IT'S NOT A MODALITY GAP: CHARACTERIZING AND ADDRESSING THE CONTRASTIVE GAP

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning jointly from images and texts using contrastive pre-training has emerged as an effective method to train large-scale models with a strong grasp of semantic image concepts. For instance, CLIP (Radford et al., 2021), pre-trained on a large corpus of web data, excels in tasks like zero-shot image classification, object detection, geolocalization, and more. These contrastive models embed input images and texts into a shared representational space. Recently, it was claimed that models like CLIP show a *modality gap*, where image and text embeddings occupy disjoint areas in the representational space. Previous studies attribute this gap to factors like data artifacts (mismatched pairs), model architecture artifacts (the cone effect), and the nature of the loss landscape (getting stuck in local minima). We demonstrate that, even after accounting for these factors, and even when using the *same modality*, the contrastive loss actually *creates* a gap during training. As a result, we propose renaming this phenomenon the *contrastive gap*. We show that the contrastive gap is exacerbated by training with small batch sizes in high-dimensional spaces, causing embeddings of each modality to occupy small disjoint portions of the latent space. Our experiments show that minimizing the contrastive gap via the addition of uniformity and alignment terms optimizes the representational space and conveys better performance on downstream tasks such as zero-shot image classification and multi-modal arithmetic.

1 INTRODUCTION

Multi-modal models map inputs from different modalities into a unified representational space, such that semantically similar inputs from different modalities map to nearby points in the representational space. This can be practically useful because it allows transfer between modalities, but it is also more consistent with the human sensory experience, which collects information across many modalities to better understand the world. In the context of learning from paired images and text, CLIP (Contrastive Language Image Pre-training) (Radford et al., 2021) establishes a strong proof-of-concept for this reasoning. CLIP's multi-modal contrastive loss allows the model to predict the text associated with an image and vice versa. CLIP scales this approach to a very large dataset of 400M image-caption pairs, learning embeddings that cover a wide variety of visual concepts applicable to many downstream tasks.

But, while CLIP is powerful, it suffers from a *modality gap* (Liang et al., 2022), wherein image embeddings reside in a space disjoint from that of the text embeddings. This phenomenon is also seen in multi-modal contrastive models in other domains such as medical images (Zhang et al., 2021), videos (Xu et al., 2021), amino acid sequencing (`https://github.com/MicPie/clasp`) and brain decoding (Luo et al., 2024). Prior work has shown that performance on downstream tasks improves when we minimize this gap, which can be achieved by simply shifting one modality's embeddings in CLIP space to change the size of the gap (Liang et al., 2022) through projection (transforming the embeddings through some projection operation) (Zhou et al., 2023), or through fine-tuning (Oh et al., 2023). Recent work has additionally shown that learning latent spaces of visual embeddings via alignment with human similarity judgments improves downstream task performance (Muttenthaler et al., 2023). Similarly, work relating CLIP embeddings to human brain data found that representations that had reduced the gap between modalities also led to improved downstream model performance (Luo et al., 2024). Therefore, analyzing and closing this gap is a promising direction to improve upon the strong representational capacity of CLIP and its variants.

In this paper, we conducted a comprehensive study of what causes the *modality gap* and simple ways to close it. We make the following contributions:

- **It's not a modality gap.** After summarizing the common purported causes of the modality gap, we perform comprehensive experiments that show that accounting for these factors does *not* close the gap, suggesting that the present understanding of modality gap may be flawed.
- **It's a *contrastive* gap.** We present experiments that demonstrate that the gap is a byproduct of a high dimensional CLIP space, combined with contrastive loss that encourages CLIP embeddings to occupy a lower dimensional manifold relative to the latent space.
- **The contrastive gap can be closed.** We show that simply fine-tuning CLIP by adding a factor for uniformity and alignment can reduce the size of the gap by distributing the embeddings more uniformly throughout CLIP's latent space.
- **Closing the contrastive gap improves downstream performance.** Finally, we present experiments to show that closing the contrastive gap, and thereby creating more aligned and uniformly distributed representations, creates a representational space that is better for most downstream tasks, including zero-shot image classification and multi-modal embedding arithmetic.

2 Background

A multi-modal contrastive model learns a representational space in which semantically paired samples from different modalities (*positive pairs*, e.g., an image and its associated caption) are closer together in the latent space compared to other randomly chosen pairs (*negative pairs*, e.g., an image and the caption associated with another randomly chosen image). Contrastive learning was originally developed for the unimodal setting to learn self-supervised representations. One such popular model is SimCLR (Chen et al., 2020), which uses the NT-Xent (Normalized Temperature-scaled Cross entropy) loss to efficiently learn robust image representations. The NT-Xent loss is different from previous contrastive learning methods as it does not need explicit negative samples. The NT-Xent loss works by treating samples other than the positive pair in a training mini-batch as soft-negatives, pushing them away from the positives. Unlike previous contrastive methods, the NT-Xent loss also normalizes the embeddings to lie on a unit hypersphere in the representational space.

Wang & Isola (2020) introduced *uniformity* and *alignment* factors as desirable properties of the representational space and showed that the NT-Xent loss optimizes the uniformity and alignment properties in the limit of infinite negative samples, which is equivalent to infinite batch size for NT-Xent loss.

Many models have been recently proposed that generalize the contrastive loss to the multi-modal setting (Zhang et al., 2021; Jia et al., 2021; Radford et al., 2021). Before these, several works have explored the integration of different modalities using contrastive and triplet losses. Wang et al. (2018) uses a two-branch neural network architecture with a bidirectional retrieval ranking loss to learn joint embeddings for images and text. Wang et al. (2018) also included neighborhood-preserving constraints, sampling mini-batches such that they contained multiple positive matches for a given target, in order to properly capture the many-to-many relationships between images and texts (i.e multiple images matching one text sequence or vice versa). Mahajan et al. (2019) employs joint Wasserstein autoencoders to align multi-modal embeddings in a shared latent space. Further, Lee et al. (2018) proposes a stacked cross-attention mechanism to discover fine-grained correspondences between image regions and words for more accurate image-text matching. However, CLIP stands out from these approaches due to its simplicity and scalability. While prior work requires complex architectures and sampling strategies, CLIP uses a simple contrastive loss and random mini-batch sampling from a vast dataset (400M image-text pairs). CLIP's approach circumvents the requirement for complex architectures and additional loss terms, yet achieves strong zero-shot results on many standard image benchmarks. CLIP's efficiency and ability to generalize across tasks have led to its widespread adoption across a wide range of applications.

While CLIP has demonstrated remarkable advances in zero-shot learning and cross-modal tasks, Liang et al. (2022) observed that a *modality* gap appears in many models that use multi-modal contrastive learning. There have since been several attempts to close the gap while monitoring its effects on downstream task performance. Liang et al. (2022) show that altering the gap by simply trans-

lating the embeddings of one modality onto the other can positively impact zero-shot classification performance. Further, Oh et al. (2023) attribute the modality gap to the absence of hard-negatives in a training mini-batch and synthetically generated hard-negative samples from existing data points, showing that training with the hard-negatives improves representational quality. In this work, we frame the gap between the multi-modal embeddings as a problem of low uniformity and show that by simply optimizing for more alignment and uniformity (properties known to be desirable in uni-modal contrastive learning), we can significantly reduce the size of the gap, while increasing the quality of the representations learned.

2.1 Contrasting Images and Text using CLIP Loss

CLIP is an example of a contrastive model that learns image and text embeddings. CLIP loss (L_{CLIP}) is based on the *multi-class N-pair loss* that was first introduced in Sohn (2016). This loss has been termed as the NT-Xent loss (Chen et al., 2020) and has also been used in prior works (Wu et al., 2018; van den Oord et al., 2019). While NT-Xent loss is designed to work on data points from a single modality, L_{CLIP} is adapted to work on two different modalities of data.

In our scenario, the multi-modal dataset contains N images and corresponding captions. We obtain image embeddings $E_j^I \in \mathbb{R}^d$ by passing image I_j through the image encoder. Similarly, we produce the text embedding $E_j^T \in \mathbb{R}^d$ by passing caption T_j through the text encoder. CLIP aims to bring image embeddings and their corresponding caption embeddings closer together in CLIP latent space (*CLIP space*) by increasing the similarity (inner product $\langle ., . \rangle$) between the corresponding embeddings. The image and text embeddings are normalized to lie on a unit hypersphere in $\mathbb{R}^d$.

The full CLIP loss is:

$$L_{\text{CLIP}} = -\frac{1}{2N}\sum_{j=1}^{N}\log\left[\frac{\exp\left(\langle E_j^I, E_j^T\rangle/\tau\right)}{\sum_{k=1}^{N}\exp\left(\langle E_j^I, E_k^T\rangle/\tau\right)}\right] - \frac{1}{2N}\sum_{k=1}^{N}\log\left[\frac{\exp\left(\langle E_k^I, E_k^T\rangle/\tau\right)}{\sum_{j=1}^{N}\exp\left(\langle E_j^I, E_k^T\rangle/\tau\right)}\right] \tag{1}$$

Where the left term **contrasts images with the texts** ($\sum_{k=1}^{N}$ in the denominator loops over text embeddings as negatives for the jth image) and the right term **contrasts texts to images** ($\sum_{j=1}^{N}$ in the denominator loops over image embeddings as negatives for the kth text). τ represents the temperature parameter (τ = 0.01 at the end of CLIP pre-training).

3 The Gap in Multi-modal Contrastive Learning

Though CLIP effectively associates images with related texts, it also creates representational spaces with a *modality gap*, a phenomenon that has generated some interest. Liang et al. (2022) attributes this modality gap to two independent factors. First, they describe **the cone effect** of deep neural networks, in which different random initializations cause the embeddings of two encoders to occupy two non-overlapping narrow cones in CLIP space. Second, they note the existence of **mismatched pairs** in the dataset, where the pairing of images and captions is incorrect for some data points. Shi et al. (2023) studied the gap in three dimensions and attributed it to conflicting uniformity and alignment terms in the contrastive loss, claiming that this leads to the existence of local minima that encourage the gap. Further, (Oh et al., 2023; Liang et al., 2022) show that the modality gap persists even after fine-tuning. In our first experiment in Section 3.2, we will show that these factors cannot fully account for the modality gap by simulating an ideal scenario where all of the above factors are controlled for, but the gap still exists at the end of training. The goal of this work is to suggest that the *modality gap* is not caused by trying to align different modalities in a unified representational space, but instead arises as a consequence of the contrastive training that is used to align both modalities. We therefore suggest that the term *contrastive gap* better describes the underlying phenomenon.

	At initialization	After 25 epochs
Centroid distance	0.01	0.34
Linear separability acc.	0.52	1.00
Contrastive loss	3.42	0.00

Table 1: **Modality gap persists even when all factors are controlled for:** Modality gap metrics and CLIP loss values before and after training CLIP from scratch in ideal dataset conditions. **At initialization:** Distance between image and text centroids is almost zero and the embeddings are *not* linearly separable, meaning that there is no modality gap. **After training:** Centroid distance increases slightly, but the text and image embeddings are *perfectly* linearly separable. Thus, the modality gap is *created* by the contrastive loss.

3.1 Measuring the gap

To show that we have closed the gap between the embeddings of the two encoders, we must first find a way to quantify the gap. We introduce the following two metrics to measure the size and severity of the gap:

Distance between modality centroids (from Liang et al. (2022)) Given N images and N captions, we denote the centroid of the image embeddings as $C^I = 1/N \sum_{j=1}^{N} E_j^I$, and similarly for the centroid of the text embeddings. We compute the distance between centroids as $\|C^I - C^T\|^2$, and note that the centroid distance can vary from 0 to 2.

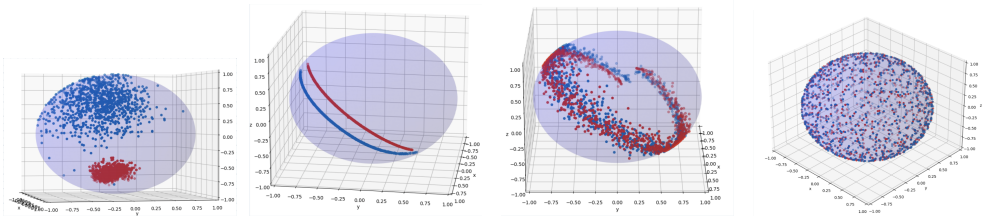
Linear Separability (from Shi et al. (2023)) is the percentage of image and text embeddings that can be distinguished by a linear classifier operating in CLIP space. We used 80% of the dataset to train a linear model to classify CLIP embeddings as originating from either "image" or "text" input. We then tested the performance of the classifier on the remaining 20% of the dataset and reported the accuracy. If a set of embeddings are 100% linearly separable, this means that the space occupied by each modality is completely disjoint. Conversely, 50% linear separability means that the image and text embeddings are overlapping in CLIP space, meaning that they occupy the same region of the latent space; i.e. there is no gap between the embeddings. To summarize, *if we can effectively close the gap we will find that the distance between centroids is small and linear separability is close to 50%*.

3.2 The Modality Gap Persists Even When All Factors Are Controlled For

We now systematically remove the factors commonly known to contribute to the modality gap. We started with the default CLIP architecture and used the MSCOCO (Lin et al., 2014) dataset. We created an idealized scenario where:

1. *There is only one modality.* We replaced the text encoder in CLIP with another copy of the image encoder and trained the model on pairs of images instead of text-image pairs. Thus, for this experiment, the CLIP encoders are identical image encoders with different random initializations.
2. *Embeddings from the two image encoders occupy the same cone at initialization*. In our setup, after we initialize the two image encoders, we computed a fixed transformation matrix that translates the embeddings of the second image encoder to overlap with those of the first image encoder (following Liang et al. (2022)). Thus, the second encoder's embeddings are translated to occupy the same narrow cone as the first encoder's embeddings at initialization. This way, there is no modality gap at initialization.
3. *There are no mismatched pairs.* The positive pairs in our constructed dataset are actually identical images. This eliminated the possibility that there would be mismatched pairs in the dataset.

We ran this experiment on the full MS COCO training set. We used a batch size of 64, and the default CLIP dimensionality of 512D (i.e. Image embeddings, $E^I \in \mathbb{R}^{512}$). We trained the model for 25 epochs and noted the final training loss value as close to zero.



(a) **Epoch 0**
$I \to T$ accuracy: 0.0

(b) **Epoch 37**
$I \to T$ accuracy: 0.0

(c) **Epoch 150**
$I \to T$ accuracy: 0.1

(d) **Epoch 275**
$I \to T$ accuracy: 0.87

Figure 1: Visualizing the training stages of 3D CLIP on 1000 image-text pairs from MS COCO. Red points are image embeddings, and blue points are text embeddings. $I \to T$ accuracy represents the text retrieval accuracies: Higher $I \to T$ accuracies mean that positive pairs are well contrasted in the latent space relative to the negative pairs. The embeddings of each modality are initialized to reside in separate cones due to the cone effect, before they form *arcs* and then eventually merge together as rings and spread out to fill the sphere.

We observe in Table 1 that even when the dataset is idealized and almost trivial to optimize on, and the model is initialized without a gap, after training to zero loss the CLIP embeddings are perfectly linearly separable. Thus, there is a contrastive gap that is *created* as a byproduct of the contrastive loss, even when the two encoders learn to align the same modality, with overlapping initializations and no mismatched pairs.

3.3 Visualizing the gap in 3D CLIP

We further investigate the persistence of the gap in CLIP space by training CLIP in 3-dimensional space. Surprisingly, in this lower-dimensional setting, we found that the contrastive gap can be closed even in a non-idealized setting (similar to typical CLIP training). To illustrate this, we visualized the training process using 1,000 randomly sampled image-text pairs from the MS COCO dataset in Figure 1. This finding suggests that closing the contrastive gap may be easier in lower-dimensional spaces vs. in high-dimensional spaces.

We report the text-retrieval accuracies in Figure 1 which shows how accurately we can retrieve an image's caption using the image's embedding and comparing it against all the caption embeddings in the dataset. We see that points on the 3D sphere are best aligned when they are evenly distributed on the sphere (as evidenced by the very high text-retrieval accuracy when Figure 1d). Based on this observation, we speculate that it is desirable to close the contrastive gap *and* to distribute the embeddings more uniformly on the unit sphere in $\mathbb{R}^d$ to improve the quality of learned representations.

3.4 Why does the contrastive loss induce a gap?

In Section 3.2, we observed that the contrastive gap persists even when training CLIP on idealized conditions. We speculate that this is because of the small value of the learned temperature parameter $\tau = 0.01$ in CLIP. As shown by Wang & Liu (2021b), when the temperature approaches zero, the contrastive loss becomes a triplet loss with a margin of zero. This means that the loss only focuses on the nearest negative sample within the batch: Once the similarity between a positive pair is greater than that between the positive and the nearest negative sample, the loss stops pushing the positive pair closer together. Wang & Liu (2021b) uses the following approximation to illustrate this point:

$$\begin{aligned} &\lim_{\tau \to 0^+} -\log \left[\frac{\exp(s_{i,i}/\tau)}{\sum_{k=1}^{M} \exp(s_{(i,k)}/\tau)} \right] \\ &= \lim_{\tau \to 0^+} \frac{1}{\tau} \max[s_{\max} - s_{i,i}, 0] \end{aligned} \tag{2}$$

Where $s_{(i,i)}$ is the similarity between the ith sample and its positive pair, $s_{(i,k)}$ is the similarity between the ith sample and the kth sample (forming a negative pair), and $s_{\max}$ is the similarity

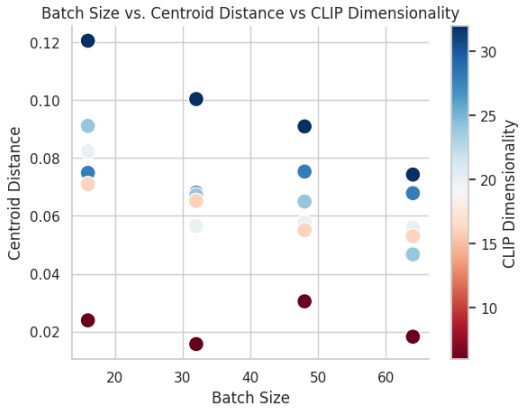


Figure 2: Interplay between batch size, CLIP dimensionality, and size of the contrastive gap. Points trend downwards towards the right (as batch size increases) and are more red (lower CLIP dimensionalities) towards the bottom. This indicates that large batch sizes and smaller CLIP dimensionalities lead to smaller contrastive gaps.

between the ith sample and the nearest negative pair to the ith sample (i.e, the maximum of all the negative similarities). While a reasonable idea would be to increase the temperature to close the gap between embeddings, this requires manual hand-tuning for each downstream task (Oh et al., 2023; Al-Jaff, 2023), and introduces a trade-off between uniformity and alignment (Wang & Liu, 2021a). Further, Udandarao (2022) shows that increasing temperature to close the gap between embeddings could degrade the quality of the representational space and hurt downstream task performance.

3.5 Interplay between dimensionality and batch size in CLIP

Referring to Equation 2, one way to reduce the gap would be to increase $s_{\max}$ which would push negative samples closer to the positive pair. This, in turn, would force the loss to increase $s_{(i,i)}$, potentially closing the gap. Following this reasoning, we can analyze how dimensionality and batch size might influence the contrastive gap by considering their impact on $s_{\max}$:

1. **Increasing batch size** adds more negative samples to the mini-batch, which raises the chance of having more similar negative samples. Let $s^m_{\max}$ represent the maximum similarity to a negative pair for a batch size of m. If we increase the batch size to $m+1$, we have $s^m_{\max} \leq s^{m+1}_{\max}$, as the additional negative sample can only increase or maintain the maximum similarity. As a result, increasing batch size is likely to increase $s_{\max}$ and thereby *reduce* the contrastive gap.
2. **Increasing dimensionality of CLIP** expands the space in which the embeddings are positioned, represented by a unit hypersphere. As dimensionality increases, the distance between randomly selected points on the unit sphere converges to $\sqrt{2}$, and most points start to get far apart (Blum et al., 2020). This means that in higher dimensions, $s_{\max}$ tends to decrease. Consequently, increasing CLIP dimensionality is likely to *increase* the contrastive gap.

We empirically demonstrate in Figure 2 how the interplay between CLIP dimensionality and batch size affects the contrastive gap. Our findings confirm that smaller batch sizes and higher dimensionalities are associated with a larger contrastive gap. This helps explain why the contrastive gap is so common in multi-modal contrastive models: They are trained with relatively small batch sizes in very high-dimensional spaces.

Wang & Isola (2020) show that the uni-modal contrastive loss, in the limit of infinite batch size, asymptotically optimizes for two properties: *uniformity* and *alignment*. In the next section, we adapt these properties to multi-modal contrastive learning and incorporate them into our loss function to reduce the contrastive gap without needing to increase batch size.

4 MULTIMODAL UNIFORMITY AND ALIGNMENT

We now introduce the concepts of uniformity and alignment in the representational space. *Uniformity* refers to the property of the embeddings being uniformly distributed throughout the contrastive latent space. *Alignment* refers to the positive pairs being close together (aligned) in the latent space. To study the effects of closing the contrastive gap, we adapt the uniformity and alignment properties from Wang & Isola (2020) to the multi-modal contrastive space as follows:

$$\text{Uniformity for Image space} : L^I_{\text{Uniform}} = \log \left(\frac{1}{N} \sum_{j=1}^{N} \sum_{k=1}^{N} \exp\left(-2\|E^I_j - E^I_k\|^2\right) \right) \tag{3}$$

We define the uniformity for text space (L^T_{Uniform}) similarly. Finally, the total L_{Uniform} term is:

$$L_{\text{Uniform}} = \frac{1}{2}(L^T_{\text{Uniform}} + L^I_{\text{Uniform}}) \tag{4}$$

L^I_{Uniform} and L^T_{Uniform} each encourage the uniformity *within* the image and text embeddings respectively. i.e., L_{Uniform} only encourages *intra-modality* uniformity. The original multi-modal contrastive loss (Equation 1) does *not* have any such term that constrains embeddings within each modality to be far apart. Instead, the denominators in Equation 1 only push negative text samples away from the positive image sample and vice versa.

To enforce a stronger constraint on the uniformity *between* negative image and text samples, we also introduce a *cross-modality uniformity* term:

$$\text{Cross-modality uniformity} : L_{\text{XUniform}} = \log \left(\frac{1}{N} \sum_{j=1}^{N} \sum_{k=1, k\neq j}^{N} \exp\left(-2\|E^I_j - E^T_k\|^2\right) \right) \tag{5}$$

Finally, to better align positive image-text pairs in CLIP space, we adapt the alignment term to the multi-modal setting:

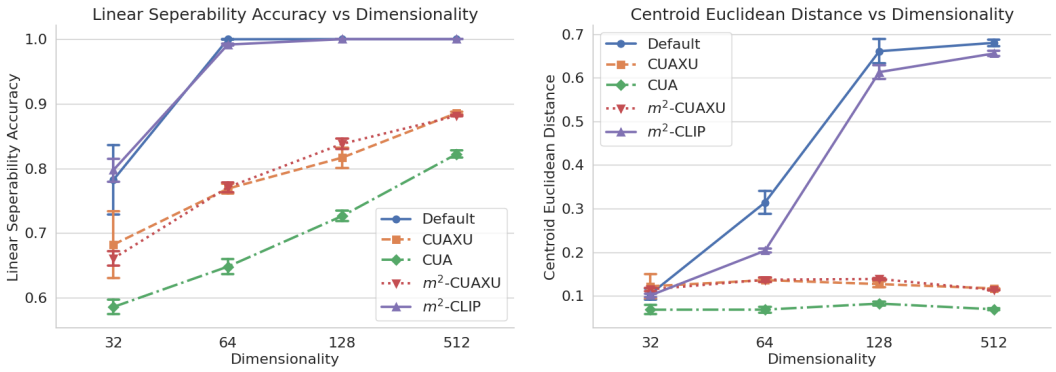
$$L_{\text{Align}} = \frac{1}{N} \sum_{j=1}^{N} (\|E^I_j - E^T_j\|^2) \tag{6}$$

Uniformity and alignment properties have been shown to be desirable properties in the uni-modal contrastive space (Wang & Isola, 2020). We validate the desirability of these two properties in the multi-modal setting. In prior works evaluating the multi-modal contrastive representational space using alignment and uniformity properties, Goel et al. (2022) and Oh et al. (2023) used cross-modality uniformity to measure uniformity in the latent space (after fine-tuning with their proposed m^2-mix loss). Meanwhile, Al-Jaff (2023) explored explicitly training multi-modal models using L_{Uniform} (i.e., only the intra-modality uniformity term) and L_{Align}. In our work, we combine both the intra-modality and cross-modality uniformity terms to encourage more uniformity in the latent space and reduce the size of the contrastive gap.

5 EXPERIMENTS

We studied the effects of reducing the contrastive gap in CLIP space by optimizing for uniformity and alignment of the latent space. We fine-tuned a pre-trained CLIP model by adding the L_{Uniform}, L_{XUniform}, and L_{Align} terms to the original CLIP loss (L_{CLIP}, Equation 1). Additionally, we compare to the m^2-mix loss from Oh et al. (2023). The m^2-mix loss requires that we generate synthetic hard negatives by interpolating (*mixing*) between the image and text embeddings. These newly generated synthetic samples are used as in place of the original in-batch negatives.

We demonstrate the effects of fine-tuning pre-trained CLIP on the following losses:



(a) Linear Separability Accuracy vs. CLIP dimensionality.

(b) Distance between image and text centroids vs. CLIP dimensionality.

Figure 3: Contrastive gap metrics after fine-tuning with the different losses for different CLIP dimensionalities.

- L_{CLIP}: The default CLIP loss
- L_{CUA}: L_{CLIP} + L_{Uniform} + L_{Align}
- L_{CUAXU}: L_{CLIP} + L_{Uniform} + L_{Align} + L_{XUniform}
- $L_{m^2\text{-CLIP}}$: $L_{m^2\text{-mix}}$ + L_{CLIP}
- $L_{m^2\text{-CUAXU}}$: $L_{m^2\text{-mix}}$ + L_{CLIP} + L_{Uniform} + L_{Align} + L_{XUniform}

Hyperparameters Overview For our experiments, we fine-tuned the CLIP model made available by Radford et al. (2021). We used the ViT-B/32 variant of the image encoder and a transformer (from OpenAI's official implementation[1]) as the text encoder. We studied the effects of fine-tuning over various sizes of CLIP space ($\mathbb{R}^d$, $d \in [32, 64, 128, 512]$). When fine-tuning, we adjusted the dimensionality of CLIP by randomly re-initializing the final linear projection layer of CLIP. For all our experiments, we fixed the temperature (τ) parameter to 0.01, as τ converges to this value after CLIP pre-training. We fine-tuned on the MS COCO (Lin et al., 2014) dataset for 9 epochs. We list all hyperparameter settings in the Appendix B.3

Experiments Structure We first show the effects the new loss functions have on the representational space by measuring the size of the contrastive gap on the MS COCO validation dataset. Then, we analyze the effects of reducing the gap by measuring image-text retrieval performance on the Extended COCO Validation (ECCV) Caption dataset (Chun et al., 2022). We then experiment to see if the same results hold for off-distribution datasets by evaluating zero-shot image classification performance on 13 standard datasets. Finally, we explore finer differences in the representational spaces learned by the different loss functions by reporting performance on multi-modal arithmetic.

5.1 Effects of the new losses on the size of the contrastive gap

First we explore the contrastive gap metrics after fine-tuning on the new losses. We compute the gap metrics on the MS COCO validation dataset (5k image-caption pairs). We use the first of the five captions available per image in MS COCO. Figure 3a shows the linear separability accuracy and Figure 3b shows the distance between centroids vs. CLIP dimensionality for each of the different losses. We observe that L_{CUA}, L_{CUAXU}, and $L_{m^2\text{-CUAXU}}$ have lower measures of both the metrics of the contrastive gap. This indicates that the size of the gap is much smaller with uniformity and alignment terms included. The differences in the size of the contrastive gap are more pronounced in higher CLIP dimensionalities. These results support our claim that increasing uniformity and alignment in CLIP space reduces the size of the contrastive gap. Further, it appears that directly optimizing for uniformity and alignment may lead to smaller contrastive gaps, compared to learning with synthetic hard negatives, as done in Oh et al. (2023).

[1]https://github.com/openai/CLIP

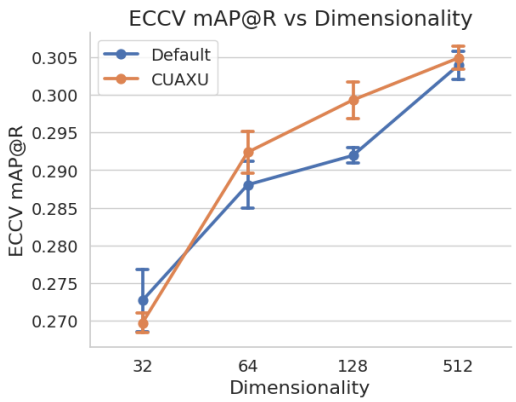


Figure 4: mAP@R for fine-tuned CLIP on different losses on the ECCV-Captions dataset.

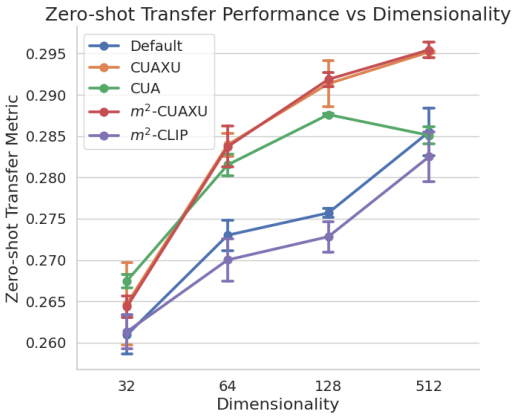


Figure 5: Average zero-shot classification accuracies for fine-tuned CLIP on the different losses.

5.2 Effects of the new losses on Image-text retrieval

Next, we evaluated the quality of the representational space after reducing the contrastive gap. We measured the image-text retrieval performance on ECCV-Captions dataset. The MS COCO dataset has many missing correspondences between images and captions: While one caption has only one matching image, that same caption could be a equally plausible description for another image in the dataset, and vice versa. The presence of these *false negatives* lead to a misrepresentation of the image-text retrieval performance if measured directly on MS COCO: We give examples in Figure 9 in Appendix A.3. We also show the Top5 image and text retrieval accuracies as measured directly on the MS COCO dataset in Figure 8 in Appendix A.3.

ECCV-Captions extends the MS COCO dataset by adding the missing image-caption correspondences using human and machine annotators, correcting the false negatives. We report the mAP@R metric for average image-text recalls in Figure 4. mAP@R is more aligned to humans than TopK Recall, and is therefore the metric recommended by Chun et al. (2022). Figure 4 shows that models fine-tuned with L_{CUAXU} slightly outperform those fine-tuned with L_{CLIP} . The general increase in retrieval performance over L_{CLIP} suggests that optimizing for uniformity and alignment leads to representational spaces with more accurate image-text correspondences.

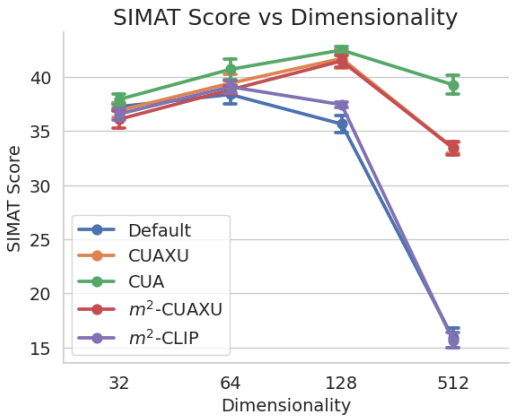
5.3 Zero-Shot Transfer

We evaluate our fine-tuned CLIP models on 13 standard image-classification datasets (listed in the Appendix A.4). We adopt the evaluation strategy of Goel et al. (2022) and recommended by Radford et al. (2021): generate prompts using the class names to form sentences like "a photo of a class name", "A sketch of a "class name", etc. and match these sentence embeddings to image embeddings output from CLIP to classify an image.

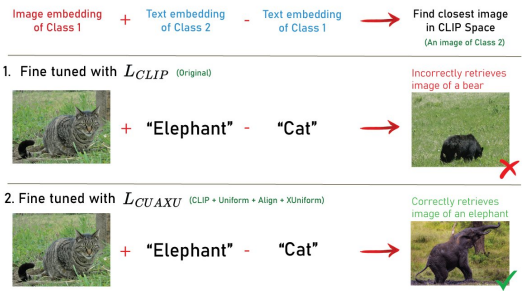
Figure 5 shows the average zero-shot accuracies across all the datasets for each of the losses and dimensionalities. CLIP losses with alignment/uniformity consistently outperform the default CLIP loss, and XUniform adds additional benefit. Thus, we argue that representational spaces with smaller contrastive gaps (as learned by models fine-tuned with added L_{CUA} and L_{CUAXU}) appear to correlate with higher performance for the zero-shot transfer task. We reason that a smaller contrastive gap likely leads to a better representation of "concepts" within images, which benefits zero-shot classification and average image-text retrieval performances.

5.4 Multimodal Arithmetic

A high-quality multi-modal representational space should have consistent structural representations across the learned modalities. We used SIMAT (for Semantic IMage Transformation) (Couairon et al., 2022) to evaluate such relationship consistencies between CLIP image and text embeddings. SIMAT computes a new image representation after transforming it with *text delta vectors* and re-



(a) SIMAT Score vs CLIP dimensionality for the different losses.



(b) Multi-modal arithmetic: Given an input image and a text transformation ("Cat" $\rightarrow$ "Elephant"), model fine-tuned with L_{CLIP} retrieves incorrect image in this example, whereas model fine-tuned with L_{CUAXU} retrieves the correct image.

Figure 6: Multi-modal arithmetic evaluation of models fine-tuned with the different losses.

trieves the closest image to the transformed embedding. (i.e., $E^I_{\text{target}} = E^I_{\text{input}} + \lambda(E^T_{\text{target}} - E^T_{\text{input}})$, where λ is a hyperparameter controlling the strength of the transformation).

Figure 6a shows SIMAT scores across the different loss functions. Adding uniformity and loss terms to the CLIP loss leads to increased SIMAT scores, indicating that learning with uniformity and alignment terms results in a more interpretable mapping between modalities in multi-modal arithmetic tasks. Notably, the difference in SIMAT scores between the baseline and the modified CLIP losses becomes more pronounced as the dimensionality of CLIP increases. We see a similar trend in Figures 3a and 3b, where fine-tuning with added L_{Uniform} and L_{Align} terms lead to representational spaces with significantly lower contrastive gap, especially in higher dimensionalities, compared to fine-tuning with L_{CLIP} . Figure 6b illustrates the multi-modal arithmetic task.

From these empirical findings, we conclude that having a smaller contrastive gap is well correlated with higher performance on the multi-modal arithmetic task and improves the relative geometry of the two spaces. Our results suggest that closing the contrastive gap by fine-tuning with added uniformity/alignment terms could benefit applications that rely on the geometric structure and consistent arithmetic properties in the latent space.

6 Conclusions

In this paper, we studied the representations learned by multi-modal contrastive learning algorithms and analyzed the contrastive gap phenomenon. Our analysis revealed that controlling for the reasons commonly thought to cause the gap does *not* close it. Thus, this is *not a modality gap*. We instead propose the term *contrastive gap* to describe this phenomenon.

We investigated the relationship between dimensionality and batch size, concluding that the contrastive gap is exacerbated by small batch sizes in high dimensional CLIP space. To reduce the gap without increasing the batch size, we proposed adding explicit uniformity and alignment terms to the CLIP loss. Our results demonstrate that directly optimizing for uniformity and alignment in the latent space significantly reduces the gap. This supports our claim that increasing uniformity and alignment in the CLIP space can reduce the contrastive gap.

Moreover, we found that closing the gap through fine-tuning with added uniformity and alignment terms improved average image-text retrieval and zero-shot image classification performances, suggesting that reducing the contrastive gap could lead CLIP to develop a more accurate understanding of concepts in images. Further, the new losses also improved multi-modal arithmetic performance, suggesting that the new representational space could have more consistent geometric structure.

While our experiments focused on fine-tuning on MS COCO, we hope to expand our scope and explore larger datasets in the future to understand the impact of uniformity and alignment at scale.

7 Reproducibility Statement

We now highlight our efforts to ensure the reproducibility of our results. We included the code we used to produce our results in the supplemental materials section of this submission. Further, we specify the hyperparameters used in our experiments in Section 5 and in the Appendix B.3. We also specify the compute hardware we used for our experiments in the Appendix B.2. Moreover, we use publicly available model architectures, and we specify the sources of their implementations in the text whenever applicable. Finally, we use freely available datasets for all our experiments, and specify the dataset details in Section 5 and in the Appendix (A.4, B.1)

References

Mohammad Al-Jaff. *Messing With The Gap: On The Modality Gap Phenomenon In Multimodal Contrastive Representation Learning*. PhD thesis, Uppsala University, November 2023.

Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of data science*. Cambridge University Press, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations, June 2020. URL http://arxiv.org/abs/2002.05709. arXiv:2002.05709 [cs, stat].

Sanghyuk Chun, Wonjae Kim, Song Park, Minsuk Chang, and Seong Joon Oh. ECCV Caption: Correcting False Negatives by Collecting Machine-and-Human-verified Image-Caption Associations for MS-COCO. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, volume 13668, pp. 1–19. Springer Nature Switzerland, Cham, 2022. ISBN 978-3-031-20073-1 978-3-031-20074-8. doi: 10.1007/978-3-031-20074-8_1. URL https://link.springer.com/10.1007/978-3-031-20074-8_1. Series Title: Lecture Notes in Computer Science.

Guillaume Couairon, Matthieu Cord, Matthijs Douze, and Holger Schwenk. Embedding arithmetic of multimodal queries for image retrieval, 2022. URL https://arxiv.org/abs/2112.03162.

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A. Rossi, Vishwa Vinay, and Aditya Grover. CyCLIP: Cyclic Contrastive Language-Image Pretraining, October 2022. URL http://arxiv.org/abs/2205.14459.

Steven M Holland. Principal components analysis (pca). *Department of Geology, University of Georgia, Athens, GA*, 30602:2501, 2008.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked Cross Attention for Image-Text Matching. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision – ECCV 2018*, volume 11208, pp. 212–228. Springer International Publishing, Cham, 2018. ISBN 978-3-030-01224-3 978-3-030-01225-0. doi: 10.1007/978-3-030-01225-0_13. URL https://link.springer.com/10.1007/978-3-030-01225-0_13. Series Title: Lecture Notes in Computer Science.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning, October 2022. URL http://arxiv.org/abs/2203.02053. arXiv:2203.02053 [cs].

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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Andrew Luo, Margaret Marie Henderson, Michael J. Tarr, and Leila Wehbe. BrainSCUBA: Fine-grained natural language captions of visual cortex selectivity. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mQYHXUUTkU.

Shweta Mahajan, Teresa Botschen, Iryna Gurevych, and Stefan Roth. Joint Wasserstein Autoencoders for Aligning Multimodal Embeddings, September 2019. URL http://arxiv.org/abs/1909.06635. arXiv:1909.06635 [cs].

Lukas Muttenthaler, Lorenz Linhardt, Jonas Dippel, Robert A. Vandermeulen, Katherine L. Hermann, Andrew K. Lampinen, and Simon Kornblith. Improving neural network representations using human similarity judgments. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/9febda1c8344cc5f2d51713964864e93-Abstract-Conference.html.

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

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. 2021.

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

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

Vishaal Udandarao. Understanding and Fixing the Modality Gap in Vision-Language Models. 2022.

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

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2495–2504, 6 2021a.

Feng Wang and Huaping Liu. Understanding the Behaviour of Contrastive Loss. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2495–2504, Nashville, TN, USA, June 2021b. IEEE. ISBN 978-1-66544-509-2. doi: 10.1109/CVPR46437.2021.00252. URL https://ieeexplore.ieee.org/document/9577669/.

Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning Two-Branch Neural Networks for Image-Text Matching Tasks, May 2018. URL http://arxiv.org/abs/1704.03470. arXiv:1704.03470 [cs].

Tongzhou Wang and Phillip Isola. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. 2020.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6787–6800, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.544. URL https://aclanthology.org/2021.emnlp-main.544.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis Langlotz. Contrastive learning of medical visual representations from paired images and text, 2021. URL https://openreview.net/forum?id=T4gXBOXoIUr.

Chenliang Zhou, Fangcheng Zhong, and Cengiz Oztireli. Clip-pae: Projection-augmentation embedding to extract relevant features for a disentangled, interpretable, and controllable text-guided face manipulation, 2023.

A Appendix

A.1 Motivation behind training CLIP in 3D

We train CLIP in 3D in Section 3.3 to better visualize the persistence of the gap on the 3D hypersphere. Our study builds on the work in Shi et al. (2023), where the authors also studied the behaviour of CLIP loss by visualizing training on a 3D sphere. The authors devised this proof of concept experiment to show that the CLIP loss induces a gap even in low dimensional space without external factors coming from the dataset or neural network architecture. They randomly generated 1,000 points on the 3D sphere and trained with a batch size of 10, learning rate of 0.01, using the SGD optimization algorithm. The parameter space for the SGD optimization algorithm was the euclidean coordinate space of the generated points. This allowed the algorithm to directly optimize the positions of the points to reduce the CLIP loss.

Their experiment highlights that the modality gap can arise even in low-dimensional spaces and without any external factors from the dataset or neural network architecture, but can eventually be closed after a large number of training steps.

Shi et al. (2023), we studied the behaviour of CLIP loss when we reduce the CLIP dimensionality from 512D to 3D. We extended the proof-of-concept experiment done by Shi et al. (2023) by

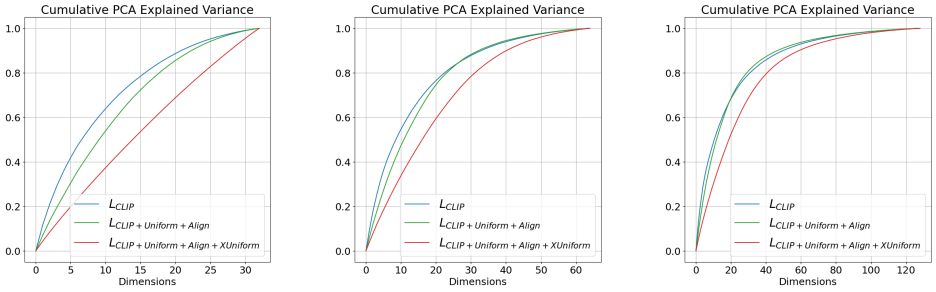
- **Using Real Data**: We sampled 1,000 image-text pairs from the MS-COCO dataset rather than using synthetic points. This way, we can better capture the nuances of a real-world data distribution.
- **Optimizing with CLIP architecture**: In the original experiment by Shi et al. (2023), the authors bypassed the neural network entirely by directly optimizing the positions of points in a 3D Euclidean space. However, in our modified experiment, we utilized the CLIP model to project image and text representations into a 3D space (by adjusting the output projection layer from 512D to 3D). By doing so, we could examine how the CLIP model and its loss function behave together, allowing us to assess the impact of both the neural network and the loss function in a more realistic setting.

We visualize the results of our 3D experiment in Section 3.3 of the main text.

A.2 Measuring distribution of embeddings on the CLIP hypersphere

We showed in Section 5.1 of the main text the effects of the various new losses on the size of the contrastive gap between image and text embeddings. Here, we evaluate the distribution of embeddings across latent space dimensions by applying Principal Component Analysis (PCA) and analyzing the explained variance ratios (Holland, 2008)

PCA is a dimensionality reduction technique that summarizes the data into a smaller set of principal components (PCs). The explained variance ratio indicates how much of the original data's variance is captured, or "explained", by each PC. By examining the cumulative PCA explained variance curve, we can assess how well the embeddings are spread across the dimensions of the unit hypersphere in CLIP space. Ideally, if the embeddings are uniformly distributed across all dimensions—indicating high uniformity—the cumulative PCA explained variance curve will form a straight line. Figure 7 shows the cumulative PCA explained variances for the three different losses for CLIP space in $\mathbb{R}, d \in [32, 64, 128]$.



(a) **32D CLIP** (b) **64D CLIP** (c) **128D CLIP**

Figure 7: PCA explained variances for each CLIP dimensionality after fine-tuning

From the figure, we see that L_{CUAXU} has the lowest cumulative variance across all the principle components for all the three dimensionalities of CLIP space tested. This indicates that the the uniformity terms help to distribute the embeddings along more dimensions in the contrastive latent space in $\mathbb{R}^d$, across a wide range of d's

A.3 Image-text retrieval performances

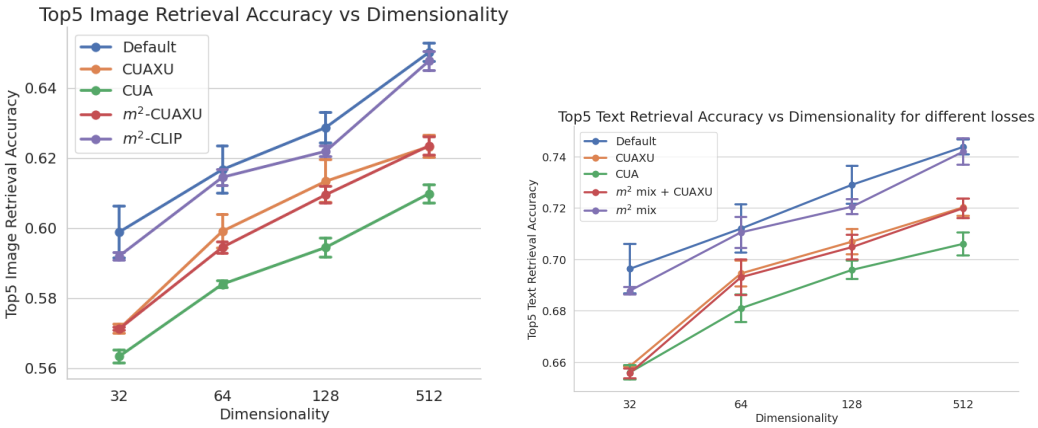
Figure 8 shows the image and text retrieval accuracies for MS COCO validation set after fine-tuning CLIP on the different losses. As described in the main text in Section 5.2, MS COCO dataset has many false negatives, which may misrepresent the retrieval performance of the models. We explain this issue in more detail here.

We speculate that smaller contrastive gap leads to better representation of "concepts" within images. However, in image-retrieval and text-retrieval tasks, the model needs to precisely match the caption to the exact ground truth image. Even if the model retrieves an image with the correct concepts, it will be considered incorrect if it does not match the exact ground truth image. Figure 9 demonstrates this, illustrating two examples where models fine-tuned with both L_{CLIP} and L_{CUAXU} retrieves the "wrong" caption for the input image. We observe that even though the retrieved captions do not match the ground truth captions exactly, they are not necessarily wrong, as they encapsulate the correct concepts that are present in the images.

The ECCV-Caption dataset accounts for these false negatives. Therefore, we use ECCV-Caption to measure image-text retrieval performance of our fine-tuned models in Section 5.2

A.4 Zero shot transfer datasets

We evaluate our fine-tuned CLIP models on the standard image-classification datasets outlined in Table 2. To test the zero-shot transfer capabilities of our CLIP models, we adopt the evaluation strategy of Goel et al. (2022), which is also recommended by Radford et al. (2021): We generate prompts using the class names to form sentences like "a photo of a {class name}", "A sketch of a {class name}", etc. We then pass these sentences through the text encoder to get prompt embeddings. We average all the prompt embeddings to get a class embedding for each class. Finally, to classify an image, we pass the input image through the image encoder and obtain its image embedding. We then determine the predicted class by finding the class embedding closest to the image embedding using cosine similarity.



(a) Top5 Image Retrieval Accuracy $(T \rightarrow I)$ vs CLIP dimensionality.

(b) Top5 Image Retrieval Accuracy $(I \rightarrow T)$ vs CLIP dimensionality.

Figure 8: Image and text retrieval accuracies for different CLIP dimensionalities.

(a) **Ground Truth Caption:** Three stuffed animals are sitting on a bed.
Retrieved Caption using L_{CLIP} : This is a picture of four stuffed animals.
Retrieved Caption using L_{CUAXU}: Two teddy bears are seated with the other stuffed animals.

(b) **Ground Truth Caption:** A black fluffy cat sitting on top of a computer keyboard.
Retrieved Caption using L_{CLIP} : A cat looking up at the television that has something interesting on it.
Retrieved Caption using L_{CUAXU}: Gray and white cat sitting next to open laptop.

Figure 9: Illustrating two instances where models fine-tuned with both L_{CLIP} and L_{CUAXU} make a mistake in retrieving the nearest caption for a given image.

B Experimental Setup

B.1 Dataset

We fine-tune on the MS COCO dataset downloaded from (Lin et al., 2014) [2]. We use the 2017 split, with 118k training images, and 5k validation images. Each image has 5 human-generated captions associated with it. In our experiments, we only take the first caption for each image.

[2] `https://cocodataset.org`

Dataset	**Classes**	**Test size**	**Evaluation metric**
CIFAR-10	10	10,000	Accuracy
CIFAR-100	100	10,000	Accuracy
SUN397	397	19,850	Accuracy
Pascal VOC 2007	20	4,952	11-Point mAP
Oxford-IIIT Pets	37	3,669	Mean Per Class
Caltech-101	102	6,085	Mean Per Class
ImageNet	1000	50,000	Accuracy
ImageNet-V2	1000	10,000	Accuracy
ImageNet-Sketch	1000	50,000	Accuracy
ImageNet-A	200	7,500	Accuracy
ImageNet-R	200	30,000	Accuracy
ImageNet-O	200	2,000	Accuracy
ObjectNet	113	50,000	Accuracy

Table 2: Datasets evaluated on to test zero-shot image-classification performance of the different CLIP losses.

Hyperparameter	Value
Image encoder	ViT/B-32
Text Encoder	Transformer (same as in [4])
Embedding dimensions	[32, 64 ,128, 512]
Temperature	0.01
Epochs	9
Batch size	64
Learning rate	1e-6
Adam beta1	0.9
Adam beta2	0.99
Adam weight decay	0.1
Scheduler	None

Table 3: Hyperparameters used for fine-tuning the CLIP models.

B.2 Computational Resources Used

training runs, we use NVIDIA RTX A5000 GPUs, and Intel(R) Xeon(R) Silver 4210 CPUs @ 2.20GHz. One run for fine-tuning CLIP from pre-trained weights on MS COCO for 9 epochs needed about 7GB GPU memory, and needed about 3.5 hours to complete (on a single GPU).

B.3 Model Architecture and Hyperparameters

We fine-tune the pre-trained CLIP model made available by OpenAI from Huggingface (Wolf et al., 2020)[3]. We list the model hyperparameters that we use below:

C Limitations

In this work, we explore the properties of uniformity and alignment in the multi-modal setting, showing that added uniformity and alignment terms help to reduce the contrastive gap. One limitation of our work is that we show this at a relatively small scale by fine-tuning CLIP on the MS COCO dataset. Training CLIP from scratch on a significantly larger dataset might help gain more insights into how the contrastive loss emerges during training at a larger scale, which is important as the success of CLIP was dependent on pre-training it on a massive dataset. Another limitation is that even after optimizing for uniformity and alignment in the MS COCO space, we see that the image-text retrieval accuracies do not change significantly from that of default CLIP fine-tuning. While we

[3] https://huggingface.co/docs/transformers/en/model_doc/clip

argue in this work that uniformity and alignment are desirable properties in the multi-modal representational space (and directly optimize for them with our losses), the image-text retrieval accuracy numbers in the MS COCO validation set indicate that there might be other properties of the latent space that correlate to performance on this task.