On the modality gap and the contrastive loss in multi-modal representation learning

Anonymous authors
Paper under double-blind review

Abstract

We study the *modality gap* in CLIP-style dual-encoder contrastive learning, where image and text embeddings remain misaligned despite being trained in a shared space. We argue that the gap is induced by a failure of the InfoNCE formulation with independent encoders. We conduct a uni-modal experiment with two independent encoders and identical initialization conditions and find that InfoNCE *actively* generates a gap at low temperatures. We provide a theoretical analysis of this phenomenon and show that the modality gap is indeed a mode-failure of InfoNCE, but only at low temperatures. We propose a simple modification called xNCE, which uses intermodal as well as intra-modality negative contrastive pairs. xNCE matches retrieval performance on MS-COCO while consistently reducing the gap even at low temperatures. Notably, xNCE improves zero-shot classification over the InfoNCE baseline across all benchmarks, whereas high-temperature InfoNCE and regularized InfoNCE both fail to do so, demonstrating that xNCE reduces the modality gap without sacrificing the discriminative geometry needed for transfer. Code availability: ...

1 Introduction

Contrastive learning has emerged as a central paradigm for self-supervised representation learning. Wu et al. introduced the view-invariant instance discrimination objective that underpins modern contrastive methods. Subsequent advances such as *SimCLR* (Chen et al.) demonstrated that strong data augmentation, large batch sizes, and a contrastive loss are sufficient to produce high-quality visual representations, while Bachman et al. broadened the theoretical foundation by defining contrastive learning as maximizing mutual information between augmentations. Parallel developments, including *CPC* (Oord et al.) and *MoCo* (He et al.), further refined contrastive objectives through predictive coding and momentum-based negative sampling. Building on these foundations, *CLIP* (Contrastive Language Image Pretraining) extended contrastive learning to the multimodal domain (Radford et al.).

Multimodal learning aims to build representations that jointly model information from heterogeneous data sources (e.g., images and language), enabling systems to exploit complementary cues and reason about shared semantics across modalities (Radford et al.). They furthermore showed that pairing images with natural-language descriptions in a contrastive learning paradigm yields highly transferable and zero-shot-capable models at scale.

CLIP uses one encoder per modality, trained using a contrastive loss function called Information Noise-Contrastive Estimation (InfoNCE). For a batch of images and captions, e.g., the loss aims to predict the correct caption for a given image and vice versa, based on pairwise cosine similarities. The authors show that such a pretraining objective produces embeddings with superior zero-shot performance compared to purely vision-based self-supervised models (Radford et al.).

Despite this success, CLIP-style dual-encoder models often exhibit a pronounced *modality gap*: image and text embeddings, while comparable via cosine similarity, tend to occupy distinct regions (or submanifolds) of the joint representation space rather than forming a fully mixed distribution. Previous analyses show that this separation has geometric and optimization roots, including inductive biases in initialization and training dynamics in contrastive learning that maintain a distance between modalities, influenced by factors such

as temperature, initialization, dimensionality, and data heterogeneity (Liang et al.; Shi et al., 2023; Fahim et al.; Role et al.; Huang et al.).

2 Related Work

Recent work has examined the *modality gap* in CLIP, i.e., the tendency of image and text embeddings to occupy distinct, non-overlapping regions of the shared representation space. Several explanations have been proposed in the literature. Liang et al. attribute the gap to modality-specific cone structures present at random initialization and optimization dynamics. They find that embeddings of ResNets, image encoders, and text transformer architectures form narrow cones on the unit sphere, which are highly separable across random initializations. The contrastive loss fails to close this gap during optimization. Under constrained optimization conditions (i.e. mismatched pairs), they find that when shifting the modalities toward each other, the contrastive loss increases, suggesting that the modality gap configuration is a local minimum of the contrastive loss at low temperatures. This local minimum disappears as the temperature increases or if the mismatched pairs are removed. Shi et al. (2023) identify conflicting properties in the InfoNCE loss function. They argue based on proof-of-concept experiments on the 3D unit sphere that the gap corresponds to local minima of the CLIP objective, making it difficult for optimization to align the modalities. Fahim et al. experimentally show that the modality gap in CLIP is neither a result of data heterogeneity nor initialization conditions. They experiment with InfoNCE in low dimensions and report that the modality gap does not arise. To circumvent the problem of the modality gap in high dimensions, they add regularization to the standard InfoNCE loss, which promotes close proximity of positive pairs as well as a uniform distribution of embeddings on the unit sphere, inspired by Wang & Isola (2020). The regularization reduces the gap and improves zero-shot performance on image classification benchmarks. Other work also shows that mitigating the gap can improve downstream performance. Role et al. provide principled metrics showing that reducing the gap can improve retrieval and zero-shot classification, while Huang et al. showed that preserving or compensating for the gap can benefit continuous learning. Eslami & de Melo (2024) proposes regularizing the contrastive loss using an intra-modal separation term, which again shows better retrieval and downstream utility. Oh et al. propose a new finetuning method, which defines hard negative samples by mixing modalities, resulting in more uniform embeddings. Other research points to deeper geometric and algorithmic factors: Levi & Gilboa find that image and text embeddings lie on distinct ellipsoidal shells encoding modality-dependent uncertainty. Collectively, these findings describe the modality gap as an inherent geometric and optimization feature of contrastive models. While some of these studies attribute the modality gap to properties of the contrastive objective, a mechanistic understanding of its emergence is still lacking. Current improvements to InfoNCE attempt to *mitigate* the gap by introducing additional loss terms with opposing effects (e.g. (Fahim et al.; Grassucci et al., 2025; Eslami & de Melo, 2024)), rather than fixing the divergent properties of InfoNCE itself.

3 Background

3.1 Contrastive learning via InfoNCE

Let $[x_i, y_i]_{i=1}^N$ be a minibatch of paired samples from two modalities. We denote the encoders by $f : \mathcal{X} \to \mathcal{U} \in \mathbb{R}^{N\times D}$ and $g : \mathcal{Y} \to \mathcal{V} \in \mathbb{R}^{N\times D}$, and write $\mathbf{u}_i = f(x_i)$ and $\mathbf{v}_j = g(y_j)$. We assume (x_i, y_i) to be positive pairs and (x_i, y_k), $k \neq i$ to be negative pairs. Let $a_{ij} \in [-1, 1]$ be the cosine similarity between $\mathbf{u}_i$ and $\mathbf{v}_j$, where a_{ii} and a_{ik} are similarity scores for positive and negative pairs, respectively. For a fixed anchor x_i (i.e., fixed $\mathbf{u}_i$) and its targets $\{\mathbf{v}_j\}_{j=1}^N$, the softmax probability $P_i(\tau)$ with temperature $\tau > 0$ is

$$P_i(\tau) = \frac{\exp\big(a_{ii}/\tau\big)}{\sum_{j=1}^N \exp\big(a_{ij}/\tau\big)} \ . \tag{1}$$

The InfoNCE loss from $\mathcal{U} \to \mathcal{V}$ is the mean over the negative log-probabilities $P_i(\tau)$:

$$\mathcal{L}_{\mathcal{U}\to\mathcal{V}}(\tau) \ = \ \mathbb{E}\left[-\log P_i(\tau)\right] \ . \tag{2}$$

The equivalent loss term $\mathcal{L}_{\mathcal{V}\to\mathcal{U}}$ is obtained by switching the indices i and j in eq. (1) and eq. (2), and the final loss value is the average of the two. The loss decreases as we increase the similarity of positive pairs a_{ii} and decrease the similarities of negative pairs a_{ik}. Typically, all embeddings are ℓ_2-normalized, and similarity is computed using the dot product. Wang & Isola (2020) investigate the properties of representations learned via a contrastive loss, and propose two quantities: *alignment* (closeness of features from positive pairs) and *uniformity* (of the distribution of features on the hypersphere). Alignment is achieved by maximizing the similarity of positive pairs, and uniformity is achieved by maximizing the entropy of the feature distribution on the hypersphere. They show that for a *uni-modal* setup, i.e. where $\mathbf{u}$ and $\mathbf{v}$ are processed by the same encoder, these properties are asymptotically optimized, and correlate strongly with downstream performance.

3.2 The Temperature Parameter

The temperature τ scales the logits in the softmax and thereby controls the sharpness of the distribution $P_i(\tau)$. Smaller values of τ amplify small similarity gaps, resulting in a winner-takes-all probability distribution, while larger values of τ attenuate those gaps and produce a more uniform probability distribution. Common values for τ are in the range $[0.01, 1]$, with CLIP having a learnable temperature parameter, which is initialized as 0.07 and converging to 0.01 during training (Radford et al.).

3.3 Mode-Failure due to Dual-Encoders

In the uni-modal scenario, $\mathcal{U}$ and $\mathcal{V}$ are outputs of a single parametrized function. Updating the parameters of this function will affect both $\mathcal{U}$ and $\mathcal{V}$. The only way to accurately classify a positive pair from negative pairs is to learn discriminative features of $\mathbf{u}$ and $\mathbf{v}$, which are invariant to noise factors (e.g. augmentations). The contrastive loss asymptotically optimizes for alignment of positive pairs and uniformity of the feature distribution on the hypersphere (Wang & Isola, 2020). In a multi-modal scenario, $\mathcal{U}$ and $\mathcal{V}$ are outputs of independently parametrized functions. Although the parameters are optimized under a joint objective, the parameter sets do not share weights. Thus, while both encoders may learn equivalently good discriminative features, it is not guaranteed that positives are mapped to the same region on the hypersphere, as they are encoded independently. While the numerator in eq. (1) encourages alignment of positives, the denominator can be satisfied by moving $\mathbf{u}$ and $\mathbf{v}$ far from each other, providing a trivial objective and explaining the phenomenon of the modality gap. Thus, we hypothesize that the uniformity term in the multi-modal InfoNCE objective is no longer only an information preservation term on $\mathcal{U}$ and $\mathcal{V}$, but also a degenerative term with a divergence objective on $\mathcal{U}$ and $\mathcal{V}$. We believe that by eliminating this degenerative divergence objective, we can eliminate the modality gap and improve the performance of the model.

This motivates the following two hypotheses.

Hypothesis 1 (H1) Induced gap by multimodal InfoNCE formulation. *The modality gap is the result of an* active inductive bias *from the InfoNCE formulation with independent encoders and the induced optimization dynamics, rather than being a byproduct of data idiosyncrasies, initialization, dimensionality or temperature.*

Hypothesis 2 (H2) Mitigating the gap by sampling intra- and inter-modality pairs. *The modality gap can be reduced by sampling negative pairs a_k interchangeably from both modalities, which removes the divergence objective of two modalities. We call this adaptation* "Mixed Noise-Contrastive Estimation"*, or* xNCE.

4 Mixing modalities - xNCE

We propose a simple and efficient modification to the InfoNCE formulation of Radford et al.. Our method builds on the following assumption for an ideal multimodal latent space: Two modalities are fully aligned and occupy the same space if the modality type of a latent variable becomes unidentifiable. If so, we can use the latent variables from both modalities interchangeably in the loss formulation in eq. (2). For a given anchor $\mathbf{u}_i$, there exists only one positive $\mathbf{v}_i$. However, we can use any $\mathbf{u}_k$ or $\mathbf{v}_k$, $k \neq i$, as a negative pair. We refer to negative pairs from the same modality, i.e. $\mathbf{u}_i$ and $\mathbf{u}_k$, as *intramodality* negatives and pairs from different modalities, i.e. $\mathbf{u}_i$ and $\mathbf{v}_k$, as *intermodality* pairs. The procedure is depicted in fig. 1 (right) and a numpy like pseudo-code is shown in fig. 2. When calculating the pairwise cosine similarity a_{ij} for mixed embeddings,

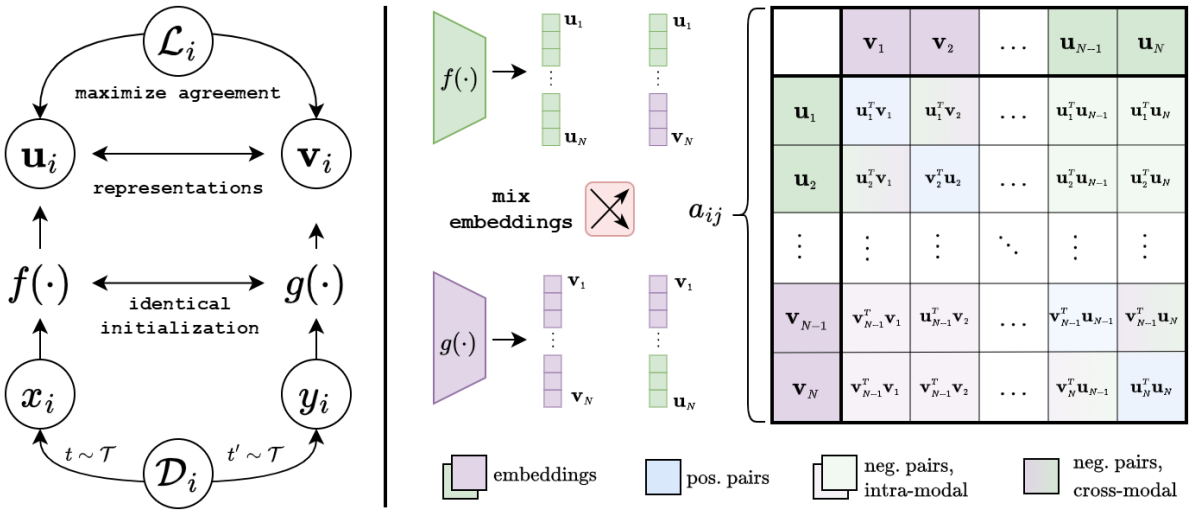

Figure 1: **Left**: A *SimCLR* like investigation of the multimodal InfoNCE loss. We generate two views of the same datapoint from the same family of augmentations $t \sim T$ and $t' \sim T$, and pass the two views to identically initialized, independent encoders $f(\cdot)$ and $g(\cdot)$ (i.e. no weight sharing). Adapted from Chen et al. **Right**: Our proposed method xNCE to avoid modality shifts. Rather than only sampling cross-modal negative pairs, we include intra- and cross-modality negative pairs. Adapted from Radford et al..

the diagonal will remain the same. The off-diagonal elements will consist of $N \times r_{mix}$ intra-modality pairs and $N \times (1 - r_{mix})$ inter-modality pairs, where $\mathrm{r_{mix}}$ is the mixing ratio applied to $\mathcal{U}$ and $\mathcal{V}$. Because the final loss is a symmetric cross entropy loss over the similarity scores, the mixing is also symmetric around $r_{mix} = 0.5$. Most importantly, this formulation removes the degenerate divergence objective of InfoNCE. In xNCE, the denominator of eq. (2) includes intermodality negatives, leading to an increase in loss as the modalities drift apart.

4.1 Unimodal: Image-to-Image using MNIST

We hypothesize that the modality gap is the result of the multi-modal InfoNCE loss formulation, rather than initialization and data idiosyncrasies (Hyp. 1). We therefore set up a *SimCLR* (Chen et al.) like experiment, where two views are generated from the same source using the same family of augmentations fig. 1 (left). In contrast to Chen et al., we use two independent encoders, i.e. without weight sharing or momentum updates, as is the case in a multi-modal setup. We use the same architecture and number of parameters for both encoders. Importantly, we initialize both encoders with the same weights, which ensures identical embeddings at initialization for the same input. If a gap is observed, it would indicate that the gap is not a passive byproduct but an active element in the optimization dynamics of InfoNCE. A short overview of this setup is given in fig. 1.

Dataset. We use the *digit* split of the Extended MNIST (EMNIST) dataset (Cohen et al.), which contains ten handwritten digit classes. To construct contrastive pairs, we generate two independently augmented views of the same image. Training augmentations include random rotation ($\pm 45°$), random resized cropping (scale $\in [0.7, 1.0]$), color jitter, and additive Gaussian noise with standard deviation $\sigma = 0.3$. Both views are normalized with the EMNIST channel statistics (μ=0.1307, σ=0.3081) and passed to its respective encoder. Thus, 'modality' in the context of the Image-to-Image contrastive learning refers to the output of each encoder.

Model. We use a minimalistic vision transformer (Dosovitskiy et al., 2020) tailored for 28×28 grayscale images, with patch size 4, embedding dimension d=48, 4 attention heads, 4 transformer layers, and an MLP ratio of 4. The transformer employs pre-norm residual blocks with GELU activations and 2D sinusoidal

```
# Inputs:
# I - Image emb. [n, d_i]
# T - Text emb. [n, d_t]
# r_mix - mixing ratio, float
# temp - temperature param.
# ce_loss - Cross Entropy Loss

# mix modalities
bs = I_e.size(0) # batch size
mask = np.arange(bs) < int(bs*r_mix)
I_mix = np.where(mask, T, I)
T_mix = np.where(mask, I, T)

# scaled pairwise cos. sim.
logits = np.dot(I_mix, T_mix.T)
logits /= temp

# symmetric loss function
labels = np.arange(n)
loss_i = ce_loss(logits, labels)
loss_t = ce_loss(logits.T, labels)
loss = (loss_i + loss_j) / 2
```

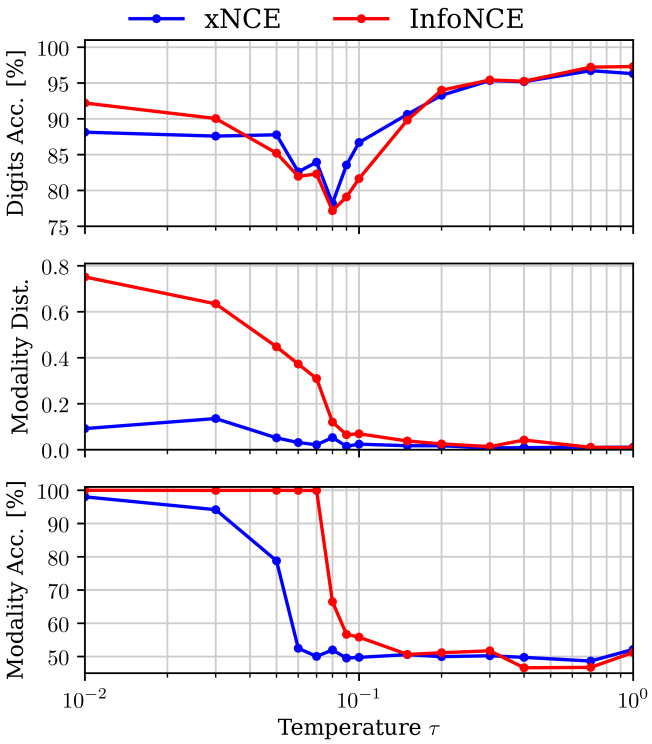

Figure 2: **Left**: The proposed xNCE loss function. Numpy like pseudo-code for mixing image and text embeddings. **Right**: Model behaviour for different temperature values. We measure classification accuracy of a linear probe, modality distance as euclidean distance of modality centroids and modality classification accuracy using a linear probe.

positional encodings. A learnable `[CLS]` token is prepended to the patch sequence, and its final-layer representation serves as the image embedding. Embeddings are ℓ_2-normalized before computing the contrastive loss.

Loss. We train several models with the InfoNCE and xNCE objective. We use a fixed mixing ratio of r_{mix}=0.5 for all MNIST experiments. We keep the temperature τ a fixed parameter and iterate on in total 14 values, ranging from 0.01 to 1.0.

4.2 Multimodal: Image-to-Text using COCO

For a truly multimodal task, we use the Microsoft Common Object in Context Dataset (COCO) (Lin et al.) and finetune CLIP (Radford et al.). COCO is widely used for benchmarking vision language models, and is not included in the pretraining corpus of CLIP. All results presented in this paper are on the Kaparthy split test set (Karpathy & Fei-Fei, 2015), obtained from Senior. COCO has five captions per image. For training, we choose a random caption among the five, whereas for validation and testing, we use the first caption.

Dataset. The COCO 2017 dataset (Lin et al.), which pairs images with natural language captions. Each image is associated with multiple captions; during training, a caption is sampled uniformly at random per image, while at validation time the first caption is used deterministically. Images are preprocessed using the standard CLIP preprocessing pipeline (Radford et al.). Captions are tokenized using CLIP's byte-pair encoding tokenizer. Object category annotations from the COCO instances set provide a multi-label classification signal over 80 object categories, encoded as multi-hot vectors.

Model. We finetune both vision and text encoders. We use the CLIP ViT-B/16 obtained from CLI.

Loss. InfoNCE serves as the baseline. As CLIP converges to a temperature of 0.01, we set this to be the initial value of τ. We used both a learnable temperature parameter where we clip the value of τ if < 0.01, equivalent to Radford et al., as well as a cosine annealed temperature parameter with $\tau_{max} = 0.7$. For all base experiments, we do not use a contrastive head (Oord et al.) as part of the loss function. The loss is calculated on the output of the encoders. We additionally compare against regularized InfoNCE (Reg) from Fahim et al., which augments the InfoNCE objective with explicit alignment, cross-modal uniformity, and unimodal uniformity terms, inspired by the work of Wang & Isola (2020). For implementation details, see section B. For xNCE, we do not know the optimal temperature value. We tried several initial temperature values in the range of 0.01 to 0.05 and found that τ=0.02 worked marginally better. We trained with three different mixing ratios: $\mathrm{r_{mix}}$ =0.01, 0.05, and 0.5.

4.3 Evaluation

To assess the structure of the learned representations, we train linear probes evaluating (i) *the separability* of the modality and (ii) *the downstream utility*. All probes are trained concurrently with the main contrastive objective on the training data. Crucially, all probes serve purely as diagnostic tools without influencing representation learning. For (i), we train a single linear layer using a binary cross entropy loss to predict the modality. As for (ii), we train the probes per modality and report the average score. For MNIST experiments, we use the digit class as target (single-label, C=10) and report classification accuracy. For COCO, we use the instances represented in the image (multi-label, C=80), available through the object-detection task in COCO and report the multilabel average precision. In addition, we measure the *alignment* and *uniformity* properties proposed by Wang & Isola (2020). Alignment is defined as the expected ℓ_2-distance between positive pairs, and uniformity is defined as the expected pairwise Gaussian potential, measured for each modality separately. For Image-to-Text experiments, we evaluate the image-to-text (I2T) and text-to-image (T2I) retrieval recalls at ranks 1, 5, and 10 (R@1, R@5 and R@10, respectively). To evaluate the transferability of the learned image representations, we also conduct zero-shot classification on four image-only benchmarks: CIFAR-10 (N=10k, C=10), CIFAR-100 (N=10k, C=100), Caltech-101 (N≈9.7k, C=100), and ImageNet (N=50k, C=1000). For each dataset, we follow Radford et al. and encode class names using the prompt template *"A photo of a {class}"* and classify images by nearest neighbour in embedding space under cosine similarity. We report top-1 and top-5 accuracy. No fine-tuning or adaptation is performed; the encoder weights are used as-is after COCO training.

4.4 Optimization

All models are trained using AdamW with a weight decay of 0.04 and a three-phase learning rate schedule: *Phase 1 - probes only*: Near-zero learning rate, allowing the linear probes to warm up on randomly initialized representations before the encoders begin learning. Probe optimizers use a fixed learning rate throughout training. *Phase 2 - warmup*: Linear warmup to the base learning rate. *Phase 3 - decay*: Cosine annealing to $1e-2\times$ the base learning rate. We used a single NVIDIA Ampere GPU for the Image-to-Image experiments and a single NVIDIA H100 GPU for the Image-to-Text experiments. All key hyperparameters for both Image-to-Image (section 4.1) and Image-to-Text experiments (section 4.2) are listed in tables 4a and 4b in section A.

5 Results

5.1 Unimodal: Image-to-Image using MNIST

We investigate how mixing modalities in the InfoNCE loss formulation, such that negative similarities include both intra- and inter-modality pairs, affects the embedding space. We name this adaptation xNCE. We measure digit and modality classification accuracy via a linear probe, as well as the modality centroid distance. fig. 2 compares these measures for InfoNCE and xNCE over a range of temperature values.

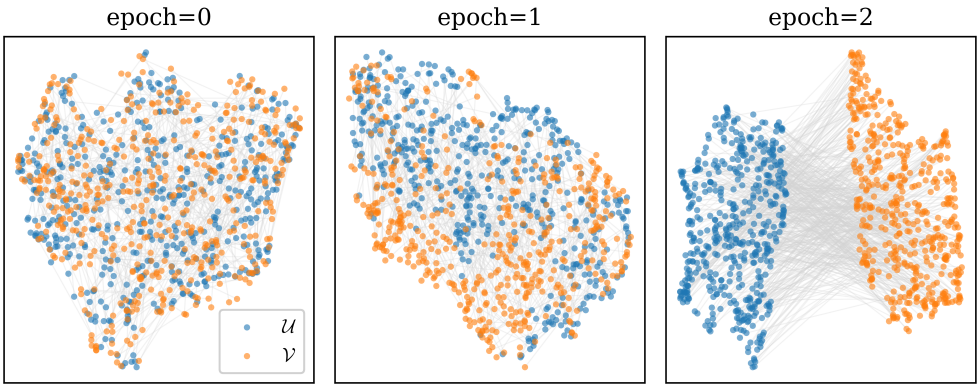

Figure 3: Two dimensional UMAP projection of MNIST embeddings for training using InfoNCE at low temperature (τ=0.01). Points are coloured by their respective encoder. Embeddings overlap at initialization as both encoders are initialized with identical weights, and get separated immediately after.

Digit classification accuracy. We find a lower classification accuracy for a temperature of $0.03 < \tau < 0.15$, and a higher classification accuracy towards low and high temperatures $\tau < 0.03$ and $\tau > 0.2$. The best digit separability of 97.2% and 96.7% is achieved at τ=1.0 and τ=0.7 for InfoNCE and xNCE, respectively. The margin of InfoNCE and xNCE is slightly larger at lower temperatures, where classification accuracy is 92.2% and 88.1% for τ=0.01, respectively.

Modality Distance. For InfoNCE, we find that the modality distance is 0.75 at τ=0.01, monotonically decreasing to a value of 0.07 at τ=0.09, and approximately constant for $\tau \geq 0.09$. Compared to InfoNCE, the modality gap for xNCE is generally less for all values of τ, starting at 0.09 and 0.13 for τ=0.01 and τ=0.03 and converging to less than 0.05 for $\tau \geq 0.05$.

Modality classification accuracy. The modality accuracy for InfoNCE is at 100% for low temperatures and ≈50% at high temperatures, with a steep transition at $\tau \approx 0.09$. The xNCE shows high classification accuracies for lower temperatures but slightly below 100% and ≈ 50% for $\tau \geq 0.07$, with a more subtle transition.

Latent geometry under different temperatures. Figure fig. 4 (left) shows the cosine similarity scores at different temperatures τ. For InfoNCE, low temperatures result in distributions with positive similarities $a_{ii} \ll 1$ and negative similarities $a_{ik} \gg 0$. For higher temperatures, the distribution is stretched and slightly skewed, where positive similarities are close to 1 and negative similarities range from -1 to 1 with a mean of ≈ 0. The distribution of pairwise similarities for low temperatures indicates that $\mathcal{V}$ is located in a narrow cone far from the query $\mathbf{u}_i$, which aligns with the observed modality gap. When using xNCE, the modality gap disappears, as the positive similarities are very close to 1. Mixing modalities has little or no effect on the similarity distribution for higher temperatures. Temperature also affects the within-modality distribution (fig. 4, right). Although the linear separability of classes for low and high temperatures is similar (fig. 2, right), the classes occupy a much narrower space for low temperatures (e.g. τ=0.01) than they do for high temperatures (e.g. τ=1.0).

5.2 Theoretical Analysis of the Temperature in InfoNCE

We analyze the limit $\lim_{\tau \to 0^+} \mathcal{L}_i$ (Equation (2)), as this is where we observe the modality gap. First, we establish a formal definition of the phenomenon. As Equation (2) acts on the similarity scores a_{ij}, we define the modality gap as follows:

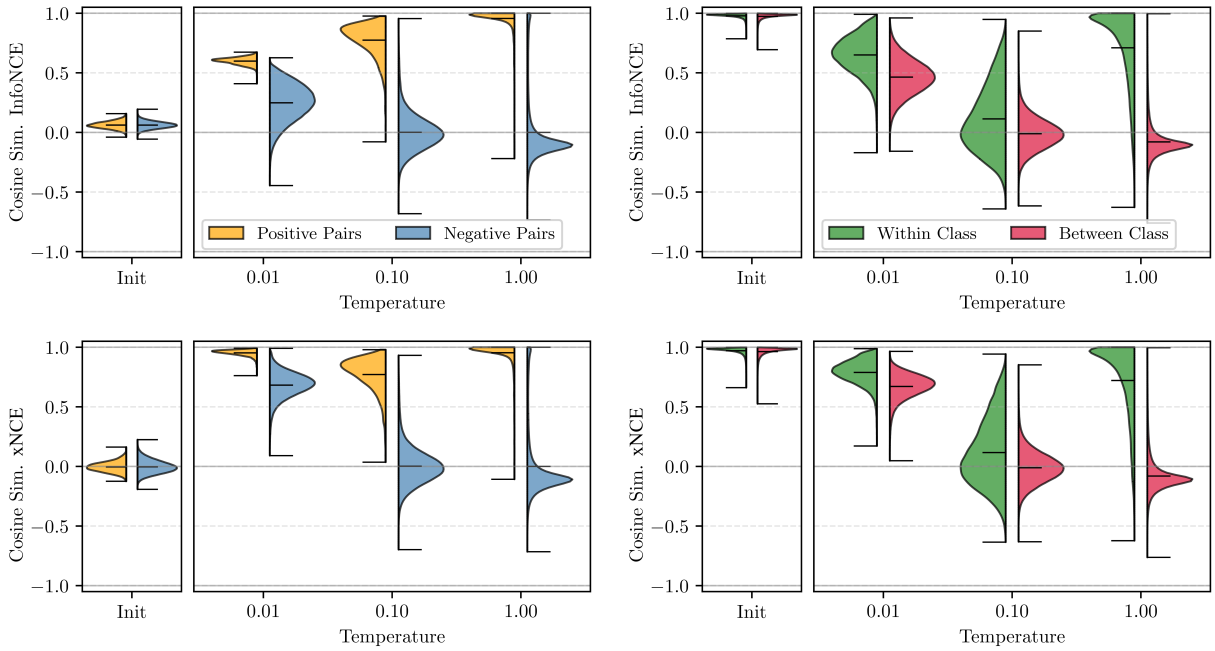

Figure 4: Distribution of similarity scores over different temperatures for the original InfoNCE loss (**top**) and our xNCE loss (**bottom**). **Left**: Cosine similarity scores for positives pairs (a_{ii}) and negatives pairs (a_{ik}). Low temperatures show short tails in the distribution of negative pairs, whereas higher temperatures show lower variance but longer tails. xNCE shows higher scores for in τ=0.01, but behaves similarly for τ=0.1 and τ=1.0. **Right**: Cosine similarity scores conditioned on class labels. The distributions are furthest apart for τ=1.0. No or very little difference between InfoNCE and xNCE.

Let $\mathcal{M} : \mathcal{X} \times \mathcal{Y} \to \mathcal{U} \times \mathcal{V}$ denote a parametrized function, which maps the inputs x and y to their corresponding latent variables $\mathbf{u}$ and $\mathbf{v}$. For pairwise cosine similarities $\mathbf{A} \in \mathbb{R}^{N \times N}$ with elements $a_{ij} \in [-1, 1]$, we denote positive similarities of $\mathbf{u}_i$ and $\mathbf{v}_i$ as a_{ii} and the negative cosine similarities of $\mathbf{u}_i$ and $\mathbf{v}_k$, where $k \neq i$ as a_{ik}.

Universal approximator . For perfectly unified and aligned spaces $\mathcal{U}$ and $\mathcal{V}$, we have $a_{ii} = 1$ for all i, and $a_{ik} < a_{ii}$ for all $k \neq i$. The expected pairwise similarity $\mathbb{E}[a_{ik}] = 0$, as this is consistent with $\mathcal{U}$ and $\mathcal{V}$ being uniformly distributed on the hypersphere. Also note that $\mathcal{L} \to 0$ for $\tau \to 0$, as $\lim_{\tau \to 0} P_i(\tau) = 1$.

Constrained approximator . If we assume an irreducible approximation error, the cost for the condition where $P(a_{ii} \leq \max_{k \neq i} a_{ik}) > 0$ depends on the margin $\max_{k \neq i} a_{ik} - a_{ii}$ and the temperature τ. For smaller values of τ, the cost for misclassified samples increases. Note that higher values of τ conversely decrease the gain of correctly classified samples. The role of τ on the local geometry is a well studied phenomenon (Manna et al., 2025; Wang & Liu).

We can rewrite eq. (2) as the sum of the negative *energy* term a_{ii}/τ and the *entropy* term $\log \left(\sum_{j=1}^{N} e^{a_{ij}/\tau} \right)$. We define $m_i = \frac{1}{N} \sum_{j=1}^{N} a_{ij}$ and write

$$\mathcal{L}_i(\tau) = -\log P_i(\tau) = -\frac{a_{ii} - m_i}{\tau} + \log \left(\sum_{j=1}^{N} e^{(a_{ij} - m_i)/\tau} \right) . \tag{3}$$

The modality gap arises for small temperatures:

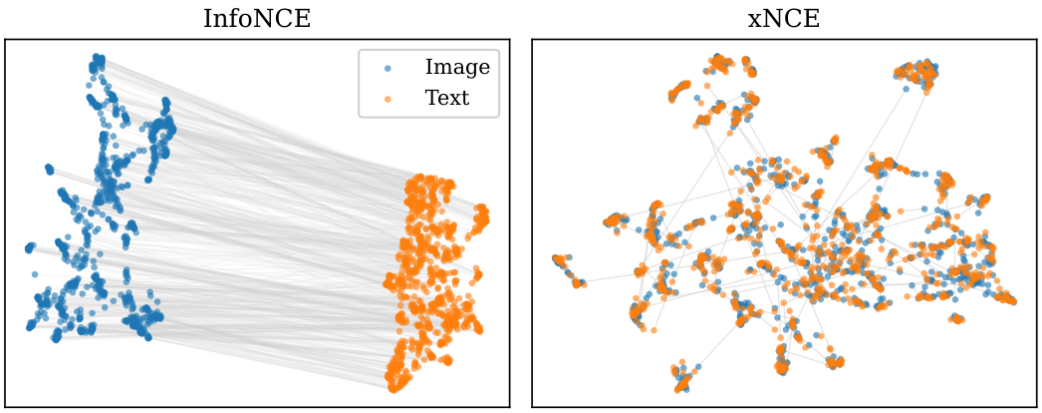

Figure 5: Visualization of Image (blue) and Text (orange) embeddings of COCO after finetuning with InfoNCE (left) and xNCE (right) using UMAP (McInnes et al., 2020). Matching pairs are connected with grey lines. InfoNCE shows a clear separation of modalities, whereas xNCE does not.

Table 1: COCO metrics using InfoNCE (Org), regularized InfoNCE (Reg), and xNCE (Mix) with different mixing ratios r_{mix}. Overall best performing model indicated using bold font, best performance xNCE (Mix) models are underlined.

Loss	τ	r_{mix}	Class mAP		Modality		Embeddings	
			Image (↑)	Text (↑)	Dist. (↓)	Acc. (%, ↑)	Align. (↓)	Unif. (↓)
Org	.01*	-	18.24±3.25	11.26±0.82	0.88±0.02	99.98±0.02	1.38	0.06
Org	.70$^\diamond$	-	**70.45**±0.04	**67.38**±0.04	**0.02**±0.00	**50.12**±0.12	**0.16**	0.03
Reg	.01*	-	59.98±0.43	53.65±0.50	0.12±0.00	80.40±0.15	0.70	**0.02**
Mix	.02*	.01	17.91±0.60	17.53±0.09	0.12±0.01	74.32±2.45	0.83	<u>0.04</u>
Mix	.02*	.05	23.32±1.67	21.83±1.67	0.08±0.00	61.03±1.31	0.67	0.05
Mix	.02*	.50	<u>28.89</u>±3.96	<u>27.82</u>±3.34	<u>0.05</u>±0.00	<u>54.02</u>±1.00	<u>0.48</u>	0.08

* are learnable parameter $^\diamond$ fixed parameter

$$\lim_{\tau \to 0^+} \mathcal{L}_i(\tau) \approx -\frac{a_{ii} - m}{\tau} + \frac{\max(a_{ij}) - m}{\tau}. \tag{4}$$

If we assume an irreducible approximation error, i.e. the optimal ranking $[a_{ij}]_{i,j}^N$ with $a_{ii} \neq \max(a_{ij})$, we have $\mathcal{L}_i \approx (\max(a_{ij}) - a_{ii})/\tau$. The cost can then be further reduced without changing the ranking of $[a_{ij}]$ by letting all a_{ij} converge toward their mean m_i, reducing the pairwise margins and resulting in a degenerate optimum of eq. (4).

The modality gap phenomenon. For a given vector $\mathbf{u}_i$, its positive $\mathbf{v}_i$ and negatives $\mathbf{v}_k$ for all $k \neq i$, a narrow similarity distribution $[a_{ij}]$ where $a_{ik} \approx m_i$ while fixing the ranking of $[a_{ij}]$ corresponds to compressing all $\mathbf{v}_k$ into a tight cluster. If we add $a_{ii} \to 1$ and $a_{ik} \to 0$ for $\tau > 0$, we have $0 < m_i < 1$, which implies that all $\mathbf{v}_k$ form a tight cluster distant from $\mathbf{u}_i$, which gives us the definition of the modality gap.

5.3 Multimodal: Image-to-Text using COCO

Latent-space structure and modality gap. Image-to-Text contrasts using InfoNCE (Org) and xNCE (Mix) on CLIP ViT-B/16 fine-tuned on COCO. With a learnable temperature, InfoNCE exhibits a large

Table 2: COCO retrieval performance for InfoNCE (Org), regularized InfoNCE (Reg), and xNCE (Mix), where R@K = Recall at rank K.

Loss	τ	r_{mix}	Image to Text			Text to Image		
			R@1	R@5	R@10	R@1	R@5	R@10
Org	.01*	-	75.83±0.18	**96.51**±0.05	**99.18**±0.09	**75.52**±0.02	**96.04**±0.13	**98.87**±0.04
Org	.70$^\diamond$	-	49.67±0.05	83.67±0.09	92.83±0.04	47.90±0.02	82.32±0.10	92.40±0.08
Reg	.01*	-	**76.31**±0.18	96.37±0.05	98.96±0.00	74.79±0.14	95.87±0.10	98.67±0.04
Mix	.02*	.01	75.82±0.21	96.41±0.08	99.11±0.08	75.48±0.12	95.81±0.21	**98.87**±0.02
Mix	.02*	.05	75.65±0.10	96.43±0.01	99.11±0.02	75.51±0.06	95.91±0.06	98.79±0.05
Mix	.02*	.50	75.18±0.10	96.12±0.11	98.86±0.04	75.35±0.28	95.89±0.08	98.73±0.05

* are learnable parameter $^\diamond$ fixed parameter

Table 3: Zero-shot classification performance for InfoNCE (Org), regularized InfoNCE (Reg), and xNCE (Mix), reporting Top-1 and Top-5 accuracy (%) on Caltech-101, CIFAR-10, CIFAR-100, and ImageNet

Loss	τ	r_{mix}	Caltech-101		CIFAR-10		CIFAR-100		ImageNet	
			Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Org	.01*	-	82.4±1.0	98.2±0.0	84.7±0.4	99.1±0.1	56.9±1.1	82.6±0.3	51.9±0.3	79.2±0.3
Org	.70$^\diamond$	-	55.7±0.4	84.0±0.5	80.2±0.2	99.0±0.0	32.1±0.2	60.0±0.2	15.5±0.1	34.4±0.2
Reg	.01*	-	79.6±0.2	96.5±0.2	84.2±0.6	99.4±0.1	50.7±0.6	78.0±0.6	43.6±0.2	71.6±0.2
Mix	.02*	.01	**83.1**±0.3	**98.2**±0.2	**87.6**±0.3	99.4±0.1	**61.1**±0.3	**85.1**±0.4	**54.0**±0.1	**82.4**±0.1
Mix	.02*	.05	82.7±0.4	98.2±0.0	86.1±0.6	99.2±0.1	57.4±0.3	82.8±0.6	50.7±0.1	79.6±0.1
Mix	.02*	.50	80.9±0.4	98.1±0.1	86.3±0.3	**99.5**±0.1	55.2±0.5	81.8±0.4	46.3±0.2	75.6±0.1

modality gap (distance 0.88 ± 0.02; linear modality separability $99.98 \pm 0.02\%$) and a comparatively low instance semanticity in the embeddings (image mAP 18.24 ± 3.25; text mAP 11.26 ± 0.82). Regularized InfoNCE (Reg) substantially reduces the gap (distance 0.15 ± 0.01; modality accuracy $82.51 \pm 0.62\%$) but at a higher computational cost of approximately 25–30% longer training time compared to standard InfoNCE and xNCE. In contrast, xNCE reduces the modality gap even at small mixing ratios (e.g. r_{mix}=0.01: distance 0.12 ± 0.01; modality accuracy $74.32 \pm 2.45\%$) while improving instance mAP over the InfoNCE baseline ($17.91 \pm 0.60 \rightarrow 28.89 \pm 3.96$ for image; $11.26 \pm 0.82 \rightarrow 27.82 \pm 3.34$ for text, as r_{mix} increases from 0.01 to 0.50). Increasing the mixing ratio further tightens the modalities (distance 0.12→0.05; modality accuracy 74.32%→54.02%) and continues to increase instance semanticity, revealing a controllable trade-off between modality separability and semanticity. For reference, a scheduled temperature increase in InfoNCE from τ=0.01 to τ=0.70 eliminates the modality gap (distance 0.02 ± 0.00; modality accuracy $50.12 \pm 0.12\%$) and yields the highest instance mAP (image 70.45 ± 0.04; text 67.38 ± 0.04). Figure 6 in section C shows the modality distance during training: xNCE immediately reduces the gap, whereas Reg shifts it more gradually, indicating that xNCE actively counteracts gap formation rather than regularizing it post-hoc.

Retrieval performance. Table 2 reports retrieval recall R@K for I2T and T2I. The InfoNCE model with learnable temperature attains the strongest overall retrieval compared to xNCE by a relatively small margin. Regularized InfoNCE (Reg) achieves the best I2T R@1 (76.31 ± 0.18) while T2I R@1 (74.79 ± 0.14) trails standard InfoNCE slightly, offering better modality alignment overall. xNCE with small mixing retains near-parity at all K, with a slight decrease as r_{mix} increases. We provide a more detailed sensitivity analysis of r_{mix} on retrieval performance in Figure 7 in Section D. The InfoNCE with high temperatures yields the worst retrieval performance, dropping to $\approx 15\%$ at R@1 in I2T compared to the low-temperature setting.

Zero-shot performance. Table 3 reports zero-shot classification on Caltech-101, CIFAR-10, CIFAR-100, and ImageNet. xNCE with small mixing (r_{mix}=0.01) achieves the best Top-1 accuracy on all four benchmarks, outperforming the InfoNCE baseline by +0.7, +2.9, +4.2, and +2.1 percentage points, respectively. Performance degrades monotonically as r_{mix} increases, mirroring the retrieval trade-off: larger mixing ratios reduce the modality gap but erode the discriminative structure needed for zero-shot transfer. Regularized InfoNCE (Reg) consistently trails the InfoNCE baseline despite its lower modality distance, suggesting that its alignment regularization comes at the cost of transferable visual features. The high-temperature InfoNCE variant (τ=0.70) reduces zero-shot performance sharply, most severely on ImageNet (Top-1 51.9→15.5%), confirming that the geometry favored by low temperatures improves discriminative power.

6 Discussion

Temperature and mixing control representation geometry. Our empirical and theoretical analyses point to a coherent picture: InfoNCE for independent encoders has a mode-failure at low temperatures τ, which governs the similarity distribution and, in turn, the geometry of the learned space (sections 4.2 and 4.3). Training independent encoders on digits using InfoNCE yields perfectly separable modalities at low temperatures (modality accuracy = 100%; distance $\approx$ 0.75 at τ=0.01) and a sharp transition towards modality overlap near $\tau\approx$ 0.09. In our theoretical analysis, we show that low temperatures produce a more narrow similarity distribution, as the cost of false positives can be reduced by shrinking similarity scores towards their mean. For larger τ, the distribution has more flexibility in the tails, as more points share the weight for a false positive. This aligns well with the modality gap phenomenon, as a decreasing τ with an increasingly narrow distribution eventually restricts the range of positive similarities with an upper limit smaller than 1.

Mixing modalities prevents the modality gap. Introducing xNCE moderates this effect by adding intra- and intermodal pairs to the negative sets. Mixing samples from both modalities to create negative pairs effectively removes the mode-failure, as simple shrinkage of similarity scores between modalities would change the ranking of points, resulting in a large penalty. For Image-to-Image experiments, xNCE reduces the modality gap uniformly across τ (distance $\leq$0.13 already at $\tau\leq$0.03 and distance <0.05 for $\tau\geq$0.05) and softens the transition in modality separability. In Image-to-Text experiments, xNCE reduces the gap compared to InfoNCE with a learnable temperature (distance 0.88→0.12 at r_{mix} =0.01; further to 0.05 at r_{mix} =0.50) and simultaneously increases instance-level semanticity (image/text mAP increasing steadily with r_{mix}), indicating that mixed negatives act as an implicit regularizer that promotes cross-modal overlap without erasing within-modality class structure. The results reveal a tunable tradeoff: (i) modality separability, (ii) within-modality semanticity (e.g., linear-probe mAP/accuracy), and (iii) retrieval precision. In Image-to-Text, xNCE with small mixing ($\text{r}_{\text{mix}} \in$ 0.01,0.05) substantially reduces the gap while maintaining near-parity with InfoNCE in R@5/10 and trailing only slightly at R@1. In contrast, increasing τ aggressively within InfoNCE also reduces the modality gap and increases the linear-probe mAP but significantly degrades retrieval (e.g. $\approx$ 15 percentage points at R@1 versus low-temperature InfoNCE). Although both temperature and mixing influence the same underlying geometry, mixing offers a more favorable training method for retrieval: small r_{mix} materially lowers the gap and increases semanticity with negligible loss at R@5/10 and only minor impact at R@1, especially when paired with a projection head. It is also worth mentioning that we finetune a model using xNCE on top of pretrained weights using InfoNCE, which is not favoring our method.

Mixing modalities improves zero-shot performance. The zero-shot results reinforce the picture established by the retrieval and embedding metrics, but with an important additional finding: xNCE with small mixing (r_{mix}=0.01) improves zero-shot Top-1 accuracy over the InfoNCE baseline on all four benchmarks, despite the two methods sharing the same low-temperature regime. As r_{mix} increases, zero-shot accuracy degrades monotonically, consistent with the mAP and retrieval trends, indicating that larger mixing ratios over-regularize the embedding space, smoothing out the class-discriminative geometry alongside the modality gap. Notably, this trade-off is qualitatively different from increasing τ: high-temperature InfoNCE also reduces the modality gap but devastates zero-shot performance (ImageNet Top-1 51.9→15.5%). Regularized InfoNCE (Reg) similarly fails to improve over the baseline, suggesting that explicit alignment terms penalize

the geometry in ways that are detrimental to transfer. Mixing, by contrast, achieves gap reduction through a change in the negative set rather than the loss landscape, preserving the low-temperature geometry that benefits zero-shot transfer while removing the degenerative divergence objective of InfoNCE.

Practical implications. For applications prioritizing strict top-1 retrieval (e.g., high-precision search), our results suggest small mixing ($r_{mix} \approx 0.01$–0.05); this configuration preserves CLIP-level recall while reducing the modality gap. For tasks that require modality-invariant representations or stronger semantic organization (e.g., linear probing, transfer, etc.), larger r_{mix} (e.g., 0.5) offer further gap reduction and higher class mAP at the cost of modest top-rank recall. Finally, when tuning τ in the original InfoNCE formulation we find that high temperatures can mimic the geometric effects of mixing but result in a steeper retrieval penalty.

Future work. Promising directions include (i) training a vision-language model using xNCE from scratch using larger batch sizes, as well as applying hard negatives from both modalities to improve discriminative performance (Faghri et al., 2018). In addition, it would be interesting to explore the shared semantic space of both image and text, e.g. by semantic editing. In CLIP, Radford et al. state that no performance gains were made by adding a contrastive head to the CLIP model. Whether this is also true for xNCE remains to be seen.

7 Conclusion

We introduced xNCE, a simple modification of InfoNCE that mixes intramodal and intermodal negatives, and analyzed how temperature and mixing jointly shape similarity distributions and, consequently, representation geometry. Empirically, xNCE consistently reduces the modality gap across temperatures and mixing ratios, while improving within-modality semanticity. For I2T and T2I retrieval in COCO, small mixing nearly preserves InfoNCE retrieval at R@5/10 and incurs only a small R@1 drop; adding a light projection head restores and even improves top-rank recall. Compared to increasing InfoNCE temperature, mixing achieves a more favorable balance between gap reduction, linear-probe quality, and retrieval.

References

Openai/CLIP. URL https://github.com/openai/CLIP.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/ddf354219aac374f1d40b7e760ee5bb7-Abstract.html.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. URL http://arxiv.org/abs/2002.05709.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. EMNIST: an extension of MNIST to handwritten letters. URL http://arxiv.org/abs/1702.05373. version: 1.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Investigating approaches for improving cross-modal alignment in clip. *arXiv preprint arXiv:2406.17639*, 2024.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives, 2018. URL https://arxiv.org/abs/1707.05612.

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

Eleonora Grassucci, Giordano Cicchetti, and Danilo Comminiello. Closing the modality gap enables novel multimodal learning applications. In *Second Workshop on Representational Alignment at ICLR 2025*, 2025.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning.

Linlan Huang, Xusheng Cao, Haori Lu, Yifan Meng, Fei Yang, and Xialei Liu. Mind the gap: Preserving and compensating for the modality gap in CLIP-based continual learning.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions, 2015. URL https://arxiv.org/abs/1412.2306.

Meir Yossef Levi and Guy Gilboa. The double-ellipsoid geometry of CLIP. URL http://arxiv.org/abs/2411.14517.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. URL http://arxiv.org/abs/1405.0312.

Siladittya Manna, Soumitri Chattopadhyay, Rakesh Dey, Umapada Pal, and Saumik Bhattacharya. Dynamically scaled temperature in self-supervised contrastive learning. *IEEE Transactions on Artificial Intelligence*, 6(6):1502–1512, 2025.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. URL https://arxiv.org/abs/1802.03426.

Changdae Oh, Junhyuk So, Hoyoon Byun, YongTaek Lim, Minchul Shin, Jong-June Jeon, and Kyungwoo Song. Geodesic Multi-Modal Mixup for Robust Fine-Tuning. URL http://arxiv.org/abs/2203.03897.

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

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

François Role, Sébastien Meyer, and Victor Amblard. Fill the gap: Quantifying and reducing the modality gap in image-text representation learning. URL http://arxiv.org/abs/2505.03703.

Henry Senior. Delphboy/karpathy-splits. URL https://github.com/Delphboy/karpathy-splits.

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.

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. IEEE. ISBN 978-1-6654-4509-2. doi: 10.1109/CVPR46437.2021.00252. URL https://ieeexplore.ieee.org/document/9577669/.

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

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742. IEEE. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00393. URL https://ieeexplore.ieee.org/document/8578491/.