Masterarbeit im Fach Informatik
RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT BONN
Institut für Informatik III
Computer Vision Group
Prof. Dr. H. Kühne

Temperature and Margin Scheduling for Multimodal Contrastive Learning on Long-Tail Data

29. January 2025

vorgelegt von:
Siarhei Sheludzko
Matrikelnummer: 3092139

Gutachter:
Prof. Dr. Hildegard Kühne
Prof. Dr. Juergen Gall

Betreuer:
Dr. Anna Kukleva

I hereby declare that I have created this work completely on my own and used no other sources or tools than the ones listed, and that I have marked any citations accordingly.

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie Zitate kenntlich gemacht habe.

Bonn, 29. January 2025
Siarhei Sheludzko

Contents

Abstract

Self-supervised learning, particularly contrastive learning (CL), has revolutionized representation learning by allowing models to learn relevant representations from large-scale image and video datasets. CL encourages the model to learn discriminative features by pulling together semantically similar samples, while pushing dissimilar apart. This approach can learn more robust and generalizable embedding spaces compared to supervised methods, as it utilizes the inherent structure of the data rather than relying on explicit labels for every instance. However, the long-tailed nature of such data presents significant challenges, since majority classes can dominate the representation, obstructing generalization.

This thesis project introduces a novel method for multimodal contrastive learning that incorporates two key components: (1) Temperature Scheduling, using cosine or linear schedules to mitigate the modality gap and improve alignment in multimodal embeddings, and (2) Distribution-based Pseudo-supervision which adapts the temperature based on the semantic frequency of training samples, balancing group-wise and instance-level discrimination. Furthermore, we extend the idea of temperature scheduling in InfoNCE loss to max-margin loss formulations.

We validate our approach on several long-tailed video-language datasets: EPIC-KITCHENS-100, YouCook2, and Something-Something-v2-LT. The results demonstrate that our method effectively mitigates the challenges of modality gaps and imbalanced data distributions, achieving state-of-the-art performance across all evaluated datasets. This work particularly demonstrates the effectiveness of dynamic temperature modulation as a simple, yet versatile mechanism to improve multimodal representation learning.

Acknowledgements

I would like to express my gratitude to all the people who supported and guided me throughout the journey of writing this thesis.

First and foremost, I extend my heartfelt thanks to my professor, Prof. Dr. Hildegard Kühne, for her invaluable guidance and encouragement. Her expertise and support have been instrumental in shaping my understanding of multimodal contrastive learning and ensuring the success of my research.

I am very grateful to my supervisor, Dr. Anna Kukleva, for her unwavering support and insightful feedback throughout this process. Her dedication and expertise have greatly enriched my learning experience, especially during our collaboration on the lab project and the continuation of this work in my thesis.

A special thanks goes to Angel Villar-Corrales for introducing me to the fascinating world of computer vision during the "Cuda Lab" project. This experience was not only one of the highlights of my Master's studies but also a pivotal moment that inspired me to focus on this exciting field in my thesis. The "Cuda Lab" was an exceptional opportunity that truly demonstrated how captivating and impactful computer vision can be.

Finally, I want to express my deepest gratitude to my fiancée, Lisa Michels, for her unwavering love and support. Throughout the most challenging moments of this journey, she has been my pillar of strength and a constant source of motivation, encouraging me when I needed it the most.

Siarhei Sheludzko

CHAPTER 1

Introduction

1.1 Introduction

Representation learning, especially in the self-supervised paradigm, has shown significant advancements in recent years. Among the various approaches, contrastive learning (CL) has become a cornerstone for the extraction of robust feature representations. Through the optimization of agreement between different manifestations of the same instance and encouraging separation from unrelated instances, contrastive learning has proven itself to be remarkably versatile. It has found success in single-modality tasks including image representation learning[1–4] and multimodal tasks that align the vision and language modalities in a shared embedding space[5–7]. These abilities have been essential for achieving State-of-the-Art performance on various tasks like classification[1], retrieval[6, 8] and visual question answering [9].

Self-supervised learning (SSL) has become one of the most promising approaches due to its power to utilize large-scale unlabeled data without costly human labeling. Nonetheless, the dataset collected from web-scale sources follows a long-tail distribution, where few classes dominate the dataset while others are underrepresented. This imbalance creates significant challenges for model training, as overrepresented classes dominate the learned representations, leading to suboptimal generalization. Addressing these challenges requires specialized techniques that can effectively handle the distributional imbalances inherent in large-scale datasets.

The core mechanism of contrastive learning involves pulling positive pairs (*e.g.* augmented views of the same instance) closer together in the representation space while simultaneously pushing negative pairs apart [1]. This push-and-pull dynamic is controlled by the temperature parameter within the contrastive loss formulation, which modulates the strength of separation between negative samples. Despite its simplicity, the temperature parameter plays an important role in shaping the representation space and has primarily been applied with a fixed value in many existing frameworks. While a higher temperature reduces the repulsive force on

negatives, encouraging cluster formation, a lower temperature enforces more uniform distribution of the representations in the embedding space. The approach of Qiu *et al.* [10] demonstrated that the InfoNCE loss inherently penalizes negative pairs based on their hardness, and the temperature parameter controls the strength of these penalties. This property enables "semantic harmonization" in long-tail data: samples with frequent semantics are assigned larger temperature values to better preserve local semantic structures (group-wise discrimination), while rare samples receive smaller temperature values to enforce a more uniform distribution (instance-wise discrimination) in the embedding space.

Recently, studies [11, 12] have explored dynamic temperature schedules during training to enhance representation learning, especially in scenarios with imbalanced data distributions. In single-modality settings, Kukleva *et al.* [11] propose temperature schedules where the temperature is dynamically adjusted to allow for alternate group-wise and instance-wise discrimination. This cyclic adjustment is shown to improve representation quality, particularly in imbalanced distribution scenarios, where no prior information about the distribution of the training data is available. For the multimodal context, the modality gap poses another challenge, which refers to the difference between the representation distributions in different modalities. Such gap may obstruct alignment in a shared embedding space. Some studies [12, 13] link the modality gap to the temperature parameter, offering valuable insights into its role in shaping multimodal embeddings. Notably, Yaras *et al.* [12] analyze the modality gap through the lens of gradient flow learning dynamics and demonstrate that linearly increasing of temperature over the course of training can effectively reduce this gap.

To address the challenges of multimodality learning on imbalanced data, we present Temperature Schedules++ (TS++), a framework for multimodal contrastive learning that utilizes temperature scheduling and distribution-aware individualization of temperature. Our framework includes two key innovations:

- **Temperature Scheduling:** In our approach we use two scheduling methods to dynamically change the temperature: a cyclic cosine schedule [11] and a linear schedule, building on the recent work of Yaras *et al.* [12], that linearly increases the temperature during the training to decrease the modality gap.

- **Distribution-based Pseudo-supervision:** Further, we refine the temperature adjustment using the text modality to estimate the distribution of the training set. Samples with more frequent semantics would have a higher temperature, enabling group-wise discrimination. In contrast, samples with rare semantics enforce stronger repulsion, emphasizing instance-wise discrimination.

Additionally, we examine the parallels between the roles of temperature in InfoNCE loss and margin in max-margin loss, focusing on their impact on penalties for negative samples of different hardness. Based on this analysis, we extend the temperature scheduling and distribution-based pseudo-supervision used in InfoNCE loss to the max-margin contrastive loss and empirically validate the effectiveness of this approach.

To investigate the efficacy of our framework, we perform experiments on long-tailed video-language datasets, including EPIC-KITCHENS-100 [14], YouCook2 [15], and Something-Something-v2-LT [16]. By integrating our method into the state-of-the-art contrastive learning methods for each dataset, we demonstrate its ability to mitigate the modality gap and balance group-wise and instance-wise discrimination. Our findings reveal that combining unsupervised dynamic scheduling with distribution-aware temperature adjustments improves performance on all used datasets.

In summary, this work shows the importance of temperature schedules in addressing the challenges of multimodal contrastive learning. By mitigating the modality gap and accommodating long-tailed data distributions, TS++ provides a versatile and effective solution for improving representation quality in complex multimodal settings.

1.2 Thesis Structure

This thesis is structured as follows:

Chapter 2 presents an overview of related work in contrastive learning, covering both unimodal and multimodal approaches. It highlights state-of-the-art methods in video retrieval and multi-instance retrieval while also exploring strategies to address challenges posed by long-tail data distributions in contrastive learning.

Chapter 3 presents the foundational concepts and methods of contrastive learning, including InfoNCE and max-margin losses, along with their key parameters, such as temperature and margin. It also covers multimodal contrastive learning and challenges posed by long-tail data distributions.

Chapter 4 introduces TS++, a method to address long-tailed distributions in multimodal data by combining temperature/margin scheduling with class-specific adjustments derived from semantic distributions.

Chapter 5 describes the downstream tasks and reviews the state-of-the-art contrastive learning frameworks used for these tasks. It provides an overview of the datasets used for the evaluation, details the implementation, presents experimental results, and offers a thorough analysis.

Chapter 6 highlights the main contributions of this thesis and outlines promising avenues for future research to improve multimodal contrastive learning in long-tailed data scenarios.

1.3 Contributions

The key contributions of this work are as follows:

- We apply the cosine temperature scheduling proposed by Kukleva *et al.* [11]

to the multimodal scenario and compare with linear temperature scheduling proposed by Yaras *et al.* [12].

- We extend the temperature scheduling method (for InfoNCE loss) to margin scheduling (for max-margin loss).
- We enhance effect of temperature/margin schedules on imbalanced data by tuning the temperature/margin based on the semantic frequency of each training sample within the training data.

CHAPTER 2

Related work

Recent advancements in SSL, especially in contrastive learning, have greatly contributed to various fields, including multimodal representation learning and video retrieval. This chapter overviews contrastive learning methods for unimodal and multimodal scenarios. Additionally, it explores SotA methods in video retrieval and multi-instance retrieval, highlighting the challenges and innovations in this domain.

2.1 Contrastive learning

Contrastive learning uses instance discrimination to distinguish individual data samples instead of assigning them to predefined classes. Wu *et al.* [17] introduced an approach leveraging a noise contrastive estimator (NCE), enabling the model to differentiate between samples by comparing their features. This method relies on contrasting representations of positive pairs while simultaneously discouraging similarity to features of other samples.

2.1.1 Role of negatives in contrastive learning

Contrastive learning relies on the interplay between positive and negative pairs. Positive pairs are well-defined: they either represent different views of the same data in an unimodal setting or share the same underlying content in different modalities in a multimodal context. Negative pairs, on the other hand, consist of samples that do not share the same underlying content. In the unimodal scenario, they are views of different original samples, while in the multimodal context, they involve unrelated pairs across modalities. The significance of negative pairs has been emphasized in numerous prior studies [1, 18–22]. Some of the experiments [1, 21, 22] indicate that increasing the number of negative examples can significantly enhance learning performance. Manna *et al.* [23] propose to use individual temperatures for each pair based on similarity values. Hardness of negatives also plays an important role in

the learning process [24]. Robinson *et al.* [25] propose a strategy of hard negatives mining. Some studies [26] have noted that hard negatives mining can hinder training convergence. To overcome it, Zhang *et al.* [27] propose a new loss function to assign the penalty strength of negatives based on their hardness. Wang *et al.* [28] introduce a Non-negative Contrastive Learning (NCL) inspired by Non-negative Matrix Factorization. It enforces non-negativity on feature representations, leading to semantically consistent and sparse features that align with human understanding.

2.1.2 Unimodal contrastive learning

Unimodal CL in the visual domain focuses on learning visual representations by comparing individual data samples using instance discrimination. This involves generating positive pairs of images through data augmentations, where two random augmentations are applied to the same raw data sample, producing two distinct views. These views are projected into the embedding space using a shared visual encoder and a projection head. This idea is utilized by Simple Framework for Contrastive Learning of Visual Representations (SimCLR) [1]. The framework combines a variety of data augmentations, the application of learnable non-linear transformations, and the utilization of large batch sizes. The performance of SimCLR is strongly influenced by batch size, as the number of available negative samples is directly constrained by it. To address this limitation, He *et al.* [2] introduced a new framework MoCo that decouples the number of negative samples from the batch size. It is done by employing a dynamic queue dictionary powered by a momentum-based encoder. This approach enables a larger and more diverse set of negative samples, independent of the batch size. BYOL [29] simplifies SSL by removing the need for negative pairs while achieving competitive performance. It is done using two networks: online and target. Online network learns to predict representations, target network provides stable targets by using a slow-moving average of the online network's parameters. The objective of the loss function is to minimize the mean squared error between the predicted and target representations. DINOv2 [4] offers a significant advancement in self-supervised contrastive learning by scaling training and improving efficiency to produce robust general-purpose visual features. DINOv2 leverages a curated dataset of 142 million diverse images, which ensures high-quality training data, and incorporates several technical improvements for stabilizing and accelerating training. DINOv2 excels across a wide range of vision tasks - such as image classification, segmentation, and instance retrieval - without requiring fine-tuning, often surpassing the performance of earlier self-supervised methods. Wang *et al.* [30] analyze effects of generated data in contrastive learning and introduce an Adaptive Inflation strategy which optimally balances real and synthetic data while adjusting augmentation strength to improve downstream accuracy.

2.1.3 Contrastive learning on long-tail data

Contrastive learning experiences great challenges posed by imbalanced data distributions, as underrepresented classes receive fewer training samples, leading to biased representations. Addressing these challenges by integrating long-tail learning

techniques into CL frameworks can increase model performance, ensuring better generalization across both frequent and rare semantics. Recent studies, including those by Kang *et al.* [31], Liu *et al.* [32] and Yang & Xu [33] discovered that self-supervised methods compared to fully supervised methods perform stably well and learn a more robust embedding space. This is particularly apparent in situations dealing with imbalanced datasets. Cui *et al.* [34] extend the idea of MoCo [2] by adding supervision to contrastive learning. It pulls samples of the same class closer to their center, ensuring that low-frequency classes get equal representation in the learned embeddings. Their method moderates the gradient disparity for low-frequency classes by adaptively rebalancing the influence of classes during training, resulting in a smoother gradient distribution across all classes. Kukleva *et al.* [11] show that tail classes benefit more from the lower temperatures, leading to a more uniform distribution in the embedding space. The head classes, however, have better results with the higher temperatures, as it forces the clustering more. For the SSL scenario they introduce the temperature schedule. The temperature is cyclically changing during the training, helping the model to adapt more effectively to diverse data distributions. Miao *et al.* [35] address class imbalance in datasets by combining a contrastive learning branch with an imbalance learning branch. This architecture enhances feature learning, particularly for underrepresented classes, by employing tailored sampling strategies and data augmentation techniques. Du *et al.* [36] propose the probabilistic contrastive learning algorithm to address challenges in visual recognition on long-tailed data. This method estimates data distribution in the feature space and samples pairs accordingly, improving the performance of the tail classes.

2.2 Multimodal contrastive learning

Multimodal CL is a method that learns to map the representations of different data modalities (*e.g.* images and text) into shared embedding space [5]. In this scenario, the positive pairs are samples that share the same underlying content across different modalities, like image and corresponding textual description of the image. All other unrelated samples from different modalities are interpreted as negatives. This approach enables pretrained multimodal models to perform effectively across various downstream tasks, including retrieval, zero-shot classification, and generative modeling [37]. Despite its success, multimodal CL creates some challenges, which are not presented in the unimodal CL. One of them is the modality gap [12, 13], where differences in the characteristics of modalities make it difficult to align their representations in the shared embedding space. Additionally, Bravo *et al.* [38] shows that multimodal models are biased towards object classes and struggle to capture fine-grained details.

2.2.1 Modality gap

Modality gap [13] refers to the phenomenon of preserving a reasonable distance between the representations of samples from two different modalities. This separation is influenced by both the model's initialization state and temperature parameter in

contrastive loss, which contributes to a "cone effect" - the tendency of single-modality encoders to utilize only a narrow, cone-shaped subspace within the total shared embedding space. To address this issue, Zhang *et al.* [39] propose the C3 method (**C**onnect, **C**ollapse, **C**orrupt) to reduce the modality gap. This approach first creates a shared representation space with CL, then it eliminates the modality gap by subtracting modality-specific means from embeddings and finally adds Gaussian noise to improve robustness and alignment. Fahim *et al.* [40] argue that the modality gap is not specific to the differences between modalities, but rather an inherent consequence of the contrastive loss function in the multimodal settings. For its elimination, they suggest to add explicit alignment and uniformity terms to the CLIP loss. It ensures that embeddings are more uniformly distributed across the embedding space. Yaras *et al.* [12] analyze the phenomenon of the modality gap from the perspective of the gradient flow learning dynamics and show that linear increasing of the temperature during the training helps to mitigate it.

2.2.2 SotA in Video Retrieval and Multi-instance Retrieval

We provide here a brief overview of the State-of-the-Art methods for Video Retrieval and Multi-instance Retrieval.

UniVL [41] learns a joint representation of the video and raw subtitles in a self-supervised manner. Due to the weak correlation between raw subtitles and the abstract meaning of the video, this approach has inconsistency between pretraining and fine-tuning steps [6]. MELTR [42] improves fine-tuning step of UniVL with non-linear combination of auxiliary losses. VLM [43] overcomes a modality gap between visual and textual encoders by combining a masked frame model and a masked language model inside of a single encoder. VAST [8] is an omni-modality video-text foundational model that works with video, audio, subtitles and text. It leverages an off-the-shelf large language model (LLM) to extract abstract meanings from subtitles, effectively mitigating inconsistencies between the pretraining and fine-tuning stages. JPoSE [44] employs distinct embedding spaces for different parts of speech (verbs and nouns) and subsequently integrates them into a unified embedding space, enabling more effective action retrieval. EgoVLP [45] mainly focuses on the egocentric videos. They use the EgoNCE pretraining objective for mining of positive and negative samples which are specific for egocentric data. LaViLa [7] addresses the sparsity of annotations in human-annotated video datasets. They use an LLM as an automatic narrator. The framework uses both human annotations and generated narrations during contrastive learning. AVION [8] builds upon the LaViLa architecture and focuses on optimizing IO, CPU and GPU bottlenecks to minimize resource usage and significantly reduce training time.

CHAPTER 3

Background

This chapter demonstrates the key concepts, methods, and challenges that underlie this thesis. It begins with a discussion of contrastive learning and its widely used loss functions, including InfoNCE and max-margin losses, which form the foundation of representation learning. The role of temperature and margin in shaping the embedding space is also explored. It also shows the parallels between InfoNCE and max-margin losses, showing how different values of temperature and margin affect training dynamics. Building on this, the chapter introduces multimodal contrastive learning and its application in aligning representations across different modalities, such as video and text. Specific losses like the CLIP loss and Multi-Instance Max-Margin loss are briefly outlined.

3.1 Contrastive losses

Contrastive learning is one of the most important frameworks to learn expressive representations that are used in various downstream tasks [46]. At its core, CL aims to structure the embedding space such that similar instances are brought closer together while dissimilar ones are pushed apart. Given a set of samples $\{x_1, ..., x_N\}$ and single modality encoder f, the similarity score between the two samples can be computed as (Eq. 3.1):

$$s_{ij} = f(x_i)f(x_j)^T, \tag{3.1}$$

which then can be used to calculate the contrastive learning objective. Among the various contrastive loss functions, InfoNCE and max-margin losses are the most widely used. The following subsections provide a detailed discussion of these objectives, their underlying principles, and role of their parameters in CL.

3.1.1 InfoNCE loss

InfoNCE[47] is the most commonly used contrastive loss and is formulated as (Eq. 3.2):

$$\mathcal{L}_{\text{InfoNCE}}(s) = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{N}\exp(s_{ij}/\tau)}, \tag{3.2}$$

where s_{ii} represents cosine similarity between two different representations of the same input, s_{ij} is the similarity between representations of x_i and x_j inputs, N is number of samples in the batch, and τ denotes the temperature parameter.

Role of the temperature. Wang *et al.* [18, 48] demonstrate that contrastive loss is a hardness-aware function and the temperature parameter plays there a pivotal role. This parameter regulates the penalty intensity for hard negative samples (negative samples with high similarity to the anchor) and influences the uniformity of the embedding space. In this context, the temperature serves as a control mechanism, balancing the "uniformity-tolerance dilemma." In contrastive learning, different approaches adopt different strategies for handling the temperature parameter. In some methods [1, 2], the temperature is treated as a static hyperparameter tuned manually. Others [5] make the temperature a learnable parameter, allowing the model to adapt it dynamically during training to optimize performance. Qiu *et al.* [10] introduced a novel contrastive loss that automatically individualizes the temperature for each sample based on the distribution frequency of its semantics. This method assigns higher temperatures to samples with frequent semantics to maintain local structure, while allocating lower temperatures to samples with rarer semantics to promote more distinguishable features. Additionally, Yaras *et al.* [12] argue that using a learnable temperature in multimodal contrastive learning can inadvertently preserve the modality gap. To address this, they propose a linearly growing temperature strategy which helps to mitigate the modality gap by gradually aligning embeddings from different modalities throughout training.

3.1.2 Max-Margin loss

The most popular contrastive loss alternatives to InfoNCE loss is the max-margin loss [49]. The objective of max-margin loss is to maximize the distance between positive and negative pairs and is formulated as (Eq.3.3):

$$\mathcal{L}_{\text{max-margin}}(s) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1,i\neq j}^{N}[\max(0,\, s_{ij} - s_{ii} + m)], \tag{3.3}$$

with s_{ii} -similarity between positive samples, s_{ij} - similarity between negative samples and m - margin, which defines the minimum similarity required between the positive and negative pairs.

Figure 3.1 illustrates the categorization of negative samples in the context of the max-margin loss function. The samples are categorized into three distinct groups according to their similarity scores in relation to an anchor (a) and a corresponding positive sample (p):

- **Hard Negatives:** These are the negative samples (n_H) that exhibit high similarity to the anchor, with similarity scores $s(a, n_H)$ greater than $s(a, p)$.
- **Semi-Hard Negatives:** These negative samples (n_S) have similarity scores that fall between $s(a, p) - m$ and $s(a, p)$, where m is the margin.
- **Easy Negatives:** These are negative samples (n_E) with similarity scores $s(a, n_E)$ lower than $s(a, p) - m$. These samples are far from the anchor and do not contribute to the loss, as they are already well-separated in the embedding space.

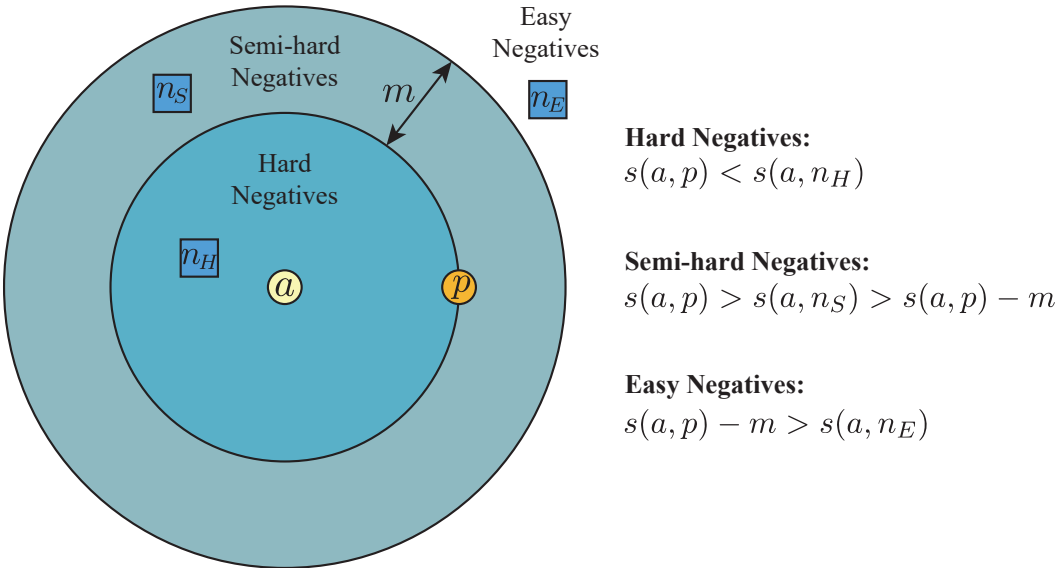

Figure 3.1: Visualization of negative sample categories in the max-margin loss function: hard negatives, semi-hard negatives, and easy negatives, defined based on their similarity scores relative to the anchor (a) and the positive sample (p).

Role of the margin. The margin in the max-margin contrastive loss plays an important role in determining which negative samples contribute to the loss and how strongly they influence the training process. Margin establishes a threshold for the similarity scores between the anchor and negative samples, shaping how different types of negatives are treated by the loss. The margin determines how many "non-hard" negatives (those closer to easy negatives) are included in the loss calculation. Increasing the margin broadens this range, allowing more semi-hard negatives to influence the training process.

3.1.3 Parallels between InfoNCE and Max-Margin

The temperature and margin parameters in InfoNCE and max-margin losses play a critical role in shaping the training dynamics. Although the losses are different, both τ and m exhibit similar influences on contribution of the negatives with different

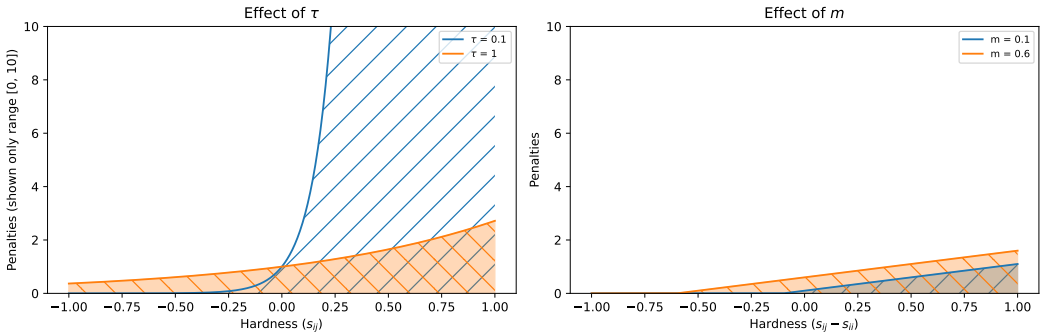

Figure 3.2: Effect of τ (left) and m (right) on penalties for individual negatives based on their hardness.

hardness to the loss calculation. In the InfoNCE loss, the hardness of negative samples is defined as their similarity to the anchor, expressed as $h_{ij} = s_{ij}$. In contrast, for the max-margin loss, the hardness of negatives is determined by the difference between the similarities of positive and negative samples, given by $h_{ij} = s_{ij} - s_{ii}$.

Figure 3.2 illustrates the impact of a single negative sample on the loss as a function of its hardness, considering different values of τ and m. The blue lines represent low values of τ and m, while the orange lines correspond to high values.

To better understand the influence of τ and m, we group their effects based on the penalties imposed on negative samples with different hardness:

- Lower τ and lower m amplify the gradients for hard-negative pairs, increasing their impact during training.
- In contrast, higher τ and higher m include more "less hard" negatives to the contribution, reducing relative influence of hard negatives.

3.2 Multimodal contrastive losses

In multimodal scenario (*e.g.* video and text) the core objective of CL remains to align semantically similar content from different modalities into a shared embedding space, while preserving the ability to distinguish between distinct samples. The similarities for video to text and text to video are computed as (Eq. 3.4):

$$s_{v \to t} = f_v(v) f_t(t)^T, \quad s_{t \to v} = f_t(t) f_v(v)^T, \tag{3.4}$$

where f_v and f_t are modality-specific encoders for video and text, respectively, and v and t represent the video and text embeddings.

The multimodal contrastive loss is typically formulated symmetrically to ensure bidirectional alignment between modalities (Eq. 3.5):

$$\beta \left(\mathcal{L}(s_{v \to t}) + \mathcal{L}(s_{t \to v}) \right), \tag{3.5}$$

where $\mathcal{L}(s)$ can take different forms, such as InfoNCE or max-margin losses, depending on the design of the learning framework, and β is a constant with typical values $1/2$ or 1.

3.2.1 CLIP Loss: Combining InfoNCE losses

The CLIP loss, introduced in the CLIP framework [5], is a widely used multimodal contrastive loss that extends the InfoNCE formulation to align text and image (or video) embeddings. In this approach, InfoNCE losses are applied independently for both visual-to-text and text-to-visual directions, with the combined objective encouraging bidirectional alignment (Eq. 3.6):

$$\frac{1}{2}\left(\mathcal{L}_{\text{InfoNCE}}(s_{v\to t}) + \mathcal{L}_{\text{InfoNCE}}(s_{t\to v})\right). \tag{3.6}$$

3.2.2 MI-MM Loss: Combining Max-Margin losses

The Multi-Instance Max-Margin (MI-MM) loss [44, 45] builds on the max-margin principle to align embeddings across modalities. It combines max-margin losses for both video-to-text and text-to-video directions (Eq. 3.7):

$$\mathcal{L}_{\text{max-margin}}(s_{v\to t}) + \mathcal{L}_{\text{max-margin}}(s_{t\to v}). \tag{3.7}$$

CHAPTER 4

Methodology

In this chapter, we present the methodology underlying our proposed approach which is based on recent advances in contrastive learning and temperature scheduling techniques [10–12]. TS++ is specifically designed to enhance multimodal contrastive learning on long-tailed datasets by leveraging a combination of dynamic temperature and margin schedules and a pseudo-supervised strategy informed by semantic class distributions.

4.1 Approach

Building on our observations on the impact of temperature and margin across modalities and the insights from [11], [12] and [10], we introduce the multimodal **T**emperature **S**chedules (TS++), a method designed to address long-tailed distributions in multimodal data.
Our approach includes two key components:

- **Temperature / margin Schedule**: This component dynamically changes τ or m during training. We explore two variants for scheduling:
  - **Cyclic Cosine Schedule:** Inspired by [11], this approach modulates τ or m cyclically, following a cosine pattern.
  - **Linear Schedule:** Based on [12], this variant linearly increases τ or m throughout training, mitigating the modality gap.
- **Distribution-aware Pseudo-supervision:** This component assigns individual τ or m values for every sample in the batch based on the semantic distribution of the corresponding sample in the training data.

4.1.1 Temperature / margin schedule

We adopt two temperature schedules methods proposed by Kukleva *et al.* [11] and Yaras *et al.* [12]. Instead of defining the range of the τ / m schedule with minimal and maximal values, we use an α parameter to calculate the correction of temperature (τ_{corr}) / correction of margin (m_{corr}).

For Cyclic Cosine Schedule τ_{corr} changes in the range $[-\alpha/2, \alpha/2]$ based on the current iteration t and predefined period T (Eq. 4.1):

$$\tau_{corr}(t) = \frac{\alpha(1 + \cos(2\pi t/T) - \alpha}{2}. \tag{4.1}$$

This schedule allows the model to alternate between focusing on group-wise and instance-wise discrimination. We have discovered empirically that highest performance is achieved with the highest values τ and m in the end of the training. To ensure this, we select the period T for the cyclic cosine schedule such that three to five complete cycles occur during training, allowing the model to finish training with τ and m at their peak values.

For Linear Schedule τ_{corr} changes in the range $[-\alpha/2, \alpha/2]$ based on the current iteration t and total number of iterations in the training t_{max} (Eq. 4.2):

$$\tau_{corr}(t) = -\alpha/2 + \alpha t/t_{max}. \tag{4.2}$$

This schedule helps to close the modality between matched pairs from different modalities.

Figure 4.1 illustrates both scheduling methods. For simplicity we visualize the cyclic cosine schedule only with 2 periods.

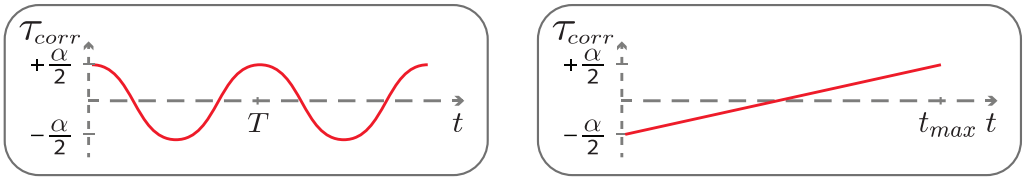

Figure 4.1: Cyclic cosine (left) and linearly growing (right) temperature corrections.

Margin correction is calculated in the same manner, following either the Cyclic Cosine or Linear Schedule.

4.1.2 Distribution-aware Pseudo-supervision

We assign individual temperature or margin values to each anchor sample based on the frequency of its semantic class in the training data.

Temperature in the InfoNCE loss

For InfoNCE loss we follow the idea of Qiu *et al.* [10] where frequent classes get higher temperatures to better capture the local semantic structure and less frequent classes get the lower temperatures to make their representations more separable.

For a semantic class c with K_c samples in class statistic K, the class temperature $\tau(c)$ is calculated as:

$$\tau(c) = \left(\frac{K_c - \min(K)}{\max(K) - \min(K)}\right) \cdot (\tau_{max} - \tau_{min}) + \tau_{min}. \tag{4.3}$$

Margin in the max-margin loss

Building on our observations regarding the influence of temperatures and margins on loss calculation, as discussed in Subsection 3.1.3, we adopt a similar approach for computing class-specific margins. The margin $m(c)$ for a semantic class c is defined as follows (Eq. 4.4):

$$m(c) = \left(\frac{K_c - \min(K)}{\max(K) - \min(K)}\right) \cdot (m_{max} - m_{min}) + m_{min}. \tag{4.4}$$

Class distribution sources

Datasets used for video-text retrieval tasks typically consist of video fragments paired with corresponding textual annotations. Many of these datasets are versatile and support additional tasks such as action recognition, instance segmentation, or question answering. Annotations for these auxiliary tasks often include class labels, which can be used as a source of information about the distribution of semantic classes.

However, some datasets lack such explicit class information, offering only video fragments and textual descriptions. In these scenarios, we extract semantic class distributions directly from the text annotations using an unsupervised approach. Specifically, we leverage a model capable of generating semantically meaningful sentence embeddings. These embeddings allow us to represent each annotation in a high-dimensional space, where semantic similarity can be quantified using cosine similarity. To approximate class distributions, we cluster the embeddings of all annotations in the training split, with each cluster corresponding to a semantic class. The distribution of semantic classes is estimated based on cluster sizes, defined by the number of samples within each cluster. This clustering-based method provides a proxy for class distribution, even in datasets without explicit semantic annotations.

4.1.3 Combination of two components

The final temperature for every sample in the batch is calculated as a sum of the semantic class temperature and correction temperature for the current training iteration (Eq. 4.5):

$$\tau_i = \tau(c)_i + \tau_{corr}(t). \tag{4.5}$$

The same idea is applied for final margin calculation (Eq. 4.6):

$$m_i = m(c)_i + m_{corr}(t). \tag{4.6}$$

Figure 4.2 illustrates the process of computing the class-specific temperature for each semantic class. The pipeline consists of the following steps:

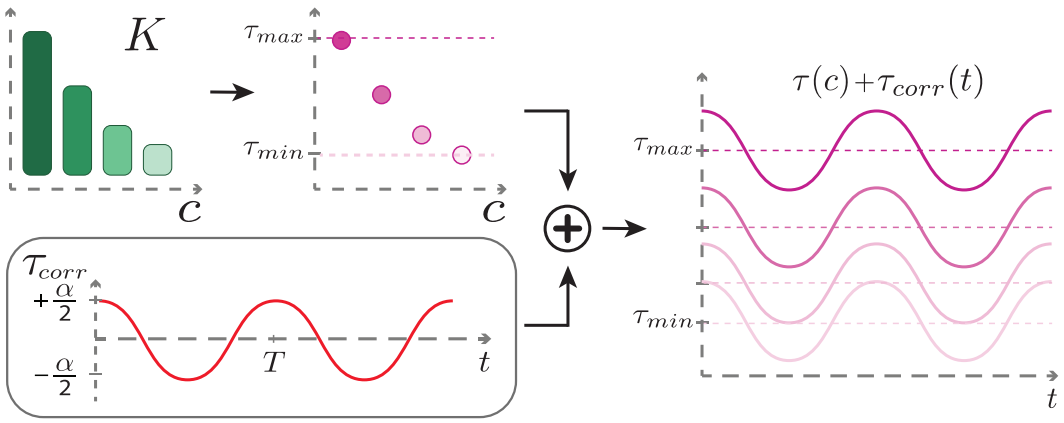

Figure 4.2: Detailed visualization of temperature calculation for every class c on every training iteration t based on the class distribution K, given temperature oscillation amplitude α and oscillation period T.

- First, each class is assigned a base temperature derived from its distribution in the training dataset, ensuring that frequent and rare categories are treated differently.
- A correction term is then applied, which evolves dynamically throughout training. In this illustration, a cyclic cosine schedule with two periods is used to modulate the correction term over time.
- The final temperature is obtained by summing these two components.

In the case of max-margin loss, the class-specific margins for each iteration are calculated in a similar manner.

4.1.4 Modified losses

We introduce slight modifications to the standard contrastive loss functions to incorporate individualized temperatures and margins into the training process. For the InfoNCE loss, the temperature parameter is now individualized per sample, leading to the following formulation (Eq. 4.7):

$$\mathcal{L}_{\text{InfoNCE}}(s) = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s_{ii}/\tau_i)}{\sum_{j=1}^{N}\exp(s_{ij}/\tau_i)}. \tag{4.7}$$

Similarly, we introduce an individualized margin for the max-margin loss, resulting in (Eq. 4.8):

$$\mathcal{L}_{\text{max-margin}}(s) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1,i\neq j}^{N}[\max(0,\, s_{ij} - s_{ii} + m_i)]. \tag{4.8}$$

In both equations, τ_i and m_i are adaptive parameters assigned to each sample. These values influence the loss calculations for positive and negative pairs of the corresponding anchor.

4.2 Illustration of the learning process

Figure 4.3 visually represents our approach. The right side of the figure depicts the standard contrastive learning framework, where two modality-specific encoders process text and video inputs to generate text embeddings (U) and video embeddings (V). These embeddings are then used to compute similarity scores.

On the left side, the figure highlights our proposed method. It demonstrates how semantic class distributions - derived from annotation embedding clusters - interacts with the cosine temperature schedule to compute the τ-vector. This individualized temperature is then combined with the similarity scores to compute the final loss, making the learning process more adaptive to long-tailed data distributions.

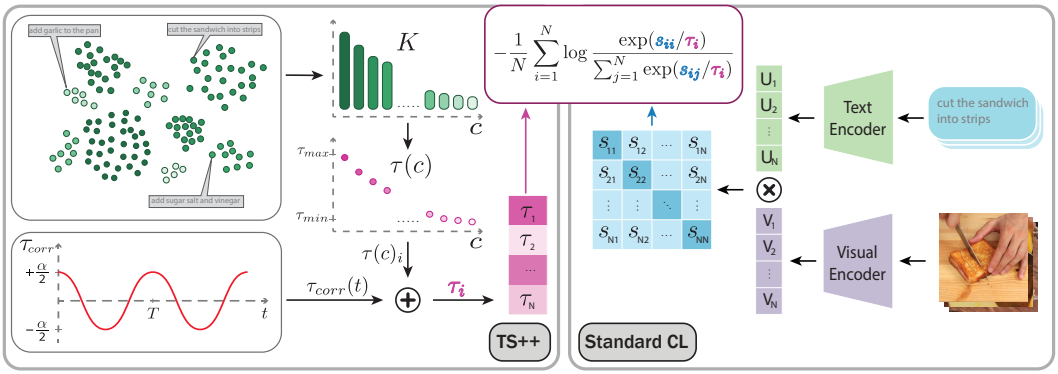

Figure 4.3: Visualization of our approach to computing the InfoNCE loss, incorporating semantic class distribution (based on annotation embedding clusters) and cosine temperature schedule into the process.

The same approach is applied to loss calculation with the max-margin loss, where the individualized margin is computed in a similar manner and incorporated into the learning process.

CHAPTER 5

Evaluation

This chapter evaluates the proposed method in the context of multimodal contrastive learning, focusing on long-tailed data scenarios. We begin by introducing the downstream tasks and associated evaluation metrics, focusing on the video retrieval and multi-instance retrieval tasks as benchmarks. Following this, we detail the models selected for evaluation, which serve as baselines for comparison. Subsequently, we provide an overview of the datasets used in our experiments, including EPIC-KITCHENS-100, YouCook2, and Something-Something-v2-LT. Key implementation details, including hardware configurations and training setups are also outlined. Finally, we present the experimental results, demonstrating the performance of our method in comparison with existing approaches. Through detailed ablation studies, we analyze the impact of specific components of the method on performance. This analysis demonstrates the strengths and limitations of our approach and its applicability to real-world scenarios involving long-tailed multimodal data.

5.1 Downstream tasks and evaluation metrics

Our method has the potential to be applied to various combinations of modalities and a wide range of contrastive learning tasks. In this study, we focus specifically on the text and video modalities, with an emphasis on video retrieval and multi-instance retrieval tasks. The following subsections provide a brief description of these tasks and their associated evaluation metrics.

5.1.1 Video retrieval task

Video retrieval is the task of identifying and retrieving relevant content from a database based on a given query [50]. It operates bidirectionally, including both text-to-video retrieval, where a textual query is used to find matching video content, and video-to-text retrieval, where a video is used as a query to find corresponding

textual descriptions. This process typically includes: Video representation extraction, text representation extraction, feature embedding and matching, and loss function[51]. The model learns to map related text and video features close to each other in the common embedding space. Retrieval results are generated by ranking similarities between video and text features. The most popular metrics for this task are recall at rank 1, 5 and 10 (R@1, R@5, R@10)

5.1.2 Multi-Instance retrieval task

Multi-instance retrieval is similar to the video retrieval task but extends it by incorporating the notion of relevance among the samples in the collection, in addition to text-video pairs. The objective is to assign the highest ranks to the most relevant samples. Performance in this task, including relevance assessment, is evaluated using metrics such as mean Average Precision (mAP) and normalized Discounted Cumulative Gain (nDCG) [14].

5.2 Models for evaluation

For the evaluation of our proposed methods, we select state-of-the-art models as baselines in the respective tasks of Video Retrieval and Multi-Instance Retrieval. Specifically, we use VAST [6] as a baseline for the Video Retrieval task, and for the Multi-Instance Retrieval task, we adopt AVION [8].

5.2.1 VAST

VAST [6] is an omni-modality video-text foundational model that works with video, audio, subtitles and text. VAST uses video, audio, and subtitles to represent the visual modality, enabling a richer and more detailed understanding of video content compared to models that rely solely on raw video frames. Its architecture is based on an end-to-end transformer structure that comprises three specialized encoders: a vision encoder for processing video frames, an audio encoder for handling audio (converted to spectrograms), and a text encoder that supports both unimodal and multimodal fusion through cross-attention layers. These encoders work together to extract global and local representations of each modality, which are then concatenated and projected into a shared semantic space for cross-modal alignment. Training the VAST model involves three key objectives. Firstly, the Omni-Modality Video-Caption Contrastive loss (OM-VCC) - a CLIP-like loss - aligns video and caption representations in the shared semantic space. Secondly, the Omni-Modality Video-Caption Matching loss (OM-VCM) ensures that the model can correctly identify matching video-caption pairs while employing a hard negative mining strategy for robust learning. Finally, the Omni-Modality Video Caption Generation loss (OM-VCG) enhances the model's ability to generate high-quality captions through masked language modeling.

VAST introduces significant innovations. It is the first model integrating vision, audio, subtitles, and text into a single framework, improving understanding of video content compared to previous models which focus on fewer modalities. This capability

is made possible by the VAST-27M dataset, a large-scale corpus of 27 million video clips, each paired with detailed multimodal captions. These captions are generated through an automated pipeline using modality-specific captioners and refined by a large language model. This dataset enables the model to excel in tasks requiring nuanced multimodal reasoning.

VAST demonstrates superior performance across a wide range of vision-text, audio-text, and multimodal video-text tasks, including retrieval, captioning, and question-answering. It outperforms state-of-the-art models on 22 benchmarks, proving its versatility and effectiveness.

5.2.2 LaViLa/AVION

The LaViLa [7] and AVION [8] frameworks are state-of-the-art video-language models designed to enhance video understanding through data generation techniques and efficient architecture. LaViLa focuses on leveraging large language models (LLMs) for creating detailed video narrations, addressing the sparsity of annotations in human-annotated video datasets. AVION adopts LaViLa's architecture and pipeline and focuses on optimizing training efficiency for large-scale video-language pretraining. Both frameworks use the Multi-Instance Max-Margin (MI-MM) loss for the Multi-Instance Retrieval task, as this loss can better handle the situation where a single text narration can be associated with multiple clips [44, 45].

LaViLa architecture and innovations

LaViLa (Language-model Augmented Video-Language Pretraining) introduces an architecture built around a dual-encoder model (similar to CLIP [5]). The video encoder uses TimeSformer [52] architecture. The spatial attention modules initialized with Vision Transformer [53], pre-trained on large-scale image-text data. The text encoder has a transformer architecture with twelve layers.

LaViLa's standout feature is its use of LLMs to generate dense, high-quality narrations for videos. This approach includes:

- **Narrator Module:** An LLM conditioned on video inputs generates rich textual descriptions, ensuring fine-grained and temporally aligned video-text annotations.
- **Rephraser Module:** A paraphrasing model diversifies narrations, augmenting the dataset with alternative textual descriptions of the same video content.

The framework uses both human annotations and generated narrations during contrastive learning.

These innovations enable LaViLa to leverage pseudo-supervised annotations, significantly expanding the effective training dataset. The framework achieves state-of-the-art performance across first- and third-person video tasks, including retrieval, classification and question-answering.

AVION architecture and innovations

AVION (A VIdeo model in ONe day) builds on the concept of LaViLa but focuses on computational efficiency. Key changes from LaViLa:

- **Memory-Efficient Vision Transformer:** AVION uses a Vision Transformer [53] enhanced with FlashAttention [54] to reduce memory complexity in sequence length from quadratic to linear. This optimization enables larger batch sizes and faster training without compromising performance.
- **Optimized Training Pipeline:** AVION uses chunk-based video loading and GPU-accelerated data augmentation to eliminate bottlenecks in video decoding and preprocessing.

These techniques allow AVION to train on four million video-text pairs in under 24 hours using a single 8-GPU server, achieving a 15× reduction in hardware cost compared to prior approaches. AVION achieves comparable or better accuracy with significantly lower computational costs. It outperforms previous methods on the Epic-Kitchens-100 dataset [14] for multi-instance retrieval and action recognition, demonstrating its capability for both zero-shot and fine-tuned tasks.

5.3 Datasets

In this work, we utilize three distinct datasets - EPIC-KITCHENS-100 [14], YouCook2 [15], and Something-Something-v2-LT [16] - each offering unique challenges and opportunities for learning robust representations in video-text retrieval tasks.

5.3.1 EPIC-KITCHENS-100

EPIC-KITCHENS-100 (EK-100) [14] is a large-scale video dataset, containing egocentric (first-person) videos with kitchen-based activities. The dataset captures natural, unscripted actions performed by participants in their own kitchens, providing a rich and realistic benchmark for understanding human-object interactions in daily life. EK-100 is widely used for a variety of tasks, including Action Recognition, Action Detection, Action Anticipation, and Multi-Instance Retrieval. In this work, we focus on the Multi-Instance Retrieval (MIR) task, which extends standard video-text retrieval by incorporating semantic relevancy between narrations. This task evaluates a model's ability to retrieve semantically related video segments based on textual queries. The dataset includes annotations for 97 verb classes (e.g., "cut," "open," "pour") and 300 noun classes (e.g., "carrot," "knife," "bowl"). The train split contains 67.2K annotated segments, while the validation split comprises 9.6K segments. Performance on the MIR task is measured using mAP and nDCG. These metrics capture the quality of retrieval, focusing on precision and the semantic relevance of the retrieved results. The relevance between is defined as the mean Intersection over Union (IoU) of the verb and noun classes, resulting in a value between 0 and 1. A value of 0 indicates no relevance (no overlap in verb or noun classes), while a value of 1 represents maximum relevance (complete overlap in verb and noun classes). Figure

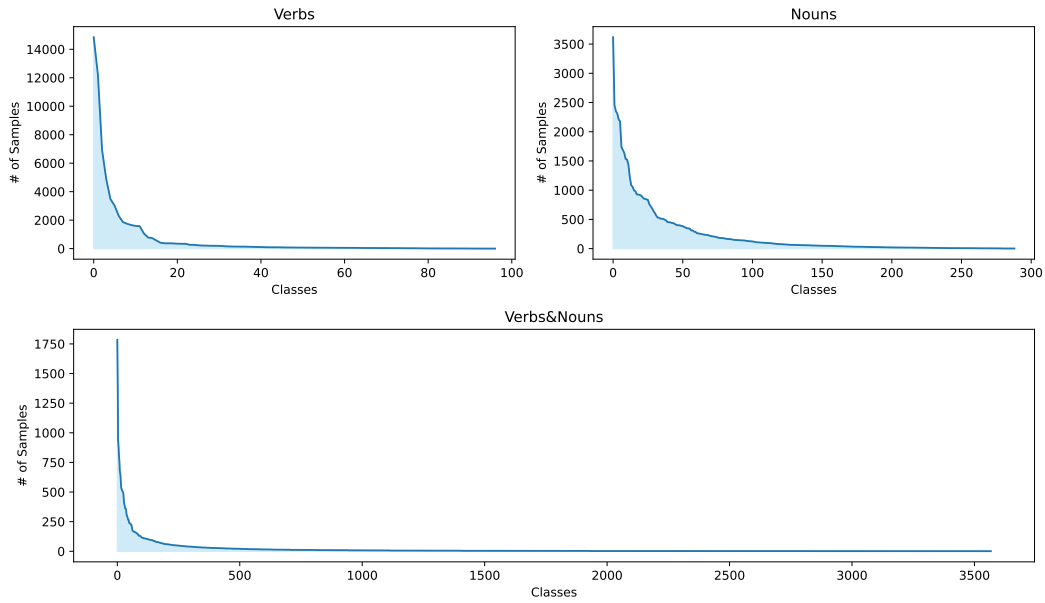

Figure 5.1: Visualizations of long-tail class distributions of Epic-Kitchens100 dataset training split. Class distributions are calculated based on verbs (top-left), important nouns (top-right) and unique combinations of verbs and nouns (bottom).

5.1 illustrates the distributions of verb classes, noun classes, and unique combinations of verbs and nouns within the training split of the dataset.

5.3.2 YouCook2

YouCook2 [15] is a large-scale video dataset focusing on cooking activities. It consists of 2000 cooking videos sourced from YouTube, annotated with textual descriptions and segmented into temporal segments corresponding to individual cooking steps. The dataset contains mainly the third person view videos. Each video is annotated with natural language descriptions that align with temporally segmented clips, where each segment represents an individual step in the cooking process. Training split contains 10K annotated video segments, validation split contains 3.5K annotated segments.[15]

As the dataset does not provide inherent labels or metadata directly suitable for extracting semantic class distribution, we employ the annotation clustering method outlined in Subsection 4.1.2. To better understand the semantic structure of the annotations, we generate embeddings of the training split using the SentenceBERT model [55]. A UMAP visualization of these embeddings is presented in Figure 5.2.

To calculate the semantic class distribution, we apply k-means clustering with $k = 200$, grouping the annotations into clusters that represent distinct semantic categories. The resulting distribution of clusters is illustrated in Figure 5.3.

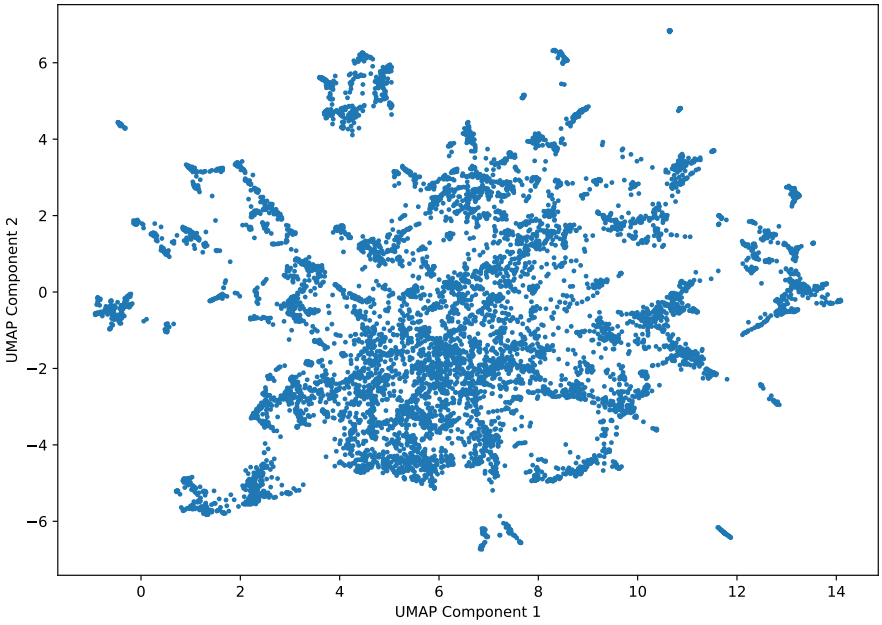

Figure 5.2: UMAP visualization of the annotation embeddings in YouCook2 dataset made using SentenceBert model [55].

5.3.3 Something-Something-v2-LT

Something-Something-v2 dataset [56] is a large collection of video clips showcasing individuals interacting with everyday objects through a set of predefined actions. Curated with contributions from a diverse group of crowd workers, the dataset is designed to foster the development of machine learning models capable of fine-grained action recognition in real-world scenarios. It comprises a total of 220,847 videos distributed across 168,913 training samples, 24,777 validation samples, and 27,157 test samples. The dataset has 174 action classes.

Something-Something-v2-LT (SSv2-LT) [16] is a long-tailed subset of the original Something-Something-v2 dataset, introduced by Perrett *et al.* [16]. The training split was resampled using a Pareto distribution with $\alpha = 6$, based on action classes, resulting in an imbalance ratio of 500. The distribution of the training split is illustrated in Figure 5.4. The validation and test splits remain balanced, containing 40 and 15 samples per class, respectively. [16]

5.4 Implementation Details

In all our experiments, we focus on fine-tuning the pretrained AVION and VAST models for the specified downstream tasks. Below are the hardware setups and dataset-specific parameters used in our implementation.

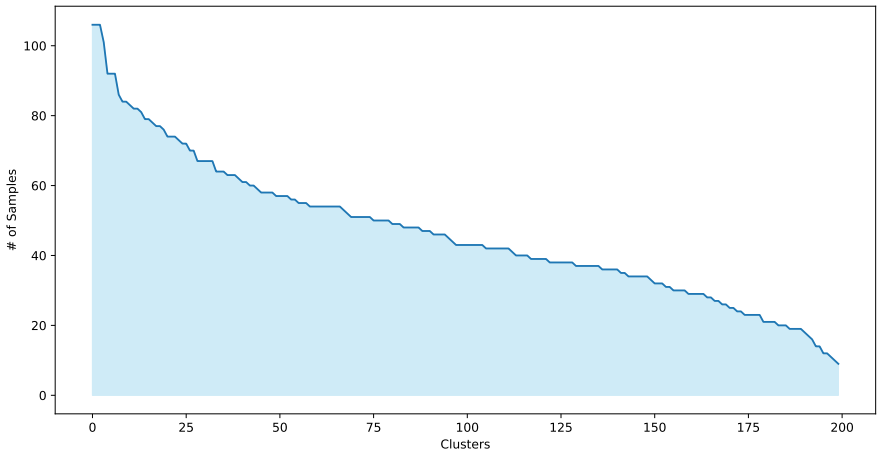

Figure 5.3: Visualization of long-tail annotations distribution of YouCook2 dataset. Annotations distribution is calculated based on k-mean clustering (200 clusters) of the annotation embeddings. Annotation embeddings are generated using SentenceBERT model [55].

Hardware Configuration

- **AVION Experiments:** Conducted on a system equipped with 4 NVIDIA RTX A6000 GPUs, each with 48 GB of memory.
- **VAST Experiments:** Performed on a system using 8 AMD Instinct MI210 GPUs, each with 64 GB of memory.

Dataset-Specific Parameters

- **EK-100:** We adhered to the fine-tuning protocol established by the AVION model [8]. This ensures consistency with prior work and enables a direct comparison of results.
- **YouCook2:** We utilized the fine-tuning configuration of the VAST model for this dataset [6]. To extract the semantic class distributions, we calculate the cluster distribution of text annotations, using the SentenceBERT model [55] the K-Means algorithm with $k = 200$.
- **SSv2-LT:** As VAST does not have a predefined configuration for this dataset, we base our setup on VAST defaults while making specific adjustments. The hyperparameters were set as follows: visual samples per video - 4, training epochs -10, batch size - 64.

In experiments utilizing only Temperature Schedule (TS) or Margin Schedule (MS), unless explicitly stated otherwise, the default values for τ and m are used.

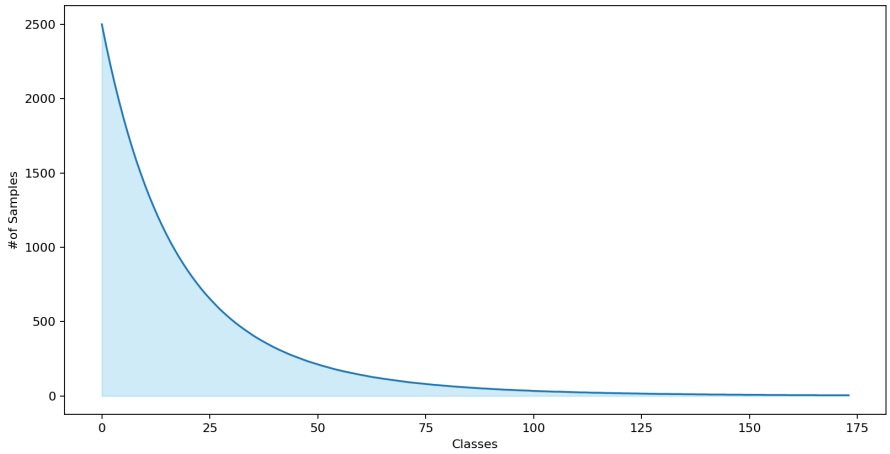

Figure 5.4: Visualization of action class distribution of SSv2-LT dataset.

Method	Back-bone	mAP			nDCG		
		V→T	T→V	Avg.	V→T	T→V	Avg.
MME [44]	TBN	43.0	34.0	38.5	50.1	46.9	48.5
JPoSE [44]	TBN	49.9	38.1	44.1	55.5	51.6	53.5
EgoVLP [45]	TSF-B	49.9	40.5	45.0	60.9	57.9	59.4
LaViLa [7]	TSF-B	55.2	45.7	50.5	66.5	63.4	65.0
AVION [8]	ViT-B	55.7	48.2	52.0	67.8	65.3	66.5
Ours	ViT-B	**59.2**	**49.4**	**54.3**	**68.9**	**66.3**	**67.6**

Table 5.1: Performance comparison of MIR on EPIC-Kitchens-100 dataset. The best iteration is selected based on mAP T→V metric.

Specifically, we set $\tau(c) = 0.07$ and $m(c) = 0.2$ as the baseline values for all semantic classes (c).

Similarly, in experiments focusing solely on Temperature Distribution (TD) or Margin Distribution (MD), the correction terms for temperature (τ_{corr}) and margin (m_{corr}) are set to zero ($\tau_{\text{corr}} = 0$, $m_{\text{corr}} = 0$) unless explicitly mentioned.

5.5 Results

First, we compare our method against state-of-the-art approaches for Multi-Instance Retrieval task on the EPIC-Kitchens-100 dataset. The results, presented in Table 5.1, demonstrate that our method achieves the highest performance across all evaluation metrics, including mAP and nDCG for both V→T and T→V directions, as well as their averages. Notably, our method achieves the highest values across all metrics, with the most significant improvement observed in mAP for the V→T direction, where it outperforms the previous best by 3.5%.

Table 5.2 presents a comparison of text retrieval performance on the YouCook2 dataset. Our method achieves the highest performance across all metrics, surpassing existing methods by a significant margin. In particular, our method achieves a Recall@1 score of 54.4%, which is a 4.0% absolute improvement over VAST [6], the previous state-of-the-art. Furthermore, improvements of 2.2% and 3.2% in Recall@5 and Recall@10, respectively.

Method	Text-to-Video		
	R@1	R@5	R@10
UniVL [41]	28.9	57.6	70.0
MELTR [42]	33.7	63.1	74.8
VLM [43]	27.1	56.9	69.4
VAST [6]	50.4	74.3	80.8
Ours	**54.4**	**76.5**	**84.0**

Table 5.2: Performance comparison on the YouCook2 dataset. The evaluation is done with Text-to-Video Recall@1,5,10. All benchmarks are multi-modal, containing both audio and subtitles.

5.6 Ablations

We conduct a set of ablation studies to empirically evaluate the effects of different components in our method.

5.6.1 Impact of TS/MS and TD/MD on the learning performance

In the table 5.3 we analyze the impact of two main components of our method (TS/MS and TD/MD) on the training performance in MIR task on EK-100 dataset.

The authors of [45] show superiority of MI-MM loss compared to CLIP loss for the MIR task, where multiple videos can correspond to a single annotation. However, we decided to check the performance of AVION with the CLIP-loss (same as used for pre-training) against the variants of CLIP-loss with temperature schedule, temperature distribution and a combination of both. MS and TS denote cyclic cosine or linearly increasing schedules for the margin and temperature, respectively. MD and TD represent distribution-based individualization of margin and temperature applied during training.

Using MS and MD with MI-MM results in consistent improvements in both mAP and nDCG scores, with the combination of linear MS and MD achieving the best performance (e.g., an average mAP of 54.3 compared to 51.4 for the base MI-MM). Replacing the MI-MM loss with CLIP loss and using TS/TD also improves performance. For instance, incorporating both TS and TD yields an average mAP of 49.2, significantly outperforming the baseline using CLIP-loss (43.1). Scheduling techniques (cosine or linear) generally outperform static approaches, with linear schedules achieving marginally better results across metrics. These findings suggest that the proposed scheduling and distribution-based techniques enhance the

Method	mAP			nDCG		
	V→T	T→V	Avg.	V→T	T→V	Avg.
MI-MM	55.3	47.6	51.4	67.6	64.8	66.2
w/ cos MS	58.4	48.6	53.6	**69.4**	64.8	66.9
w/ linear MS	57.7	**49.8**	53.7	68.6	66.1	67.4
w/ MD	56.4	48.4	52.4	68.2	66.0	67.1
w/ cos MS&MD	58.8	48.9	53.9	68.9	65.8	67.3
w/ linear MS&MD	**59.2**	49.4	**54.3**	68.9	**66.3**	**67.6**
CLIP	47.0	39.8	43.1	64.4	61.2	62.8
w/ cos TS	**55.0**	41.7	48.4	**69.1**	64.6	**66.8**
w/ TD	52.2	43.6	47.8	68.0	**64.8**	66.4
w/ cos TS&TD	54.6	**43.9**	**49.2**	68.6	64.4	66.4

Table 5.3: Performance comparison of AVION on EK-100 dataset (MIR) with modified MI-MM and CLIP - losses. MS/TS and MD/TD refer to margin/temperature schedule and margin/temperature distribution respectively. For MS cos and linear refer to the cyclic cosine and linearly growing margin schedules types.

discriminative power of the learned representations, leading to improved MIR task performance on the EK-100 dataset.

In table 5.4, we extend our analysis to the SSv2-LT dataset. These results exclude refinement steps (as VAST additionally uses matching loss for it) to isolate the impact of contrastive loss modifications. Baseline results with original VAST [6] are compared against variants employing TS, TD, and their combinations. Introducing TS and TD to VAST leads to consistent performance gains across all metrics. For example, the combination of cosine TS and TD achieves the highest R@5 (71.0) and R@10 (79.7) for Text-to-Video retrieval. Linear scheduling slightly outperforms cosine scheduling in certain metrics (e.g., R@1 for Video-to-Text retrieval).

Method	Text-to-Video			Video-to-Text		
	R@1	R@5	R@10	R@1	R@5	R@10
VAST [6]	41.7	69.1	78.9	41.2	69.3	79.3
w/ cos TS	42.6	70.0	79.5	41.3	68.7	79.3
w/ linear TS	43.0	70.1	79.4	42.8	69.8	79.3
w/ TD	42.7	70.3	79.6	41.2	69.7	79.5
w/ cos TS&TD	**43.2**	**71.0**	**79.7**	41.8	**70.3**	**80.0**
w/ linear TS&TD	**43.2**	70.3	**79.7**	**43.0**	**70.3**	79.8

Table 5.4: Performance comparison of variants of VAST on finetuning on SSv2-LT dataset. Results without refinement.

Distribution source	mAP			nDCG		
	V→T	T→V	Avg.	V→T	T→V	Avg.
verbs & nouns	58.2	48.7	53.4	68.3	**65.5**	66.9
verbs	58.5	48.7	53.6	68.0	65.3	66.7
nouns	**58.8**	**48.9**	**53.9**	**68.9**	65.3	**67.3**

Table 5.5: Performance comparison of AVION with margin distribution method based on the distribution of unique verbs+nouns combinations, verbs and nouns

5.6.2 Margin distribution based on distributions of verbs and nouns in EK-100

Here, we compare the results of the MIR on EK-100 finetuning using different sources of the class distribution. The table 5.5 shows that utilizing nous distribution as source for class distribution, leads to better results in most of the metrics. That finding supports our hypothesis about the importance of object-aware multimodal distribution.

5.6.3 Comparison of cosine and linear schedules

In this subsection, we compare the effects of a cyclic cosine margin schedule, inspired by the findings of Kukleva *et al.* [11], with a linear margin schedule, motivated by Yaras *et al.* [12]. In these experiments, we finetune the pretrained AVION model on the EK-100 dataset for the MIR task with the static, cyclic cosine, and linearly growing margin schedules. For simplicity, distribution-based pseudo-supervision with margin was not applied in these experiments. The results in Figure 5.5 show the progress of finetuning through 100 epochs. Both margin schedules (cosine MS and linear MS) greatly outperform the baseline with a static margin. For the video-to-text MIR task, the performances of cosine MS and linear MS are similar. Nonetheless, linear MS demonstrates superior results in text-to-video retrieval. Indeed, the cosine MS experiment yields its best performance when the margins are largest, which aligns with the findings of Yaras *et al.* [12], which shows the effectiveness of growing temperature for mitigation of modality gap.

To explore the performance variations between cosine MS and linear MS in the context of text-to-video MIR, we employ Principal Component Analysis (PCA) to visualize the embedding space of the evaluation dataset. This exploration indicates that samples from the text modality create more compact clusters in the embedding space than their video counterparts. It can be explained by the reduced variance in text data when compared to video data.

From the perspective of contrastive learning, the better performance of the linear MS in text-to-video retrieval can be explained as follows:

- The linear schedule incrementally raises the margin during training, resulting in a more stable optimization process. These gradual adjustments help the model to focus on fine-grained differences within text clusters, facilitating improved alignment with related video embeddings.

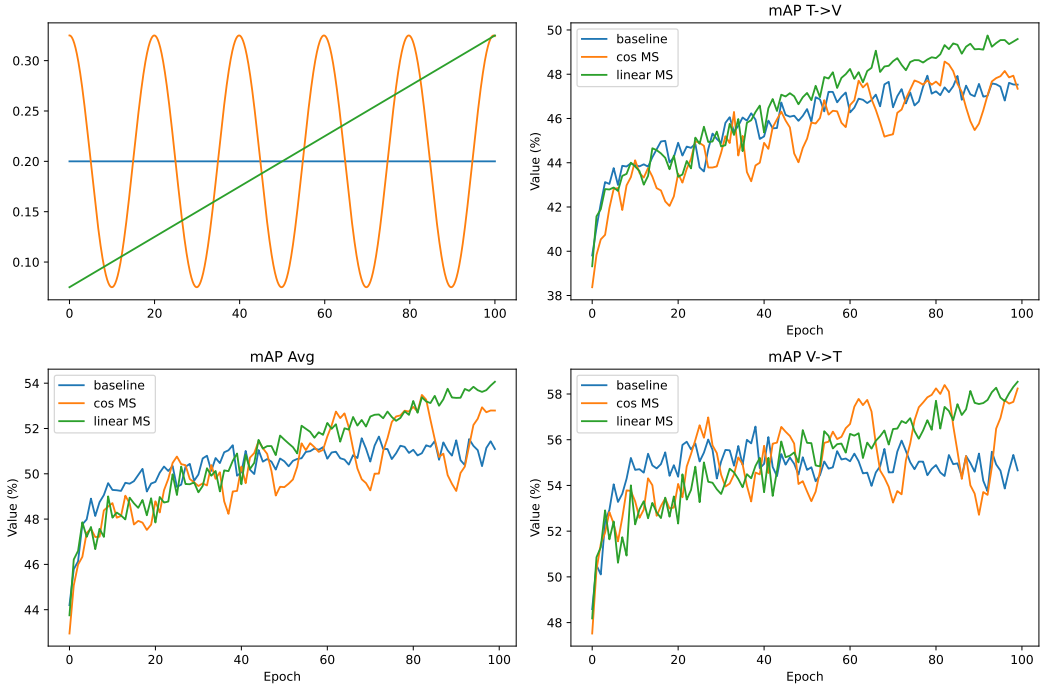

Figure 5.5: Visualization of the mAP metrics (text-to-video, video-to-text, and their average) collected during the training of the AVION framework on the EPIC-KITCHENS-100 dataset. The blue line represents the baseline experiment with a static margin, the orange line corresponds to the cyclic cosine margin schedule (cosine MS), and the green line represents the linearly growing margin schedule (linear MS).

- Periodical shifts to lower margin values create a widening modality gap, which disrupts the alignment process for closely packed text embeddings.

5.6.4 Experiments with asymmetric TS

In unimodal contrastive learning, as described in 3.2, the outer sum ensures that each sample within the input set serves as an anchor, leading to the inclusion of relations across all $N \times N$ possible pairs within the given input set. This comprehensive pairwise comparison facilitates a uniform distribution of representations in the normalized embedding space, particularly when a small temperature parameter is used [11, 18].

In contrast, multimodal contrastive learning, as formulated in 3.6, incorporates only cross-modal relations within the loss function. Unlike the unimodal case, a small temperature in this setup often encourages stronger separation between the modalities, thereby increasing the modality gap [13] instead of achieving uniform coverage of the embedding space.

We hypothesize that this behavior is a consequence of the instance-level discrimination applied to each cross-modal pair in the loss function, which increases the distance between the two modalities in the embedding space. We construct a series of experiments to test this hypothesis, seeking to characterize the structures of

multimodal embedding spaces. Specifically, we independently alter the temperature parameter τ for each direction in the CLIP-loss. We denote $\tau_{t\to v}$ to be the temperature parameter for $\mathcal{L}_{\text{InfoNCE}}(s_{t\to v})$ (loss calculation from text to video modalities), and $\tau_{v\to t}$ to be the temperature for $\mathcal{L}_{\text{InfoNCE}}(s_{v\to t})$ (loss calculation from video to text modality).

Our experiments are based on the VAST multimodal framework [6] without a refinement step (to isolate the effect of contrastive loss) and are conducted on the SSv2-LT dataset [16]. In the first experiment, we apply the cyclic cosine temperature schedule symmetrically in both directions ($\tau_{t\to v}\&\tau_{v\to t} = cos$). The values of ($\tau_{t\to v}$) and ($\tau_{v\to t}$) are modulated using a cosine schedule, oscillating in range [0.04, 0.1] over three periods during the training. This dynamic adjustment encourages the model to learn both 'easy' and 'hard' semantic features in the embedding space [11], effectively controlling the 'tightness' of semantic clusters [11, 18]. In the following two experiments we apply the temperature schedule asymmetrically (only in one direction. In the second experiment we set $\tau_{v\to t}$ to a constant value of 0.07, which has been shown to achieve optimal downstream performance in standard settings. The value of $\tau_{t\to v}$ is modulated in the same way as for the first experiment

In the third experiment, we reverse the setup of the second experiment by fixing $\tau_{t\to v}$ to 0.07 and modulating $\tau_{v\to t}$ using the same cosine schedule. By comparing these three setups, we aim to investigate how asymmetric modulation of the temperature parameter affects the alignment and structure of the multimodal embedding space.

In figure 5.6, we summarize the findings of these three experiments, the first experiment ($\tau_{t\to v}\&\tau_{v\to t} = cos$) in blue, the second experiment ($\tau_{t\to v} = \cos$, $\tau_{v\to t} = 0.07$) in orange and the third experiment ($\tau_{t\to v} = 0.07$, $\tau_{v\to t} = cos$) in green.

We find that all three losses, and hence retrieval performance, follow a cosine schedule, even when only one loss asymmetrically uses a changing temperature. Referring to the extreme temperature values (the lowest and the highest), it is interesting to note the counter-phase relationship of text-to-video and video-to-text retrieval results when the temperature is changed. In particular, in the low τ phase (dashed line), text-to-video retrieval for the 2nd experiment reaches its lowest point, while video-to-text retrieval simultaneously achieves its highest. Our observation is aligned with the findings of Liang *et al.* [57], as the modality gap also peaks at the lowest temperature, demonstrating how it dynamically changes with temperature variations. In the second experiment, the low τ phase appears to cause each video representation to repel all text representations.

Further investigation is needed to reconcile the apparent discrepancy between video-to-text loss during training and video-to-text retrieval performance on the test set. While higher video-to-text loss generally indicates poorer performance during training, the observed improvement in video-to-text retrieval during the low τ phase suggests a more nuanced relationship. A similar pattern can be observed for text-to-video loss and text-to-video retrieval, where an increase in text-to-video loss during the low τ phase is accompanied by an improvement in text-to-video retrieval performance.

Summary:

- Large τ tends to mitigate the modality gap, generally leading to improved

evaluation performance of the other modality.

- Small temperature in a multimodal (in contrast to unimodal settings) setting does not guarantee uniform distribution. Instead, it can increase the modality gap, leading to tighter clusters within each modality.

5.7 Analysis

Our method leads to significant performance gains in Video Retrieval and Multi-Instance Retrieval tasks. It outperforms state-of-the-art approaches on benchmark datasets.

Our ablation studies show the importance of each of the components - temperature/margin schedules and distribution based pseudo-supervision. Interestingly, we discovered that the method performance is influenced by the choice of semantic class distribution source, with the object-aware distribution producing the best results. Moreover, our comparison of cosine and linear schedules reveals that the type of schedule can impact retrieval performance, with linear schedules having an advantage in text-to-video MIR settings.

In the experiments with asymmetric temperature schedules we observe a counter-phase phenomenon between text-to-video and video-to-text retrieval when the temperature is asymmetrically modulated. This illustrates the intricate relationship between temperature, modality gap, and retrieval performance.

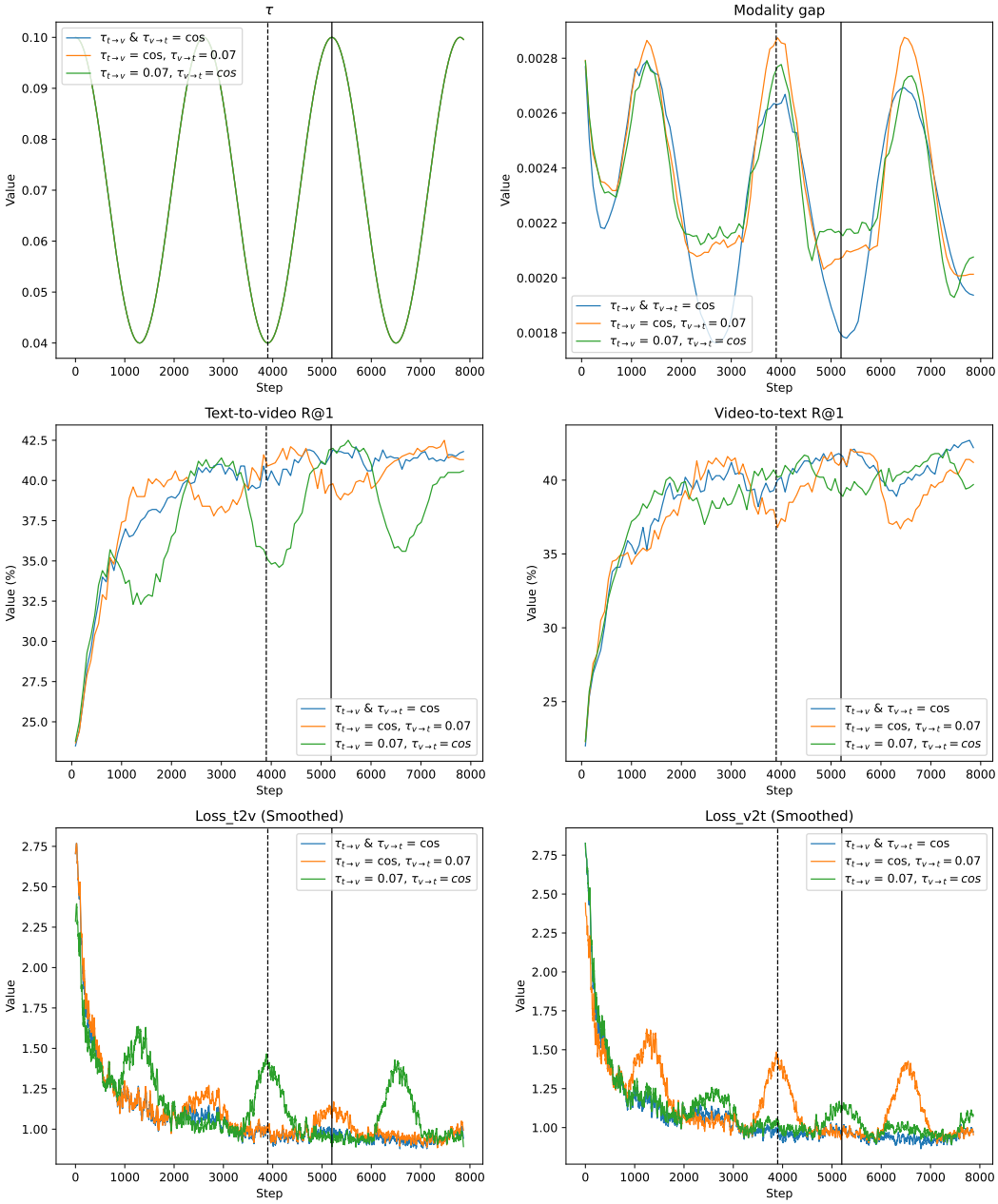

Figure 5.6: Visualization of relationships between asymmetrically applied temperature, modality gap, retrieval performance and losses in both directions.

CHAPTER 6

Summary and Future Work

6.1 Conclusion

This thesis project introduces and evaluates a novel method, Temperature Schedules++, aimed at addressing the challenges of multimodal contrastive learning in long-tailed data scenarios. By leveraging temperature and margin scheduling in conjunction with distribution-aware adjustments, TS++ provides a robust solution to improve representation quality and retrieval performance.

The efficiency of TS++ has been demonstrated across video-text retrieval tasks on EPIC-KITCHENS-100, YouCook2, and Something-Something-v2-LT datasets. Below, we summarize the key findings of this work:

- **Effectiveness of temperature and margin schedules:** In this work, we successfully adapt the cyclic cosine temperature schedule, originally proposed for unimodal contrastive learning on long-tailed data, to the multimodal setting. In addition, we made a comparison with the linear temperature schedule which was specifically designed to mitigate the modality gap. Both scheduling strategies demonstrated similar performance, effectively improving the alignment of multimodal embeddings in the shared representation space. Based on our analysis of the parallels between the temperature parameter in InfoNCE loss and the margin parameter in max-margin loss, we extended these scheduling strategies to margin scheduling. More specifically, we employ both cyclic cosine and linear schedules to dynamically adjust the margin during the training. Our empirical results on the EPIC-KITCHENS-100 dataset validate the effectiveness of this approach.

- **Distribution-aware Pseudo-supervision:** It was proven that adapting temperature and margin parameters based on the semantic frequency of training samples helps to improve long-tailed dataset performance for underrepresented classes. This approach uses not only class labels provided in the datasets

(often used for auxiliary tasks such as object detection, action classification and semantic segmentation), but also embeddings of text annotation for training split of the dataset. By grouping text annotation embeddings based on their cosine similarity, we extract information about the frequency of semantic classes only from textual data. It makes our approach more flexible, enabling its usage in the absence of explicit class labels.

- **Combination of both components:** While temperature and margin scheduling and distribution-aware adjustments independently showed significant performance improvements over the baseline, the combination of both methods resulted in the highest overall gains. TS++ consistently outperformed state-of-the-art methods, achieving higher mAP, nDCG, and Recall metrics in both Multi-Instance Retrieval and Text-to-Video Retrieval tasks.

- **Scalability and generalization:** The modular design of TS++ allows seamless integration into existing multimodal contrastive learning frameworks, as evidenced by its successful application to both AVION and VAST models. This demonstrates its versatility and potential for broader applicability.

6.2 Future Work

Multimodal contrastive learning has gained great attention in recent years due to its potential to bridge the gap between different modalities such as photo, video, text and audio. These models find applications in various task, including image and video retrieval, question answering, recommendation systems, and content understanding. However, multimodal CL still faces the persistent challenges of modality gaps and unbalanced data distributions. To overcome these obstacles and improve the effectiveness of these methods, we outline several promising avenues for future research.

In this thesis project, TS++ (and its individual components) is applied exclusively during the fine-tuning stage for retrieval tasks. Future work could investigate whether the method is equally effective when incorporated into the pretraining phase or applied to fine-tuning for other downstream tasks and to the different combination of modalities.

The other important direction is a more extensive study of asymmetric temperature schedules. We found out through our experiments that dynamically changing the temperature in one direction (*e.g.* , from the first modality to the second) affects not only the performance of the adjusted direction but also the opposite direction. Notably, these changes appear to occur in the anti-phase, suggesting that temperature adjustments in one direction influence alignment dynamics across modalities. This finding requires more research on asymmetric temperature schedules, with the potential to design individualized schedules for each direction between modalities.

Based on experiments with cyclic cosine temperature and margin schedules, we determine that the highest performance occurs when the temperature or margin is at its highest value towards the end of training. Such a conclusion closely matches the

observations of Yaras *et al.* [12], which showed that training with linearly increasing temperature can effectively narrow the modality gap. The original reason for exploring cyclic cosine temperature schedules in this work was motivated by the findings of Kukleva *et al.* [11], who highlighted their utility in unimodal contrastive learning on unbalanced data. However, the role of periodic temperature changes in multimodal settings remains an open question. It is unclear whether such cyclic schedules in multimodal settings are inducing dynamic instance-wise and group-wise discrimination, or if the improvement is solely due to the increasing temperature which helps to mitigate the modality gap. Future research should focus on disentangling these effects to better understand the underlying mechanisms of temperature scheduling in multimodal contrastive learning.

List of Figures

List of Tables

Bibliography

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[2] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[4] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[6] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems*, 36:72842–72866, 2023.

[7] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597, 2023.

[8] Yue Zhao and Philipp Krähenbühl. Training a large video model on a single machine in a day. *arXiv preprint arXiv:2309.16669*, 2023.

[9] Ngoc Dung Huynh, Mohamed Reda Bouadjenek, Sunil Aryal, Imran Razzak, and Hakim Hacid. Visual question answering: from early developments to recent advances–a survey. *arXiv preprint arXiv:2501.03939*, 2025.

[10] Zi-Hao Qiu, Quanqi Hu, Zhuoning Yuan, Denny Zhou, Lijun Zhang, and Tianbao Yang. Not all semantics are created equal: Contrastive self-supervised learning with automatic temperature individualization. *arXiv preprint arXiv:2305.11965*, 2023.

[11] Anna Kukleva, Moritz Böhle, Bernt Schiele, Hilde Kuehne, and Christian Rupprecht. Temperature schedules for self-supervised contrastive methods on long-tail data. *arXiv preprint arXiv:2303.13664*, 2023.

[12] Can Yaras, Siyi Chen, Peng Wang, and Qing Qu. Explaining and mitigating the modality gap in contrastive multimodal learning. *arXiv preprint arXiv:2412.07909*, 2024.

[13] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

[14] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022.

[15] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[16] Toby Perrett, Saptarshi Sinha, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Use your head: Improving long-tail video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2415–2425, 2023.

[17] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[18] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2021.

[19] Rui Cao, Yihao Wang, Yuxin Liang, Ling Gao, Jie Zheng, Jie Ren, and Zheng Wang. Exploring the impact of negative samples of contrastive learning: A case study of sentence embedding. *arXiv preprint arXiv:2202.13093*, 2022.

[20] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Mining on manifolds: Metric learning without labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7642–7651, 2018.

[21] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[22] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021.

[23] Siladittya Manna, Soumitri Chattopadhyay, Rakesh Dey, Saumik Bhattacharya, and Umapada Pal. Dystress: Dynamically scaled temperature in self-supervised contrastive learning. *arXiv preprint arXiv:2308.01140*, 2023.

[24] Tiffany Tianhui Cai, Jonathan Frankle, David J Schwab, and Ari S Morcos. Are all negatives created equal in contrastive instance discrimination? *arXiv preprint arXiv:2010.06682*, 2020.

[25] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

[26] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016.

[27] Hao Zhang, Zheng Li, Jiahui Yang, Xin Wang, Caili Guo, and Chunyan Feng. Revisiting hard negative mining in contrastive learning for visual understanding. *Electronics*, 12(23):4884, 2023.

[28] Yifei Wang, Qi Zhang, Yaoyu Guo, and Yisen Wang. Non-negative contrastive learning. *arXiv preprint arXiv:2403.12459*, 2024.

[29] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[30] Yifei Wang, Jizhe Zhang, and Yisen Wang. Do generated data always help contrastive learning? *arXiv preprint arXiv:2403.12448*, 2024.

[31] Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2020.

[32] Hong Liu, Jeff Z HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance. *arXiv preprint arXiv:2110.05025*, 2021.

[33] Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. *Advances in neural information processing systems*, 33:19290–19301, 2020.

[34] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 715–724, 2021.

[35] Jie Miao, Junhai Zhai, and Ling Han. Contrastive dual-branch network for long-tailed visual recognition. *Pattern Analysis and Applications*, 28(1):10, 2025.

[36] Chaoqun Du, Yulin Wang, Shiji Song, and Gao Huang. Probabilistic contrastive learning for long-tailed visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[37] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

[38] Maria A Bravo, Sudhanshu Mittal, Simon Ging, and Thomas Brox. Open-vocabulary attribute detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7041–7050, 2023.

[39] Yuhui Zhang, Elaine Sui, and Serena Yeung-Levy. Connect, collapse, corrupt: Learning cross-modal tasks with uni-modal data. *arXiv preprint arXiv:2401.08567*, 2024.

[40] Abrar Fahim, Alex Murphy, and Alona Fyshe. Its not a modality gap: Characterizing and addressing the contrastive gap. *arXiv preprint arXiv:2405.18570*, 2024.

[41] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.

[42] Dohwan Ko, Joonmyung Choi, Hyeong Kyu Choi, Kyoung-Woon On, Byungseok Roh, and Hyunwoo J Kim. Meltr: Meta loss transformer for learning to fine-tune video foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20105–20115, 2023.

[43] Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. VLM: Task-agnostic video-language model pre-training for video understanding. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4227–4239, Online, August 2021. Association for Computational Linguistics.

[44] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 450–459, 2019.

[45] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022.

[46] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[47] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[48] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[49] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[50] BV Patel and BB Meshram. Content based video retrieval systems. *arXiv preprint arXiv:1205.1641*, 2012.

[51] Cunjuan Zhu, Qi Jia, Wei Chen, Yanming Guo, and Yu Liu. Deep learning for video-text retrieval: a review. *International Journal of Multimedia Information Retrieval*, 12(1):3, 2023.

[52] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[53] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020.

[54] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

[55] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[56] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

[57] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, 2022.

Typeset January 29, 2025