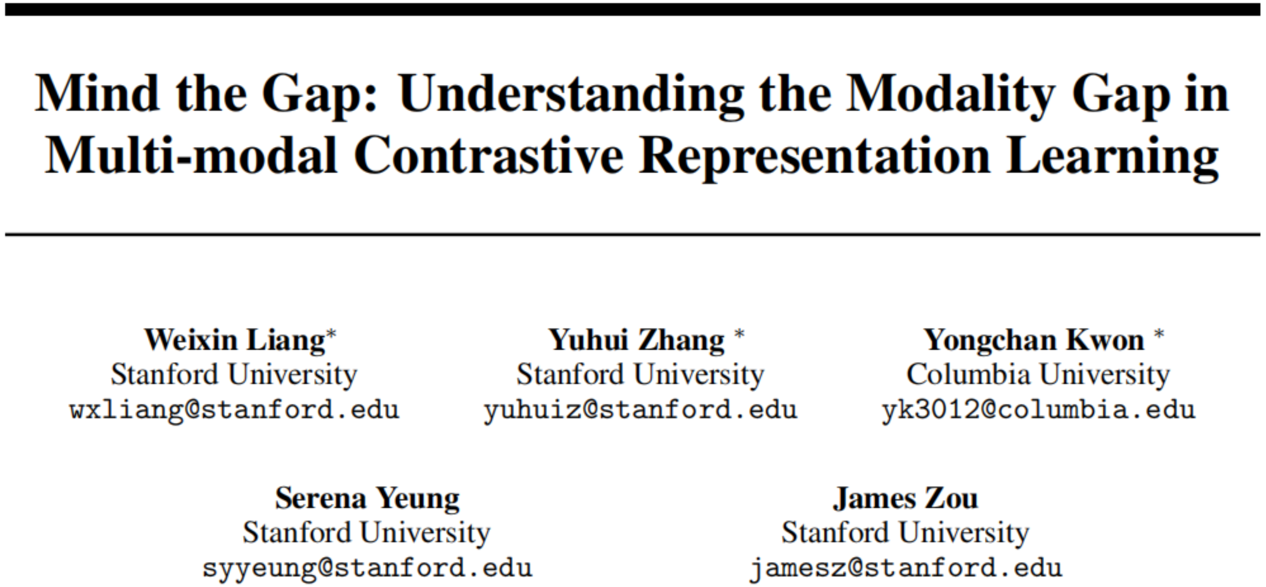
北京大學
PEKING UNIVERSITY

Modality Gap

彭天天

2025/03/14

Mind the Gap



Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning

Weixin Liang*
Stanford University
wxliang@stanford.edu

Yuhui Zhang *
Stanford University
yuhuiz@stanford.edu

Yongchan Kwon *
Columbia University
yk3012@columbia.edu

Serena Yeung
Stanford University
syyeung@stanford.edu

James Zou
Stanford University
jamesz@stanford.edu

NeurIPS 2022

Modality Gap

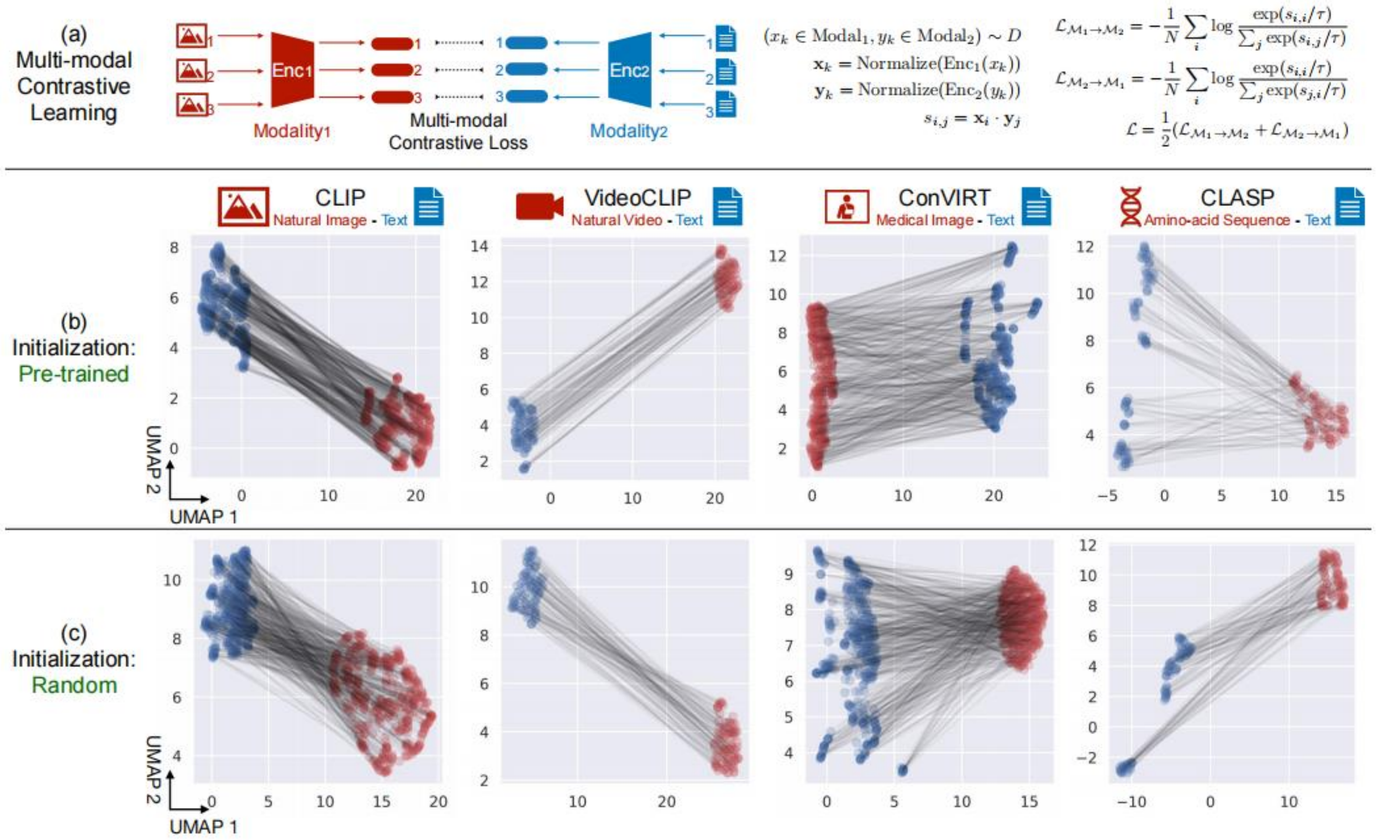
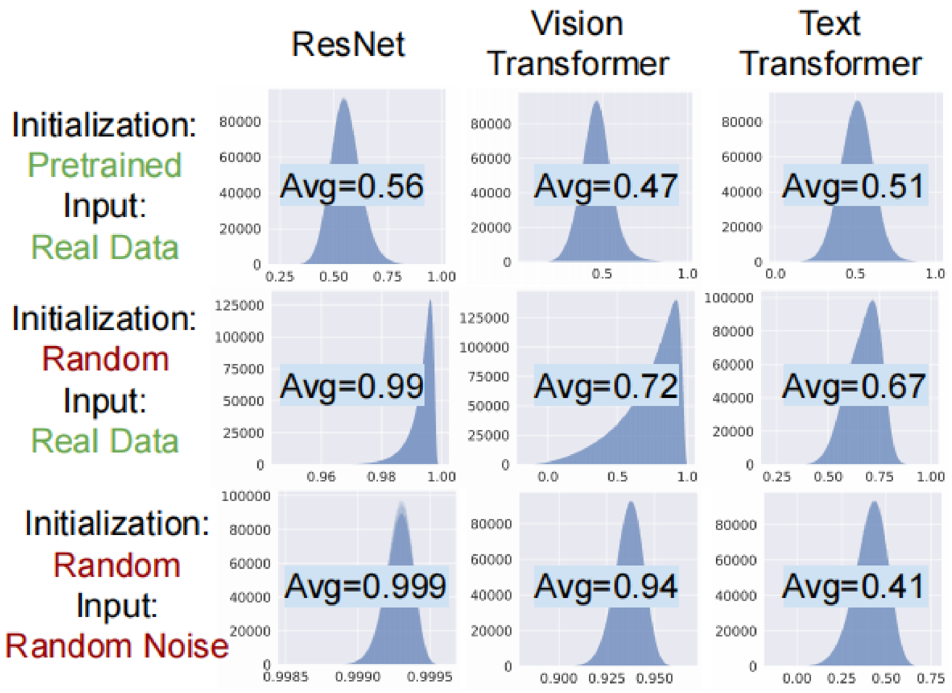


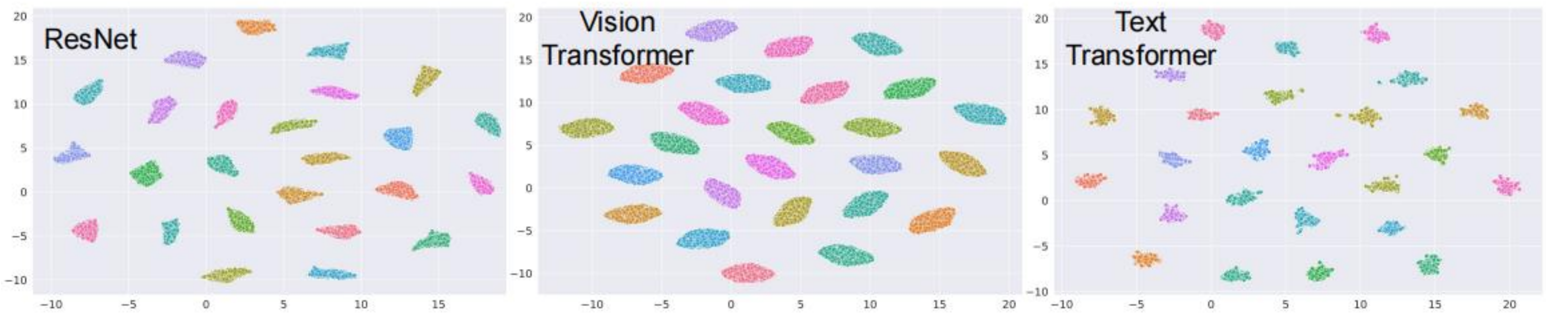
Figure 1: **The pervasive *modality gap* in multi-modal contrastive representation learning. (a) Overview of multi-modal contrastive learning.** Paired inputs from two modalities (e.g., image-caption) are sampled from the dataset and embedded into the hypersphere using two different encoders. The loss function is to maximize the cosine similarity between matched pairs given all the pairs within the same batch. **(b) UMAP visualization of generated embeddings from pre-trained models.** Paired inputs are fed into the pre-trained models and the embeddings are visualized in 2D using UMAP (lines indicate pairs). We observe a clear modality gap for various models trained on different modalities. **(c) UMAP visualization of generated embeddings from same architectures with random weights.** Modality gap exists in the initialization stage without any training.

多模态模型（无论是否训过，无论哪种模态pair）的特征空间内存在 modality gap——两个模态的特征空间之间有一定的距离

Cone Effect



(a) The cosine similarity between all pairs of embeddings



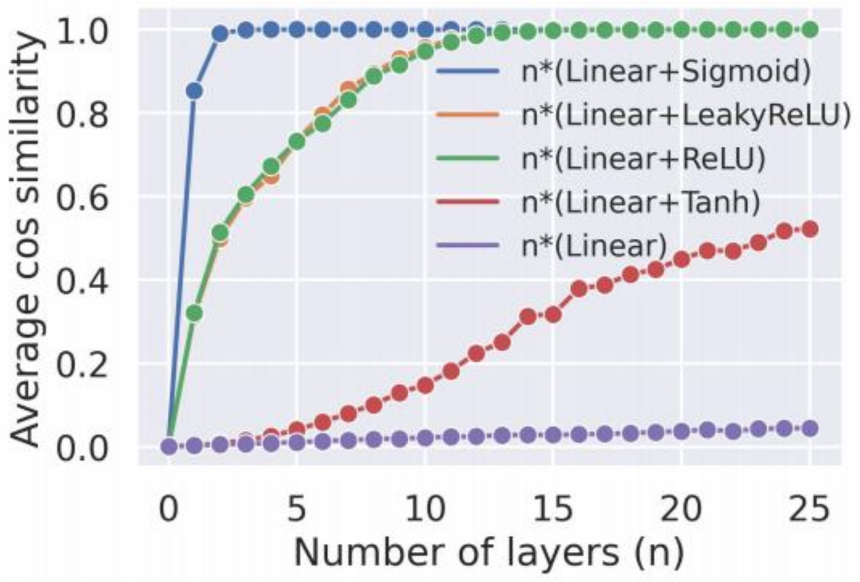
(c) UMAP visualization of embeddings of 25 randomly initialized models on real data (color indicates random seed)

cone effect:神经网络的嵌入空间在高维空间中是一个狭窄的锥形

(a)在不同设置中两两嵌入对之间的余弦相似度直方图。平均余弦相似度大于0，说明嵌入空间为一个窄锥。

(c)不同的随机初始化会让锥形朝向不同的方向

Cone Effect



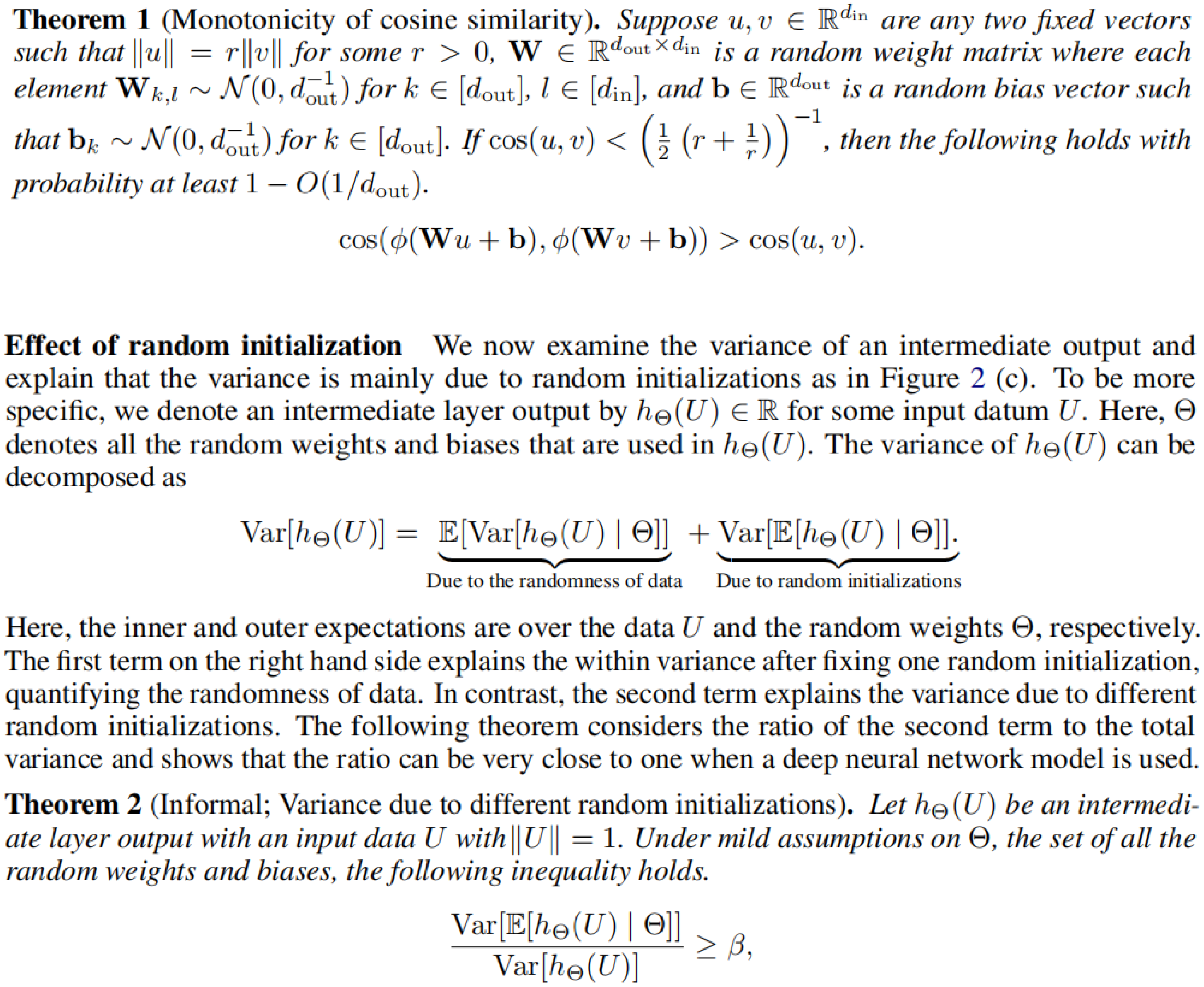
(b) Effects of nonlinear activation and depth

(b)非线性激活函数是导致锥形效应的主要原因，并且网络深度越深，锥越细

结论：1. 神经网络自带锥形效应；2. 这些锥形方向不同。

所以CLIP的 modaility gap 自随机初始化就存在

Theoretical analysis



Theorem 1 (Monotonicity of cosine similarity). *Suppose $u, v \in \mathbb{R}^{d_{\text{in}}}$ are any two fixed vectors such that $\|u\| = r\|v\|$ for some $r > 0$, $\mathbf{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ is a random weight matrix where each element $\mathbf{W}_{k,l} \sim \mathcal{N}(0, d_{\text{out}}^{-1})$ for $k \in [d_{\text{out}}]$, $l \in [d_{\text{in}}]$, and $\mathbf{b} \in \mathbb{R}^{d_{\text{out}}}$ is a random bias vector such that $\mathbf{b}_k \sim \mathcal{N}(0, d_{\text{out}}^{-1})$ for $k \in [d_{\text{out}}]$. If $\cos(u, v) < \left(\frac{1}{2}\left(r + \frac{1}{r}\right)\right)^{-1}$, then the following holds with probability at least $1 - O(1/d_{\text{out}})$.*

$$\cos(\phi(\mathbf{W}u + \mathbf{b}), \phi(\mathbf{W}v + \mathbf{b})) > \cos(u, v).$$

Effect of random initialization We now examine the variance of an intermediate output and explain that the variance is mainly due to random initializations as in Figure 2 (c). To be more specific, we denote an intermediate layer output by $h_\Theta(U) \in \mathbb{R}$ for some input datum U. Here, Θ denotes all the random weights and biases that are used in $h_\Theta(U)$. The variance of $h_\Theta(U)$ can be decomposed as

$$\mathrm{Var}[h_\Theta(U)] = \underbrace{\mathbb{E}[\mathrm{Var}[h_\Theta(U) \mid \Theta]]}_{\text{Due to the randomness of data}} + \underbrace{\mathrm{Var}[\mathbb{E}[h_\Theta(U) \mid \Theta]]}_{\text{Due to random initializations}}.$$

Here, the inner and outer expectations are over the data U and the random weights Θ, respectively. The first term on the right hand side explains the within variance after fixing one random initialization, quantifying the randomness of data. In contrast, the second term explains the variance due to different random initializations. The following theorem considers the ratio of the second term to the total variance and shows that the ratio can be very close to one when a deep neural network model is used.

Theorem 2 (Informal; Variance due to different random initializations). *Let $h_\Theta(U)$ be an intermediate layer output with an input data U with $\|U\| = 1$. Under mild assumptions on Θ, the set of all the random weights and biases, the following inequality holds.*

$$\frac{\mathrm{Var}[\mathbb{E}[h_\Theta(U) \mid \Theta]]}{\mathrm{Var}[h_\Theta(U)]} \geq \beta,$$

1.向量u、v经过带非线性激活函数的线形层后余弦相似度增加

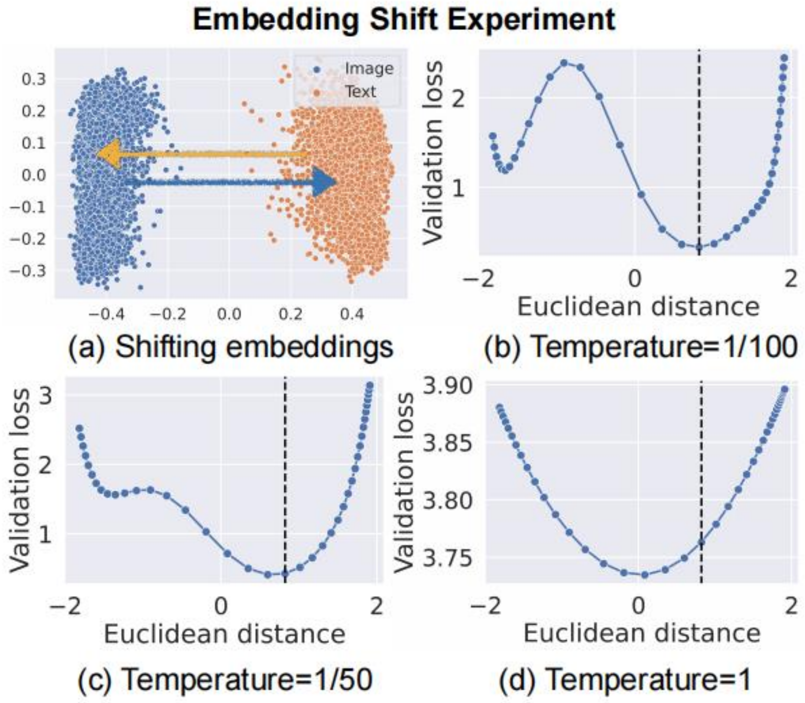
2.深度网络输出的方差主要来自于随机初始化

Contrastive learning preserves modality gap

Embedding Shift Experiment:
5000 pairs from MSCOCO Caption dataset

$$\vec{\Delta}_{\text{gap}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i - \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i \qquad \|\vec{\Delta}_{\text{gap}}\| = 0.82$$

$$\mathbf{x}_i^{\text{shift}} = \text{Normalize}(\mathbf{x}_i - \lambda\vec{\Delta}_{\text{gap}}), \quad \mathbf{y}_i^{\text{shift}} = \text{Normalize}(\mathbf{y}_i + \lambda\vec{\Delta}_{\text{gap}}).$$

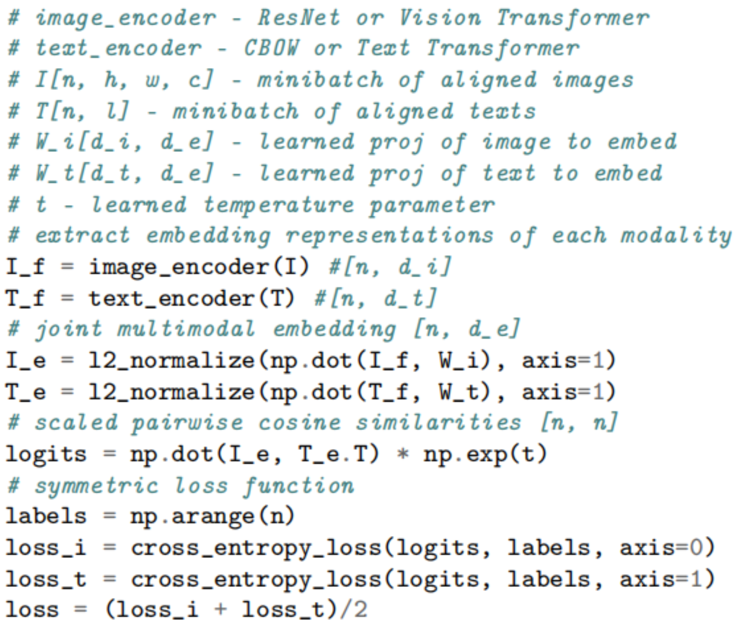

Embedding Shift Experiment

(a) Shifting embeddings

(b) Temperature=1/100

(c) Temperature=1/50

(d) Temperature=1



```
# image_encoder - ResNet or Vision Transformer
# text_encoder - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l] - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t - learned temperature parameter
# extract embedding representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T) #[n, d_t]
# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)
# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)
# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss = (loss_i + loss_t)/2
```

在不同的温度下微调CLIP:

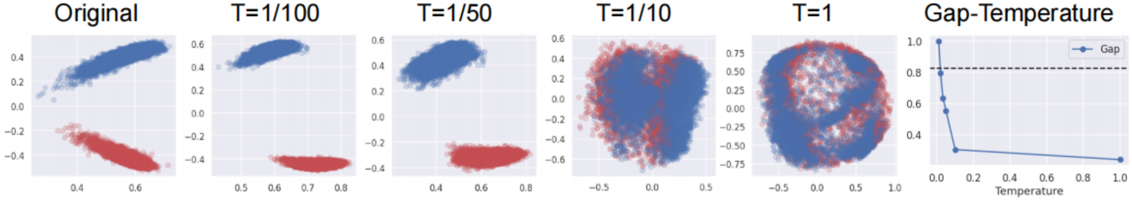


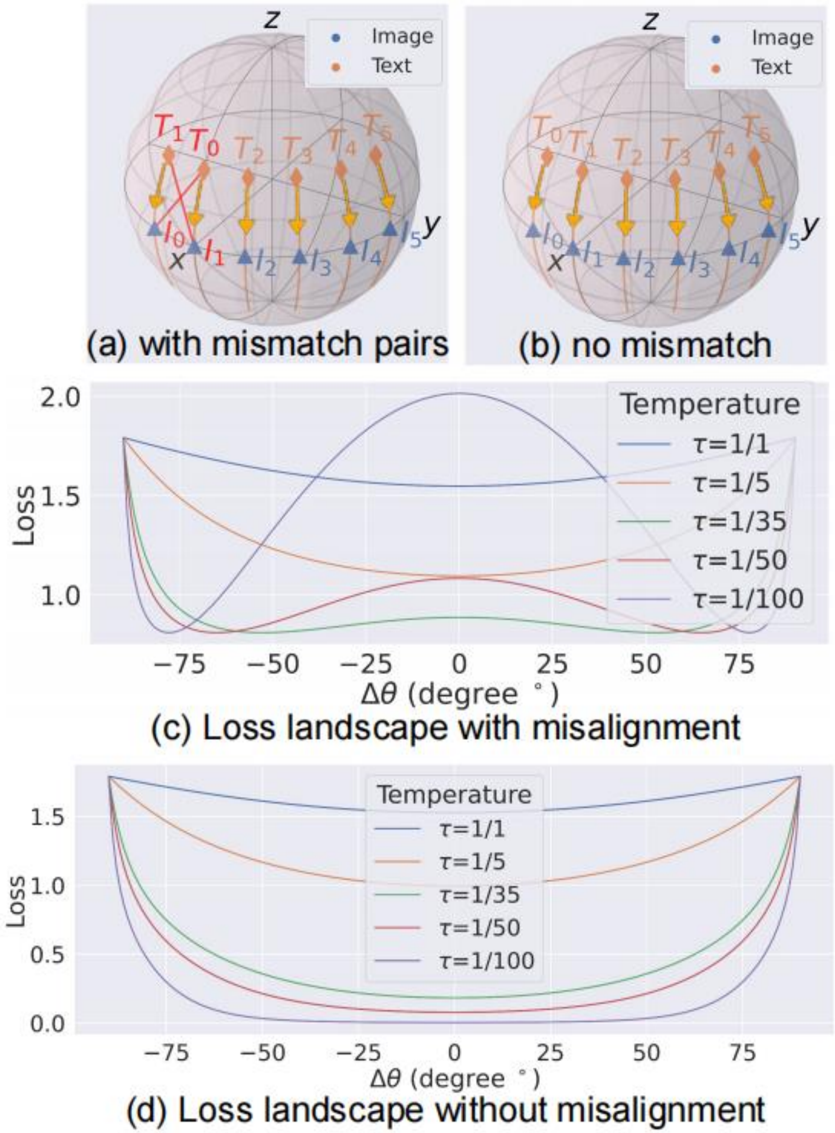
Figure 8: **Reduce the gap by fine-tuning with high temperature.** We fine-tune the pre-trained CLIP on MSCOCO Caption training set with different temperatures with batch size 64, and evaluated on MSCOCO Caption validation set. We found that a high temperature ($\tau \in \{\frac{1}{10}, 1\}$) in fine-tuning significantly reduces or closes the gap, while a low temperature does not. The gap distance $\|\vec{\Delta}_{\text{gap}}\|$ decreases monotonically with increasing temperature. The dashed line shows the original gap without fine-tuning.

Contrastive learning preserves modality gap

为什么温度可以影响gap？因为mismatched data

Simulating mismatched data

Additional Simulation Experiments



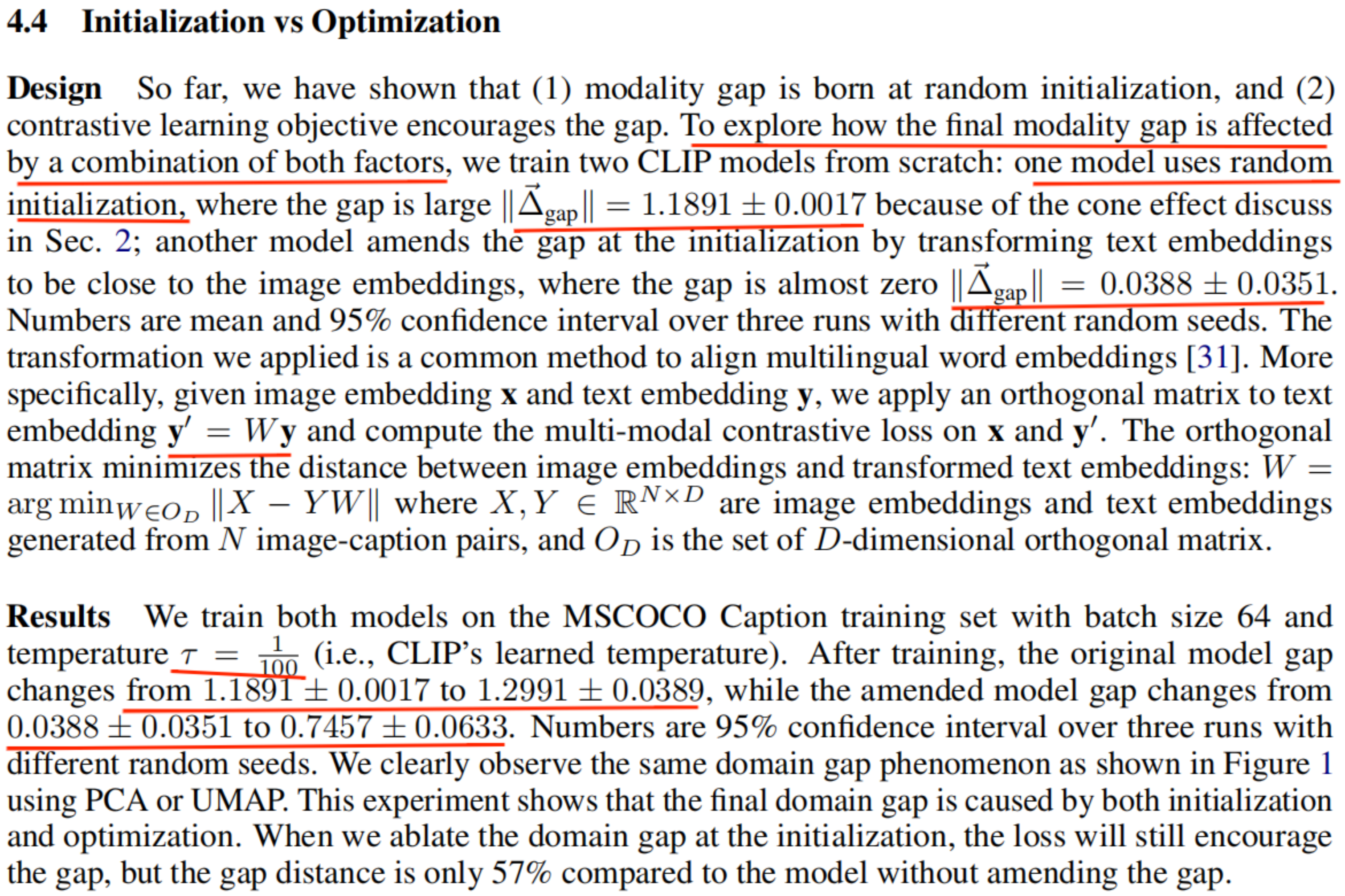
(a) with mismatch pairs

(b) no mismatch

(c) Loss landscape with misalignment

(d) Loss landscape without misalignment

Initialization vs Optimization



4.4 Initialization vs Optimization

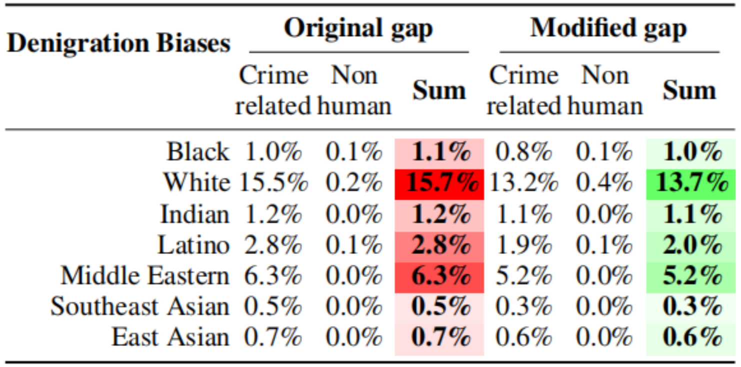
Design So far, we have shown that (1) modality gap is born at random initialization, and (2) contrastive learning objective encourages the gap. To explore how the final modality gap is affected by a combination of both factors, we train two CLIP models from scratch: one model uses random initialization, where the gap is large $\|\vec{\Delta}_{\text{gap}}\| = 1.1891 \pm 0.0017$ because of the cone effect discuss in Sec. 2; another model amends the gap at the initialization by transforming text embeddings to be close to the image embeddings, where the gap is almost zero $\|\vec{\Delta}_{\text{gap}}\| = 0.0388 \pm 0.0351$. Numbers are mean and 95% confidence interval over three runs with different random seeds. The transformation we applied is a common method to align multilingual word embeddings [31]. More specifically, given image embedding $\mathbf{x}$ and text embedding $\mathbf{y}$, we apply an orthogonal matrix to text embedding $\mathbf{y}' = W\mathbf{y}$ and compute the multi-modal contrastive loss on $\mathbf{x}$ and $\mathbf{y}'$. The orthogonal matrix minimizes the distance between image embeddings and transformed text embeddings: $W = \arg\min_{W \in O_D} \|X - YW\|$ where $X, Y \in \mathbb{R}^{N \times D}$ are image embeddings and text embeddings generated from N image-caption pairs, and O_D is the set of D-dimensional orthogonal matrix.

Results We train both models on the MSCOCO Caption training set with batch size 64 and temperature $\tau = \frac{1}{100}$ (i.e., CLIP's learned temperature). After training, the original model gap changes from 1.1891 ± 0.0017 to 1.2991 ± 0.0389, while the amended model gap changes from 0.0388 ± 0.0351 to 0.7457 ± 0.0633. Numbers are 95% confidence interval over three runs with different random seeds. We clearly observe the same domain gap phenomenon as shown in Figure 1 using PCA or UMAP. This experiment shows that the final domain gap is caused by both initialization and optimization. When we ablate the domain gap at the initialization, the loss will still encourage the gap, but the gap distance is only 57% compared to the model without amending the gap.

Modality Gap Implications

1.gap增大或减小都可能提升zero-shot准确率

2.gap可以影响模型对种族的偏见...



Denigration Biases	Original gap			Modified gap		
	Crime related	Non human	Sum	Crime related	Non human	Sum
Black	1.0%	0.1%	**1.1%**	0.8%	0.1%	**1.0%**
White	15.5%	0.2%	**15.7%**	13.2%	0.4%	**13.7%**
Indian	1.2%	0.0%	**1.2%**	1.1%	0.0%	**1.1%**
Latino	2.8%	0.1%	**2.8%**	1.9%	0.1%	**2.0%**
Middle Eastern	6.3%	0.0%	**6.3%**	5.2%	0.0%	**5.2%**
Southeast Asian	0.5%	0.0%	**0.5%**	0.3%	0.0%	**0.3%**
East Asian	0.7%	0.0%	**0.7%**	0.6%	0.0%	**0.6%**

Table 2: **Modifying the modality gap reduces biases for all races.** Number indicates the fraction FairFace images whose top-1 prediction is offensive. Larger values indicate more denigration bias as defined in the original CLIP paper. Increasing the gap from 0.82 to 0.97 reduces denigration harms consistently for all races.

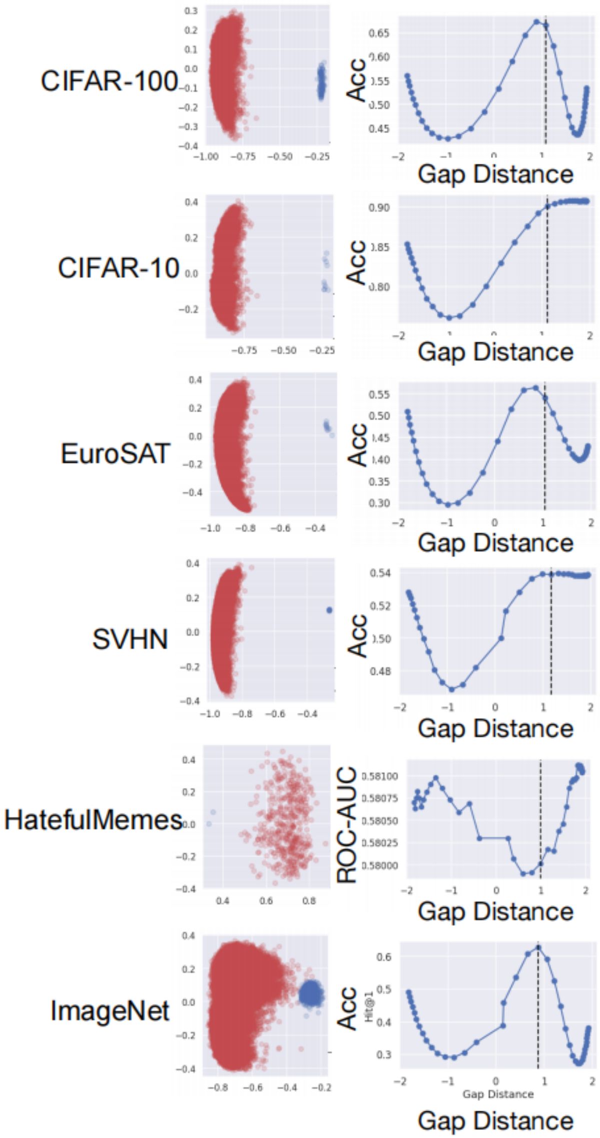
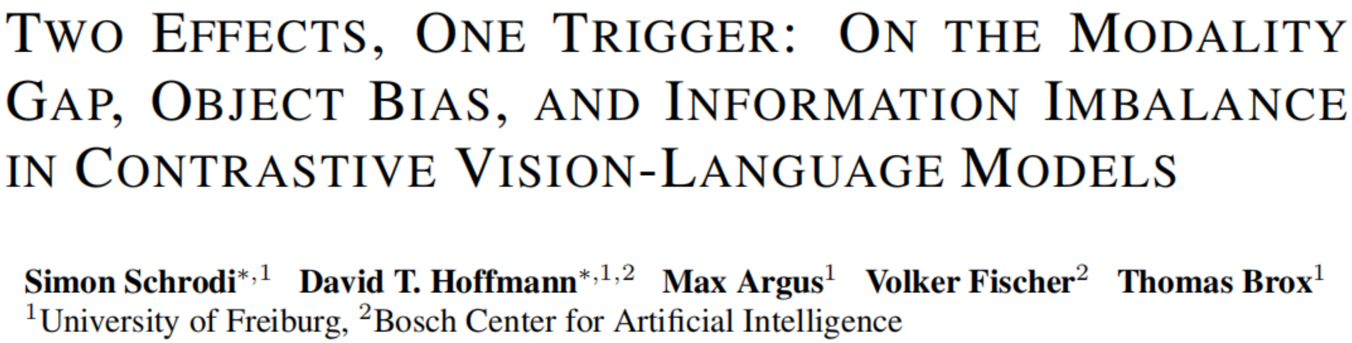


Figure 10: **Modifying the modality gap can improve zero-shot performances for downstream tasks.** Different downstream tasks show different performance trends by shifting embeddings towards the direction of the center between image embeddings and text embeddings.

Two Effects, One Trigger



Two Effects, One Trigger: On the Modality Gap, Object Bias, and Information Imbalance in Contrastive Vision-Language Models

Simon Schrodi[*,1] **David T. Hoffmann**[*,1,2] **Max Argus**[1] **Volker Fischer**[2] **Thomas Brox**[1]
[1]University of Freiburg, [2]Bosch Center for Artificial Intelligence

ICLR 2025 Oral

Two Effects, One Trigger

现象：
1.Modality Gap
2.Object bias：识别物体（object）如"狗"、"桌子"、"汽车"等时，表现得比识别这些物体的属性（attribute）如"红色的"、"光滑的"、"大的"等要好得多。

原因：
Information imbalance：availability of more information in one modality than the other

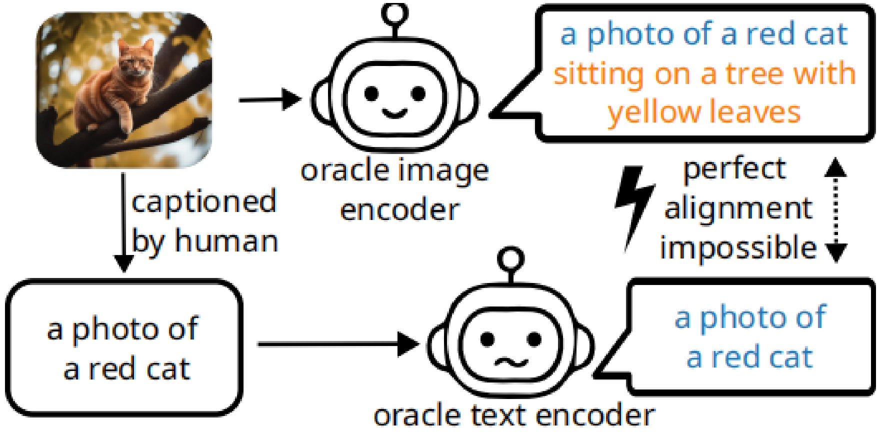


Figure 1: **Illustration of information imbalance** between images (top left) and captions (bottom left). This imbalance makes it even for an oracle image encoder virtually impossible to predict the content of a caption, leading to undesirable effects in contrastive training, such as the modality gap and object bias (see Section 6).

Relationship between the modality gap and downstream performance

新的衡量gap的指标RMG：考虑了matched image-text pairs，以及effectively used space

$$\mathrm{RMG} := \frac{\frac{1}{N}\sum_{i=1}^{N} d(\mathbf{x}_i, \mathbf{y}_i)}{\frac{1}{2N(N-1)}\left(\sum_{i,j=1;i\neq j}^{N} d(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i,j=1;i\neq j}^{N} d(\mathbf{y}_i, \mathbf{y}_j)\right) + \frac{1}{N}\sum_{i=1}^{N} d(\mathbf{x}_i, \mathbf{y}_i)} \quad ,$$

Relationship between the modality gap and downstream performance

直觉上人们觉得更小的modality gap会导致更好的performance
实验中似乎更大的modality gap有更好的performance？

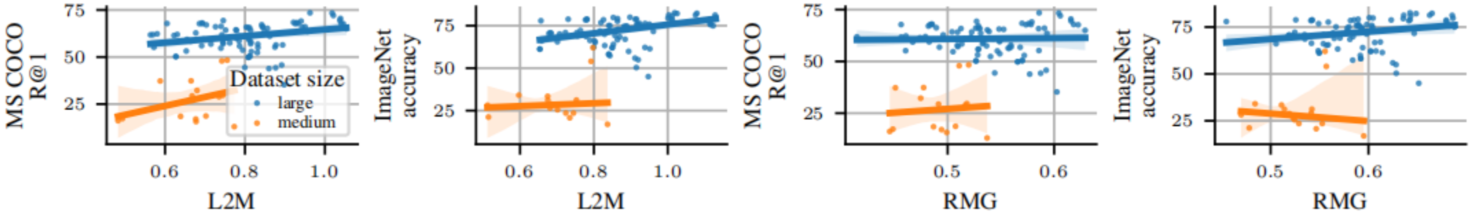
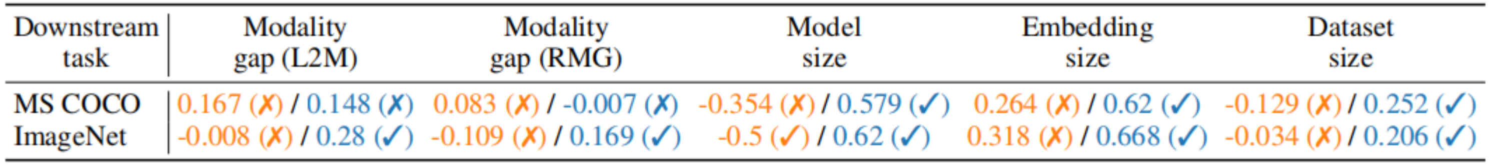


Figure 3: **Relation between modality gap (L2M & RMG, larger value → larger gap) and downstream performance** for a total of 98 contrastive VLMs pre-trained on medium- and large-scale datasets (each scatter point is a VLM). The plots indicate no to weak positive correlations between performance and modality gap (see the numbers in Table 1).

table1表明Model size、Embedding size、Dataset size对performance的影响更大，以至于掩盖了modality gap的影响。控制了这些因素后，发现modality gap与performance**似乎**存在微弱的负相关

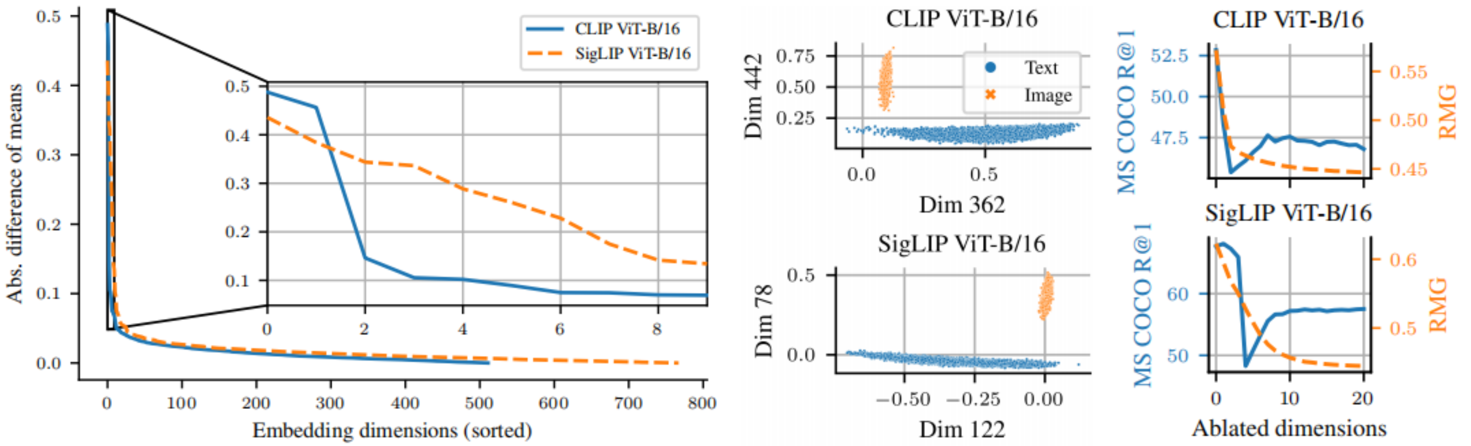
Table 1: **Kendall's τ rank correlation between downstream performance and various factors** for models trained on medium and large datasets. ✓denotes statistical significance ($p < 0.05$). Model, embedding, and dataset size correlate stronger with performance than the modality gap.



Downstream task	Modality gap (L2M)	Modality gap (RMG)	Model size	Embedding size	Dataset size
MS COCO	0.167 (✗) / 0.148 (✗)	0.083 (✗) / -0.007 (✗)	-0.354 (✗) / 0.579 (✓)	0.264 (✗) / 0.62 (✓)	-0.129 (✗) / 0.252 (✓)
ImageNet	-0.008 (✗) / 0.28 (✓)	-0.109 (✗) / 0.169 (✓)	-0.5 (✓) / 0.62 (✓)	0.318 (✗) / 0.668 (✓)	-0.034 (✗) / 0.206 (✓)

Takeaway 1: A larger modality gap has mild positive correlation with downstream performance. However, there is no indication that a larger modality gap leads to a better performance; rather, it suggests the presence of common confounders (e.g., model size).

Few embedding dimensions drive the modality gap



(a) Some dimensions have vastly different means. (b) Separating dims. (c) Removing dims.

Figure 4: **Few embedding dimensions separate the modalities.** Results on MS-COCO. (a) We plot the absolute difference in the means of each embedding dimension between the modalities. Most dimensions have similar means for both modalities, but for some the differences are huge. (b) Pairs of these high difference dimensions can perfectly separate the modalities (we show the ones with largest mean for each modality). (c) Successive removal of embedding dimensions based on the sorting of embedding dimensions from (a) leads to a sharp drop, followed by a partial recovery of downstream performance, while the modality gap gradually closes (similar results for L2M). See Appendix D.3 for results on ImageNet and the plots in (b) with the largest two dimensions of (a).

a)只有少数维度的差异很大

b）只用两个维度就可以完美区分不同模态的embdding

c）移除差异最大的维度，gap和性能急剧下降

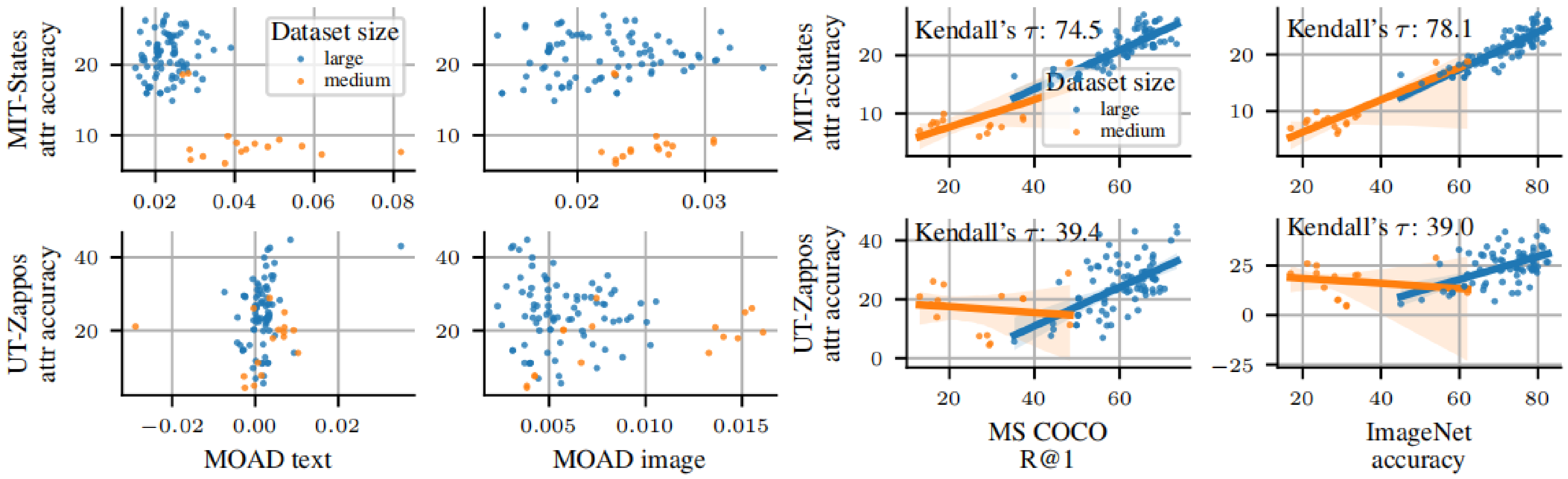
Takeaway 2: Few embedding dimensions drive the modality gap. We find that two dimensions suffice to perfectly separate the modalities.

Takeaway 3: Simple post-hoc approaches can close the modality gap but do not improve performance. One reason for this is that the modalities have different local neighborhoods.

Object bias

Matching Object Attribute Distance (MOAD)：用于衡量object bias大小
MOAD大于0表示偏向object、等于0表示无bias、小于0表示偏向attribute

$$\mathrm{MOAD}_{\mathrm{img}} := \frac{1}{2|O|}\sum_{\mathrm{obj}\in O}\left(\mathrm{sim}^{+}_{\mathrm{obj}} - \mathrm{sim}^{-}_{\mathrm{obj}}\right) - \frac{1}{2|A|}\sum_{\mathrm{att}\in A}\left(\mathrm{sim}^{+}_{\mathrm{att}} - \mathrm{sim}^{-}_{\mathrm{att}}\right),$$



(a) Object bias vs. downstream performance.

(b) Object vs. attribute performance.

Figure 5: **Object bias and performance on attribute tasks.** (a) We find a bias towards objects (positive MOAD values) but no correlation with attribute performance. We attribute this to the (b) positive correlation between performance improvements on object tasks and attribute tasks.

实验结果：

1）更大的数据集训出来的模型，MOAD更小

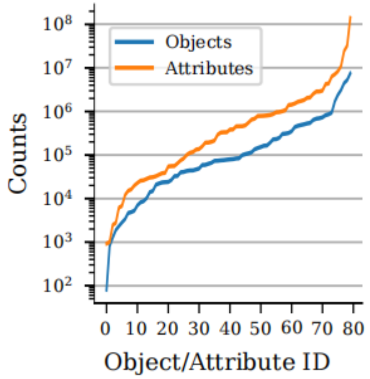
2）bias不会直接影响模型表现——模型在对象识别任务上的表现越好，属性识别能力通常也越强，表明对象任务和属性任务的提升是相互关联的

Takeaway 4: Contrastive VLMs trained on large-scale data tend to have a lower object bias than models trained on medium-scale. However, there is no clear relation between object bias and attribute performance. This can be attributed to the observation that performance improvements on object tasks correlate with improvements on attribute tasks.

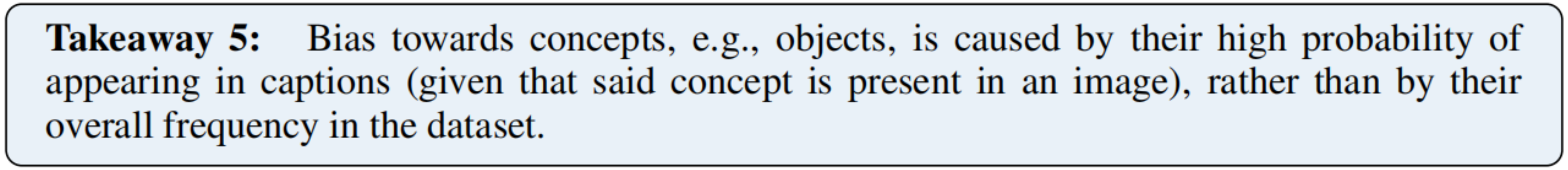
Object bias

Object bias产生的原因
不是因为数据集的captions中object出现的频率比attribute高（在LAION-2B caption中attibute出现的频次比objects高），而是因为条件概率p(word|image)
——Information imbalance

举例：含有白色的猫的图像，几乎所有caption都会包含cat，而只有一部分会包含white cat



(a) Object and attribute counts in LAION-2B.

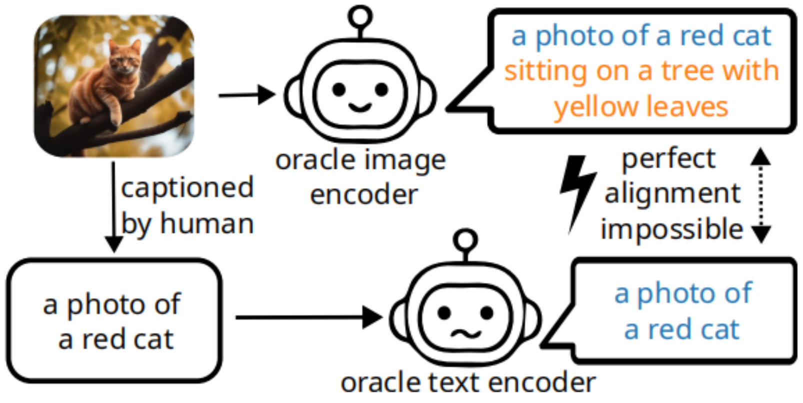


Takeaway 5: Bias towards concepts, e.g., objects, is caused by their high probability of appearing in captions (given that said concept is present in an image), rather than by their overall frequency in the dataset.

Two Effects, Information imbalance Trigger

The origin of the object bias.

Information imbalance导致编码器很难对齐它们的嵌入，因为它们无法知道在其他模态中有什么可用的信息。为了实现更好地对齐，最优编码器（尤其是图像编码器）使用更多关注caption中的object，用更少关注attribute。



The origin of the modality gap.

由于上述原因，loss的分子项（对齐）优化存在下界。于是通过增大分母来实现减小loss。

“In other words, the alignment term is bounded. With alignment being bounded(and optimized), the only way to further reduce the total loss is by maximizing uniformity (the denominators). Consequently, contrastive VLMs tend to focus more on **increasing the distance of non-matching pairs**, i.e., maximizing the uniformity”

因此，modality gap也是由于Information imbalance导致的。

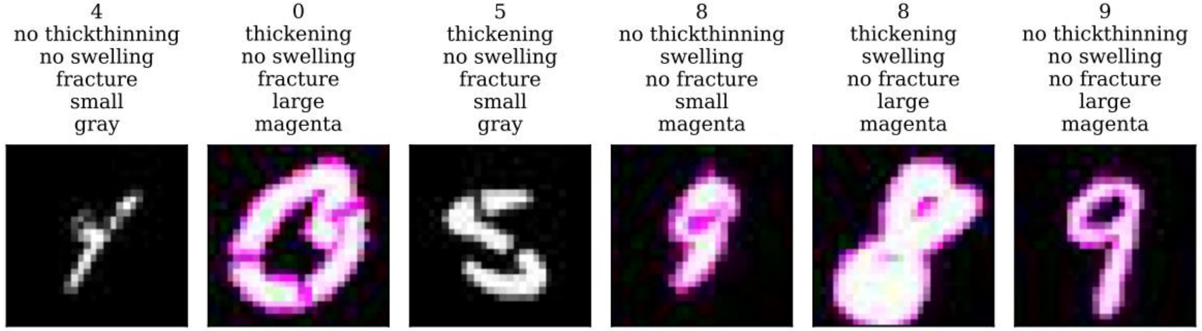
推测它通过使用少数维度来增加均匀性，对对齐的影响很小。

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2N}\sum_{i=1}^{N}\log\frac{\exp(\tau f_x(\mathbf{x}_i)^T f_y(\mathbf{y}_i)))}{\sum_{j=1}^{N}\exp(\tau f_x(\mathbf{x}_i)^T f_y(\mathbf{y}_j)))} - \frac{1}{2N}\sum_{j=1}^{N}\log\frac{\exp(\tau f_x(\mathbf{x}_j)^T f_y(\mathbf{y}_j)))}{\sum_{i=1}^{N}\exp(\tau f_x(\mathbf{x}_i)^T f_y(\mathbf{y}_j)))}$$

Experimental validation in a fully-controlled synthetic setting

通过控制caption中object和attribute出现的频率来控制Information imbalance

数据集：MINIST变体，如右
模型：small CLIP



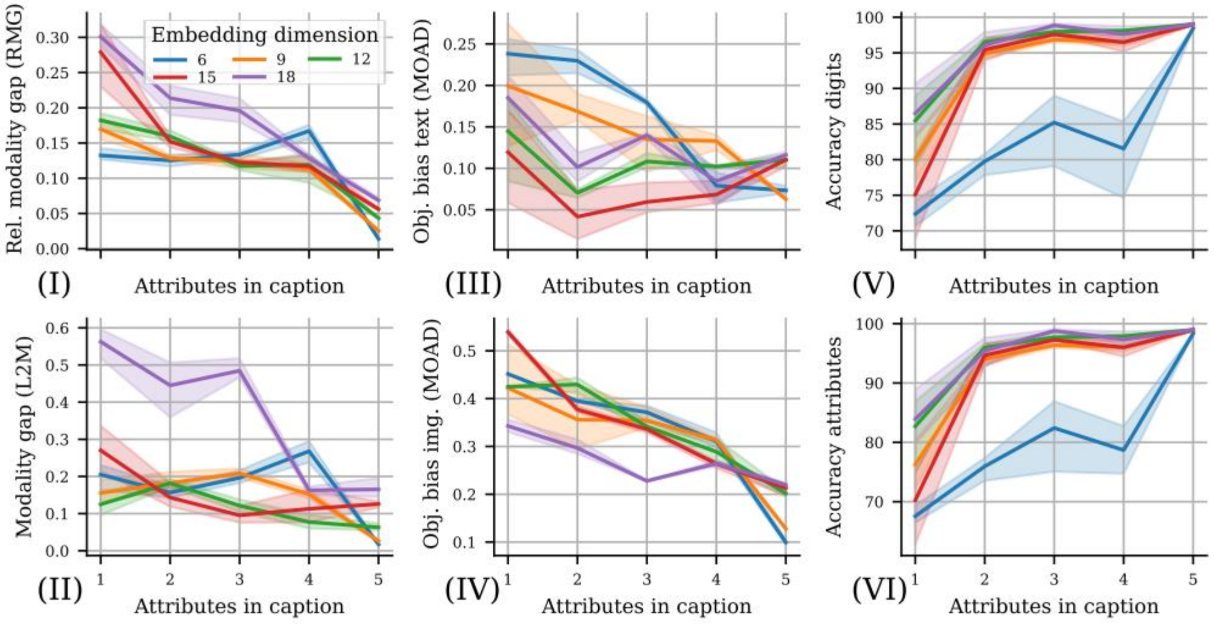
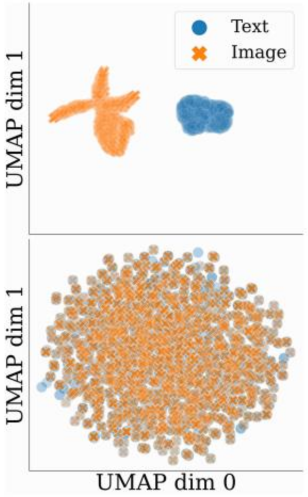


Figure 13: **Complete version of Figure 7a.** To study the influence of information imbalance between the modalities, we control the number of attributes present in the captions (the image is always affected by all attributes) in MAD. As the amount of information shared between the modalities increases, the modality gap (I-II) and bias towards objects reduces (III-IV), while downstream accuracy improves (V-VI).

训练前：

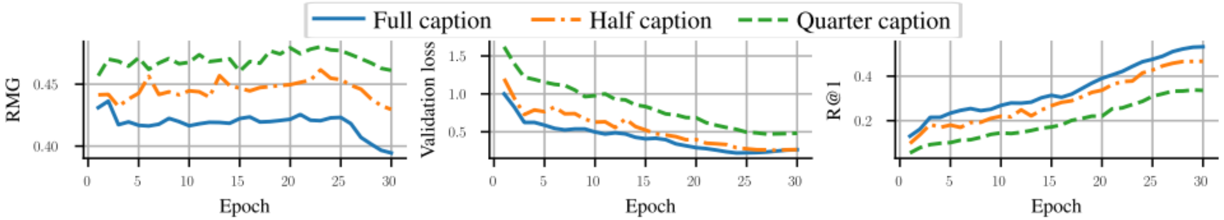
训练后：



(b) UMAP embeddings.

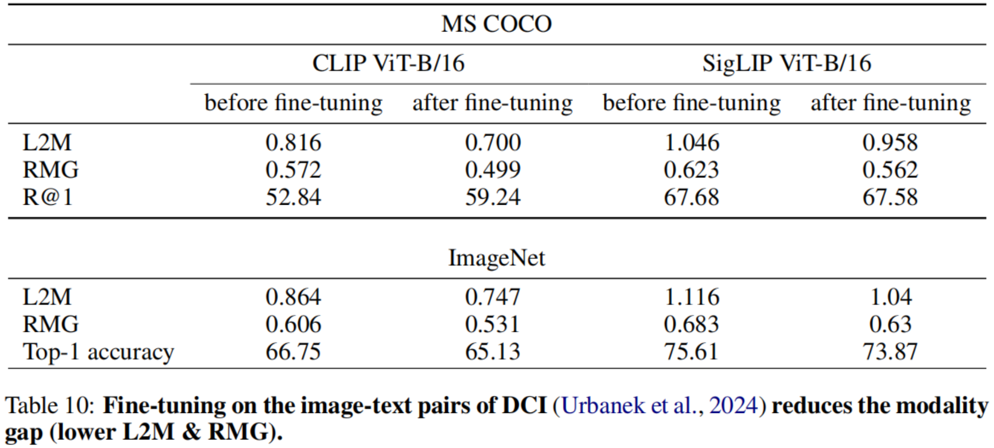
Experimental validation on real data

数据集：CC12M
模型：CLIP



(c) Larger information imbalance (shorter captions) → larger gap (real data).

数据集：Densely Captioned Images (DCI)
模型：CLIP、SigLIP



MS COCO				
	CLIP ViT-B/16		SigLIP ViT-B/16	
	before fine-tuning	after fine-tuning	before fine-tuning	after fine-tuning
L2M	0.816	0.700	1.046	0.958
RMG	0.572	0.499	0.623	0.562
R@1	52.84	59.24	67.68	67.58
ImageNet				
L2M	0.864	0.747	1.116	1.04
RMG	0.606	0.531	0.683	0.63
Top-1 accuracy	66.75	65.13	75.61	73.87

Table 10: **Fine-tuning on the image-text pairs of DCI** (Urbanek et al., 2024) **reduces the modality gap (lower L2M & RMG).**

Takeaway 6: An information imbalance between modalities leads to both modality gap and object bias. Reducing this imbalance decreases both the modality gap and object bias.

Is the modality gap a bug or a feature?

先在CC12M上使用full caption训练CLIP，再在1/4caption上进行微调

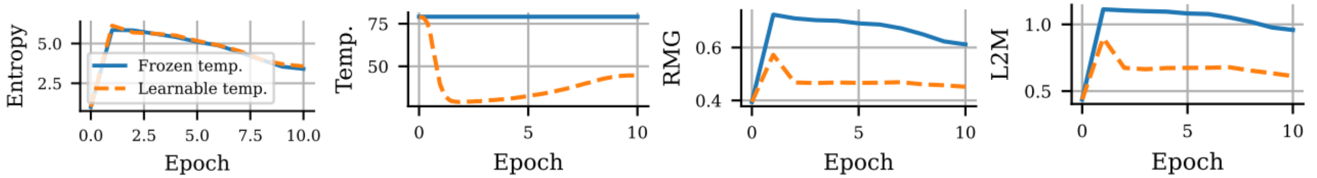
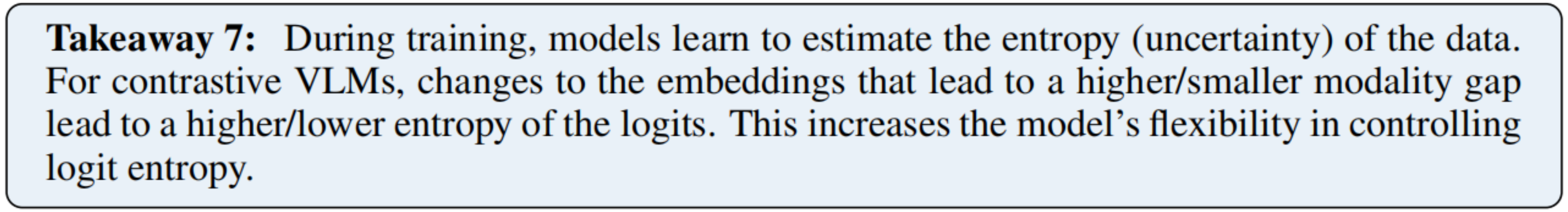


Figure 8: **A model trained with frozen temperature increases the modality gap more than a model with trainable temperature to achieve a similar logit entropy.**

模型学习估计数据的熵，改变温度参数和modality gap来改变学到的熵的估计——信息越不平衡，数据的熵越高，需要更大的gap或更小的温度参数。



Takeaway 7: During training, models learn to estimate the entropy (uncertainty) of the data. For contrastive VLMs, changes to the embeddings that lead to a higher/smaller modality gap lead to a higher/lower entropy of the logits. This increases the model’s flexibility in controlling logit entropy.