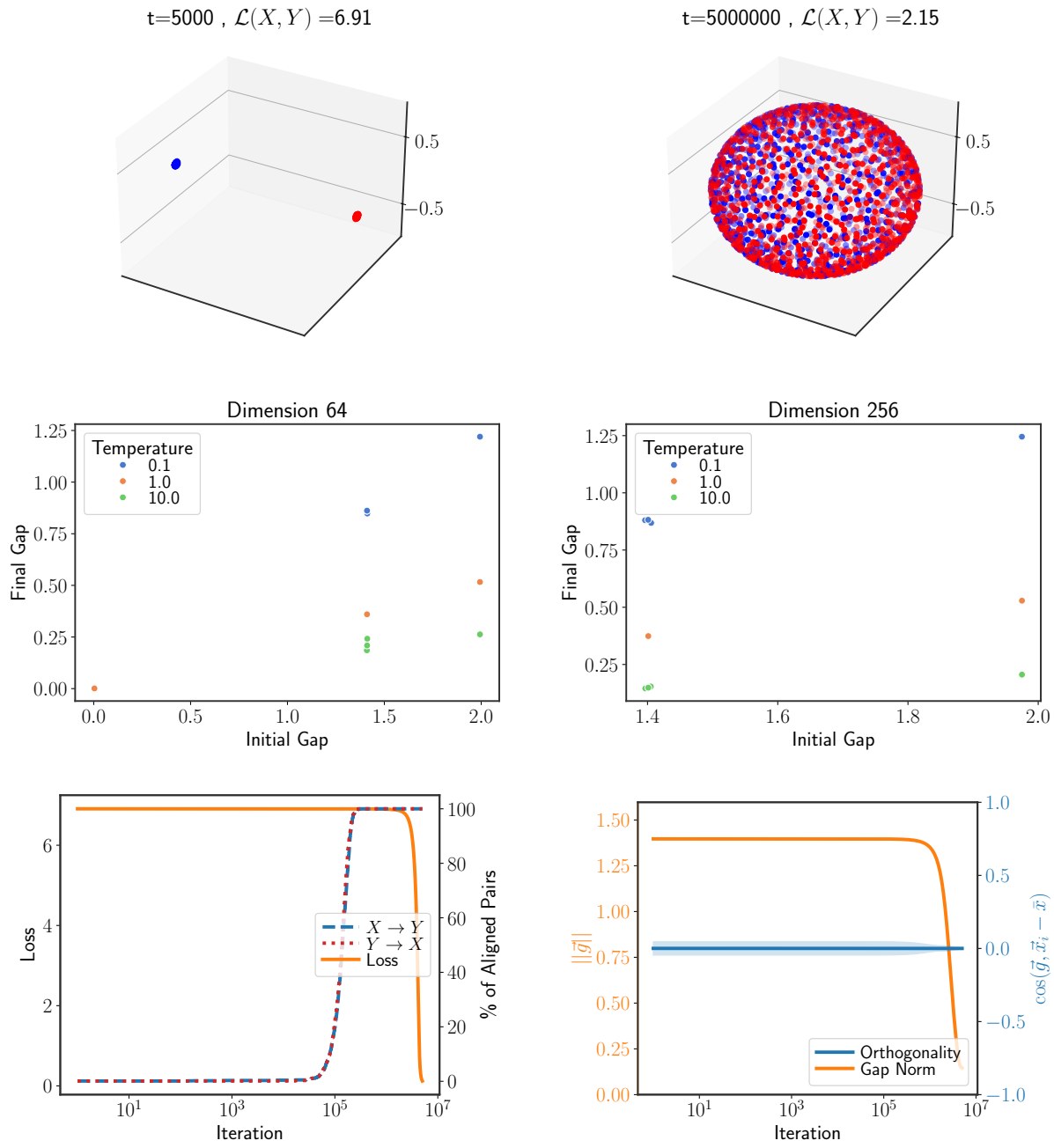
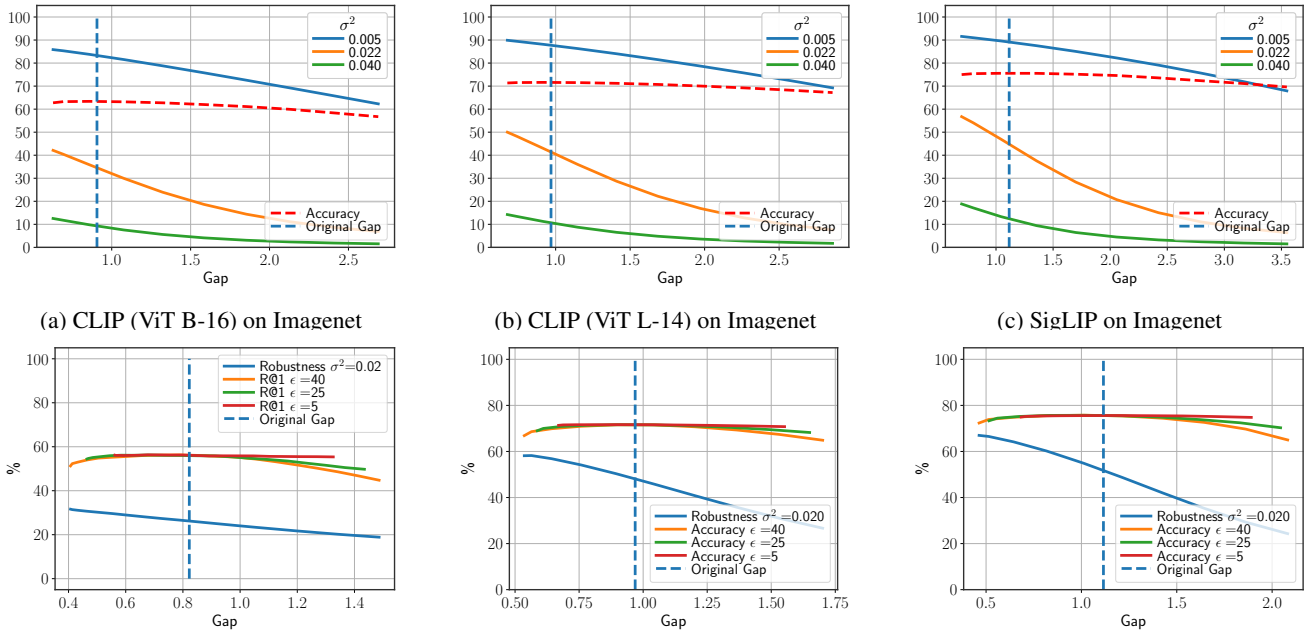


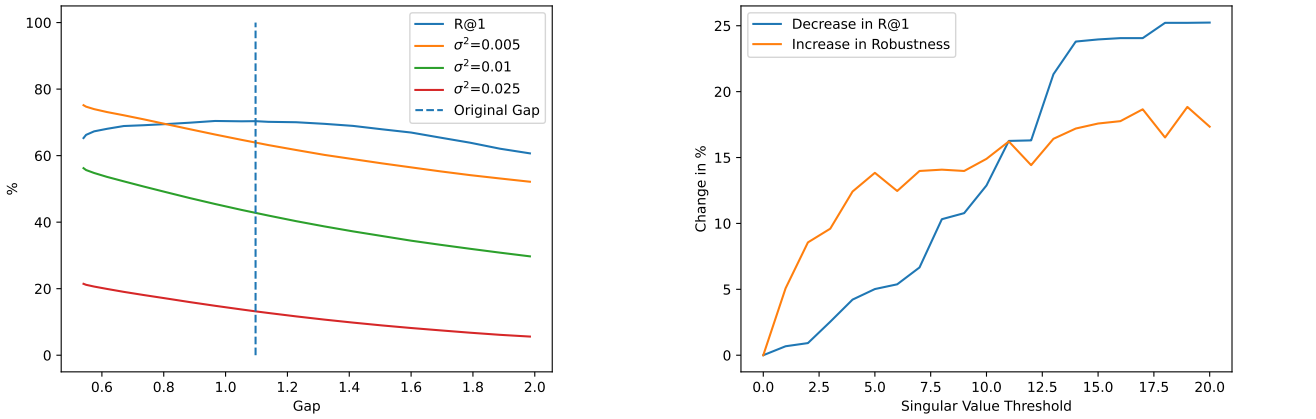
Figure 17. Top: When training with $\tau = 1$ training converges to a solution without a gap, despite existence of an initial gap and orthogonality. Middle: This is consistent for training in higher dimensions as well - different temperatures have different effects on how much of the gap is closed. When 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 18. 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 19. Even when using Algorithm 1 the drop in R@1 for SigLIP [40] on image to text retrieval on MS-COCO dataset [14] is negligible relative to the improvement in robustness for different Gaussian noises (left). Fig. 19b 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.