UNDERSTANDING THE MODALITY GAP IN CLIP

Peiyang (Yonk) Shi [*], **Michael Welle, Mårten Björkman, Danica Kragic**
KTH Royal Institute of Technology
Stockholm, Sweden
{pyshi, mwelle, celle, dani}@kth.se

ABSTRACT

This work examines the phenomenon of the modality gap observed in CLIP-based multimodal learning methods. The modality gap in this context refers to the separation of image and text embeddings in the joint latent space. Some previous research has attributed the gap to cone effect of neural network initialization and suggested closing may not be necessary. However, this study argues that the modality gap is associated with local minima in the CLIP loss function. Through a series of proof-of-concept experiments, we illustrate these local minima and the difficulty of avoiding them in practice. Overall, this work hopes to provide better insight into the root cause of the modality gap.

1 INTRODUCTION AND RELATED WORK

Recent years have seen a growing interest in contrastive learning based multimodal learning, with OpenAI's CLIP model Radford et al. (2021) being a prime example. CLIP uses a contrastively trained joint latent space for text and image modalities, which has been shown to encode rich semantics, providing a solid foundation for a wide range of downstream tasks such as text-conditional generative models Gal et al. (2022), image captioning Fang et al. (2022), and large dataset pairing Schuhmann et al. (2021). Additionally, the multimodal objective function of CLIP has been extended beyond text and image to domains such as video Xu et al. (2021) and audio Wu et al. (2022), indicating a broader potential for the CLIP contrastive framework.

Despite the popularity of CLIP, Zhou et al. (2022) first observed that a *modality gap* exists within the joint latent space, indicating that the image embeddings and text embeddings occupy separate regions of the joint latent space. This phenomenon has also been observed in CLIP models applied in other domains Zhang et al. (2020) Xu et al. (2021). One recent work by Liang et al. (2022) attributes this modality gap to the *cone effects*, where embeddings of two modalities lie on narrow cone regions of the latent space during neural network initialization. Furthermore, some prior work has shown improved performance when the modality gap is closed through projection Zhou et al. (2022), or fine-tuning So et al. (2022).

In this work, we study further what the modality gap observed in CLIP-based multimodal learning methods may be caused by. Our hypothesis is that the CLIP loss function contains local minima, directly affecting the structures of the latent space. To illustrate this, we conduct a series of proof of concept experiments using the CLIP loss to visualize and demonstrate the local minima that are associated with the modality gap. Our experiments show that certain initialization may avoid such local minima, but meeting these initialization criteria in practice is challenging.

2 BACKGROUND

Contrastive Learning is a representation learning framework that aims to organize the latent space in such a way that semantically similar samples, *positive samples*, are brought closer together, while samples with little shared semantics, *negative samples*, are pushed farther apart. One popular method in recent years is the SimCLR Chen et al. (2020) and its NT-Xent loss function . NT-Xent differs from previous contrastive methods in two ways: first, it normalizes the representations to lie on

[*] Corresponding author: Peiyang Shi (pyshi@kth.se)

a hypersphere, which confines the latent space, and second, it does not require explicit negative samples, the denominator of the loss function acts as a soft negative sample by pushing all other samples within the batch away. This allows for a more efficient and effective way of learning robust representations. NT-Xent's effectiveness has been attributedWang & Isola (2020) to two factors in the latent space: *uniformity* and *alignment*. *Uniformity* refers to the property that representations of different data points should be distributed uniformly in the latent hypersphere. *Alignment* refers to the property that the representations of positive samples should be close to each other. Together, uniformity and alignment help the model learn a more expressive and robust representation of the data.

However, the Alignment and Uniformity properties do not explain the second-order effects - the clustering of semantically similar samples. Semantically clustered latent space is attributed as the primary reason for the high-quality representation for downstream tasks Radford et al. (2021). In self supervised learning, work such as Wang & Liu (2021) and Wang et al. (2022) has attributed augmentation as the key factor in clustered representation. The idea is that certain augmentation leads to augmentation overlap - images with semantic similarities will appear similar in certain augmented views.

CLIP Radford et al. (2021) is a multimodal contrastive learning method that was trained on a web-scale dataset of text and images pairs. Based on NT-Xent loss, CLIP extends the contrastive loss to the multimodal domain:

$$\mathcal{L}_{\text{clip}} = -\frac{1}{2} \mathop{\mathbb{E}}_{\substack{(x,y)\sim P_{data} \\ \{x',y'\}_{i=1}^{M}\sim P_{data}}} \left[\log \frac{\exp(x^T y/\tau)}{\sum_i \exp(x_i'^T y/\tau)} + \log \frac{\exp(x^T y/\tau)}{\sum_i \exp(x^T y_i'/\tau)} \right] \tag{1}$$

Intuitively, the $\mathcal{L}_{\text{clip}}$ takes matching pairs of text and image embeddings, denoted as (x, y), as positive samples. Concurrently, the $\mathcal{L}_{\text{clip}}$ samples a set of M negative embedding samples, denoted as $\{x_i', y_i'\}$. Both the positive pair (x, y) and the negative set x_i', y_i' are sampled from the data distribution $P_{\text{data}} \in \mathbb{R}^n \times \mathbb{R}^n$. Like all NT-Xent loss-based methods, CLIP latent space lies on a unit hypersphere, such that $x^T x = y^T y = x_i'^T x_i' = y_i'^T y_i' = 1$.

3 Uniformity, alignment, and the modality gap

In unimodal and self-supervised learning methods such as SimCLR Chen et al. (2020), the *uniformity* and *alignment* terms of the NT-Xent loss function work in harmony. However, in multimodal learning methods, the *uniformity* and *alignment* terms may be in conflict with each other. We hypothesize that this tension between the two terms results in the modality gap phenomenon. To illustrate, we transform $\mathcal{L}_{\text{clip}}$ from eq.1 in to separated *uniformity* and *alignment* terms:

$$\mathcal{L}_{\text{clip}} = \underbrace{- \mathop{\mathbb{E}}_{(x,y)\sim P_{data}} \left[x^T y/\tau \right]}_{\mathcal{L}_{\text{alignment}}} + \underbrace{\mathop{\mathbb{E}}_{\substack{(x,y)\sim P_{data} \\ (x_i',y_i')_{i=0}^{M}\sim P_{data}}} \left[\frac{1}{2} \log \sum_i \exp(x^T y_i'/\tau) + \frac{1}{2} \log \sum_i \exp(x_i'^T y/\tau) \right]}_{\mathcal{L}_{\text{uniformity}}} \tag{2}$$

The term $\mathcal{L}_{\text{alignment}}$ maximizes the similarity between x and y modalities by aligning them onto the same points on the hypersphere. Meanwhile, the term $\mathcal{L}_{\text{uniformity}}$ maximizes the dissimilarity between negative samples in the opposite modalities but does not place any constraints within each own modality. This results in local minima in which $\mathcal{L}_{\text{uniformity}}$ and $\mathcal{L}_{\text{alignment}}$ push two modalities in opposite directions. Thus we hypothesize this eventually results in the modality gap in which the two modalities are completely disjoint in the latent space.

4 Experimental Setup

To illustrate that the CLIP loss induces the modality gap, we conduct several proof-of-concept experiments to understand the process better. We compare CLIP loss with unimodal contrastive loss

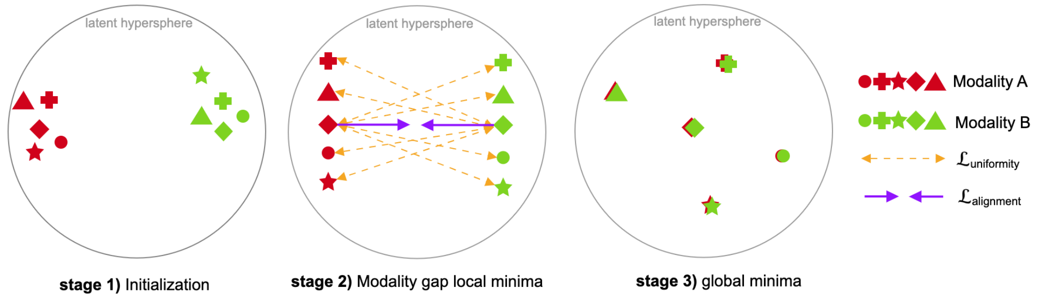


Figure 1: Illustrated modality gap. **Stage 1** - the two modalities are initialized as disjoint cones on the hypersphere. **Stage 2** - When training with the CLIP loss in eq 2, this stage is characterized by a strong and difficult-to-escape local minima - the $\mathcal{L}_{\text{alignment}}$ term serves to pull two modalities together while the $\mathcal{L}_{\text{uniformity}}$ term pushes two modalities apart, resulting in the modality gap. **Stage 3** - If the model can ”escape” the modality gap local minima, then the two modalities are aligned and is uniformly distributed across the latent hypersphere

and investigate the impact of different initialization conditions on the loss function. The proof-of-concept experiments involve optimizing two sets of random points on a 3D sphere using the CLIP loss function. We randomly generated 1,000 points on the 3D sphere and trained with a batch size of 10 and a learning rate of 0.01 using the SGD optimization algorithm. The results are then visualized by randomly subsampling 30 points. More detail on the experiment can be found in Appendix C.

4.1 CLIP LOSS LOCAL MINIMA

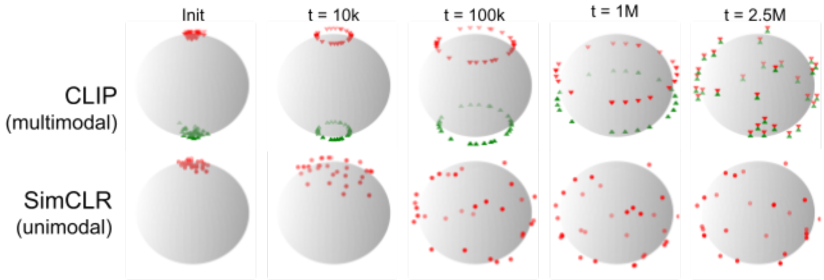


Figure 2: Proof-of-concept experiment comparing CLIP (multimodal NT-Xent) *vs* SimCLR (unimodal NT-Xent) losses, t denotes training epoch. We observe that using the CLIP loss, two symmetrical rings emerge starting at the 10k training epoch. This two rings phenomenon persisted for over 1 million epochs until it reached uniformity at 2.5 million. By comparison, the unimodal NT-Xent loss resulted in uniformity in under 100k steps. We hypothesize that the two rings which CLIP induces on the 3D sphere is the so-called modality gap.

A proof-of-concept experiment was conducted to examine the existence of local minima in the CLIP loss compared to its unimodal counterpart. The experiment compared the CLIP NT-Xent loss to a unimodal NT-Xent loss as depicted in Figure 2. Results showed that CLIP produced two disjoint rings on the 3D sphere and required significantly more training (100k vs 2.5 million epochs) to achieve uniformity compared to the unimodal loss. The emergence of the two disjoint rings is believed to indicate the modality gap phenomenon. Closing this gap is postulated to be a local minimum that requires a substantial portion of the training time.

4.2 MODALITY GAP

Like the two rings phenomenon observed above, we hypothesize that the text and image embeddings are also completely disjoint and thus linearly separatable in the full-scale pre-trained CLIP model. Some previous work Zhou et al. (2022) demonstrated the modality gap qualitatively through 2D visualization. In this study, a *linear* classifier was used to assess the linear separatability between text and image modalities in the CLIP latent space. Using the COCO Captions dataset Lin et al. (2014), which comprises 50,000 images and 250,000 corresponding text descriptions, we conduct a linear separatability test. The classifier was trained using only 0.1% of the dataset and tested on the remaining 99.9%. We report a 100% test accuracy, providing evidence of the complete disjointness of the two modalities in the CLIP latent space.

Additionally, it has been noted that after a sufficient number of training epochs, the proof-of-concept CLIP model above demonstrated the ability to close the modality gap and reach uniformity. Despite this observation, the full-scale CLIP model appears to display a persistent modality gap. We hypothesize in the actual full-scale CLIP model, given its 512-dimensional latent space and various stochasticity coming from the training data and neural network, the local minima associated with the modality gap are more complex than those encountered in the proof-of-concept experiment. This presents an intriguing query that exceeds the scope of the current study, and thus remains a topic for future investigation.

4.3 Effects of Initialization

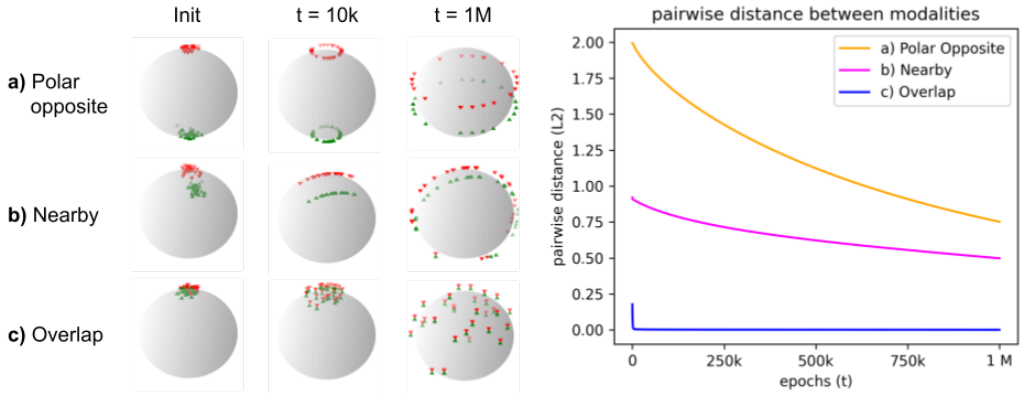


Figure 3: Three types of initialization and their impact on the embedding space. A visual comparison of the convergence characteristics of the three initializations is presented in the left figure, while the right plot illustrates the inter-modality distances throughout a million training epochs. **a) Polar Opposite Initialization** - CLIP loss consistently induces the modality gap in this case, with the pairwise distance (yellow line) between modalities reducing at a slow rate. **b) Nearby Initialization** - The study highlights the sensitivity of initialization to CLIP, as even when the two modalities are initialized in close proximity, the CLIP loss still induces a modality gap. Despite this, the nearby initialization (purple line) demonstrates a slightly faster convergence rate compared to the polar opposite initialization. **c) Overlapped Initialization** - CLIP loss does not induce a modality gap. The two modalities rapidly converge to one another within 10k training epochs, as evidenced by the rapid decline in the pairwise distance (blue line at the bottom).

Previous literature Liang et al. (2022) had attributed the cone effect of neural network initialization as the root cause of the modality gap, we show experimentally in figure 3 that, unless the initialization of two modalities overlaps completely, even the slight disjoint will result in the modality gap. The full-scale CLIP latent space has a 512-dimensional hypersphere, which is much less likely to have overlapped initialization. Therefore we attribute the gap phenomenon to the inadequate design of the CLIP loss rather than purely initialization.

4.4 Effects of Temperature τ

In this experiment, we study the effects of the learn parameter τ. We observe that τ has an effect on the modality gap - lower tau causes gap to close much sooner and higher tau significantly delays the rate at which the modality gap minimizes. 4

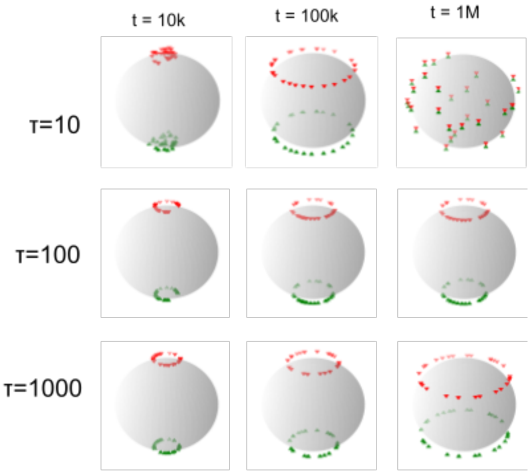


Figure 4: effects of various tau

5 Conclusion and future work

In summary, our study sheds light on the modality gap issue in CLIP-based multimodal learning methods. Our findings suggest that the CLIP loss function is susceptible to local minima due to conflicting alignment and uniformity properties of the contrastive loss. Our proof-of-concept experiments demonstrate the presence of two distinct and separate rings on a 3D sphere, which we believe is evidence of the modality gap. However, we also observe that the proof-of-concept experiment eventually resulted in the closing the gap, but this was not observed in the full-scale CLIP model. The underlying reason for the discrepancy is still unclear and requires further investigation. Moreover, we highlight that the convergence and inter-modality distances of CLIP models are sensitive to initialization, and only overlapping initialization appears to avoid the CLIP loss minima. Our research aims to advance the understanding of the modality gap and underscores the importance of ongoing work in this field.

6 Acknowledgement

This work has been supported by the ERC BIRD, Knut and Alice Wallenberg Foundation and Swedish Research Council.

References

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

Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP), August 2022. URL `http://arxiv.org/abs/2205.01397`. arXiv:2205.01397 [cs].

Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022.

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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. SLIP: Self-supervision meets Language-Image Pre-training, December 2021. URL http://arxiv.org/abs/2112.12750. arXiv:2112.12750 [cs].

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pp. 529–544. Springer, 2022.

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, February 2021. URL http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs].

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Junhyuk So, Changdae Oh, Yongtaek Lim, Hoyoon Byun, Minchul Shin, and Kyungwoo Song. Geodesic multi-modal mixup for robust fine-tuning. *arXiv preprint arXiv:2203.03897*, 2022.

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, Jun 2021. IEEE. ISBN 978-1-66544-509-2. 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 *Proceedings of the 37th International Conference on Machine Learning*, pp. 9929–9939. PMLR, November 2020. URL https://proceedings.mlr.press/v119/wang20k.html. ISSN: 2640-3498.

Yifei Wang, Qi Zhang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. May 2022. URL https://openreview.net/forum?id=ECvgmYVyeUz.

Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4563–4567. IEEE, 2022.

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. *arXiv preprint arXiv:2109.14084*, 2021.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020.

Chenliang Zhou, Fangcheng Zhong, and Cengiz Oztireli. CLIP-PAE: Projection-Augmentation Embedding to Extract Relevant Features for a Disentangled, Interpretable, and Controllable Text-Guided Image Manipulation, October 2022. URL http://arxiv.org/abs/2210.03919. arXiv:2210.03919 [cs].

A Appendix: SimCLR Loss function

SimCLR is primarily a unsupervised image representation learning framework that shows robustness for downstream tasks. The SimCLR's NT-Xent loss function is formulated as follows:

$$\mathcal{L}_{SimCLR} = -\mathbb{E}_{x_i, x'_i \sim P_{pos}, j \sim P_{data}} \left[\log \frac{\exp(x_i^T x'_i/\tau)}{\sum_{j=1, j\neq i}^{2N} \exp(x_i^T x_j/\tau)} \right] \tag{3}$$

In this formulation, the NT-Xent loss utilizes two augmentations of an image, x_i and x'_i, which are sampled from the set of positive samples P_{pos}. The negative samples are implicitly defined in the denominator of the loss function, where the sample x_i is compared with all other samples x_j within the same batch.

B Appendix:SLIP

SLIP Mu et al. (2022) is an extension of CLIP and has shown significant improvements in zero-shot classification and downstream tasks over CLIP. SLIP's method is a simple concept that combines the power of self-supervised SimCLR contrastive loss 8 to the CLIP multimodal contrastive loss in eq 1 as follows:

$$\mathcal{L}_{\text{slip}} = \mathcal{L}_{\text{clip}} + \mathcal{L}_{\text{SimCLR}} \tag{4}$$

C Appendix: Proof of Concept Experiment Setup

We devise a proof of concept experiment using a 3D sphere to illustrate that the CLIP loss function 1 induces a modality gap even in low dimensional space and without any external factors coming from the dataset or neural network architecture. To that end, we generate two sets of points scattered on a 3-dimensional euclidean space, representing the two modalities. Our parameter space for optimization is the euclidean coordinate space of the points so we can we optimize directly on the positions of the points. We randomly generated 1,000 points on the 3D sphere and trained with a batch size of 10 and a learning rate of 0.01 using the SGD optimization algorithm. The results are then visualized by randomly subsampling 30 points.

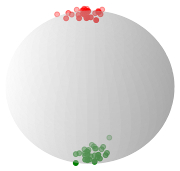


Figure 5: Example initialization of the proof of concept experiment

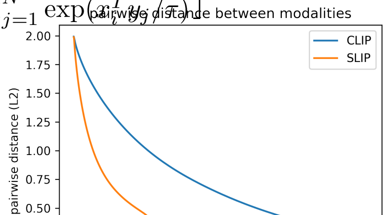
D SLIP and the modality gap

In this experiment, we conduct a proof of concept experiment involving the 3D sphere utilizing the SLIP Mu et al. (2021) loss in equation 4. The results of this experiment are depicted in Figure 6. Our findings indicate that the SLIP algorithm converges at a significantly faster rate in comparison to the CLIP algorithm. We postulate that this accelerated convergence can be attributed to the inclusion of a self-supervised SimCLR loss term for the image modality within the SLIP framework. This loss term effectively increases pressure on the uniformity of image embeddings on the hypersphere, thereby disrupting the equilibrium of the modality gap.

$$\mathcal{L}_{\text{uniformity}} = \mathop{\mathbb{E}}_{\substack{(x,y)\sim P_{data} \\ (x'_i, y'_i)_{i=0}^{M} \sim P_{data}}} \left[\frac{1}{2} \log \sum_i \exp(x^T y'_i/\tau) + \frac{1}{2} \log \sum_i \exp(x_i'^T y/\tau) \right] \tag{5}$$

$$\mathcal{L}_{\text{clip}} = -\frac{1}{2} \mathbb{E}_{(x,y)\sim P_{data}} \left[\log \frac{\exp(x_i^T y_i/\tau)}{\sum_{j=1}^{N} \exp(x_j^T y_i/\tau)} + \log \frac{\exp(x_i^T y_i/\tau)}{\sum_{j=1}^{N} \exp(x_i^T y_j/\tau)} \right] \tag{6}$$

Figure 6: Comparison of SLIP vs CLIP's modality gap with identical initialization (polar opposit), SLIP converges at a significantly faster rate than CLIP, demonstrating the importance of good alignment and closing the modality



$$\mathcal{L}_{\text{clip}} = -\frac{1}{2}\,\mathbb{E}_{(x,y)\sim P_{data}}\left[\log\frac{\exp(x_i^T y_i/\tau)}{\sum_{j=1}^{N}\exp(x_j^T y_i/\tau)} + \log\frac{\exp(x_i^T y_i/\tau)}{\sum_{j=1,j\neq i}^{N}\exp(x_i^T y_j/\tau)}\right] \tag{7}$$

$$\mathcal{L}_{SimCLR} = -\mathbb{E}_{x_i, x_i' \sim P_{pos}, x_j \sim P_{data}}\left[\log\frac{\exp(x_i^T x_i'/\tau)}{\sum_{j=1,j\neq i}^{2N}\exp(x_i^T x_j/\tau)}\right] \tag{8}$$