This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

MM-TS: Multi-Modal Temperature and Margin Schedules for Contrastive Learning with Long-Tail Data

Siarhei Sheludzko[1] Dhimitrios Duka[2] Bernt Schiele[2] Hilde Kuehne[3,4] Anna Kukleva[2]
[1]University of Bonn [2]MPI for Informatics, SIC [3]Tuebingen AI Center/University of Tuebingen [4]MIT-IBM Watson AI Lab

Abstract

Contrastive learning has become a fundamental approach in both uni-modal and multi-modal frameworks. This learning paradigm pulls positive pairs of samples closer while pushing negatives apart. In the uni-modal setting (e.g., image-based learning), previous research has shown that the strength of these forces can be controlled through the temperature parameter. In this work, we propose Multi-Modal Temperature and Margin Schedules (MM-TS), extending the concept of uni-modal temperature scheduling to multi-modal contrastive learning. Our method dynamically adjusts the temperature in the contrastive loss during training, modulating the attraction and repulsion forces in the multi-modal setting. Additionally, recognizing that standard multi-modal datasets often follow imbalanced, long-tail distributions, we adapt the temperature based on the local distribution of each training sample. Specifically, samples from dense clusters are assigned a higher temperature to better preserve their semantic structure. Furthermore, we demonstrate that temperature scheduling can be effectively integrated within a max-margin framework, thereby unifying the two predominant approaches in multi-modal contrastive learning: InfoNCE loss and max-margin objective. We evaluate our approach on four widely used image- and video-language datasets, Flickr30K, MSCOCO, EPIC-KITCHENS-100, and YouCook2, and show that our dynamic temperature and margin schedules improve performance and lead to new state-of-the-art results in the field. [1]

1. Introduction

Contrastive learning has become one of the most important tools in self-supervised learning, allowing models to learn representations from single modalities, such as images [2, 15, 56], as well as from multiple modalities, such as vision and language [5, 42]. By aligning different modalities of the same instance in a joint embedding space, these models achieve state-of-the-art performance, e.g., in the context of zero-shot classification, image retreival [9, 13, 39] and video retrieval [49, 58, 59], but also serve as a representation backbone for autoregressive frameworks [29, 30, 33].

Contrastive learning aims to bring positive pairs closer together, such as different augmentations of the same image in the uni-modal case or image-caption pairs in the multi-modal case, while pushing apart all other samples, so called negatives. In particular, in the multi-modal setting, contrastive learning aligns representations from two different modalities by minimizing the cosine distance between matching text-image pairs. This process is controlled by the temperature parameter τ in InfoNCE-based losses or by the margin definition in the max-margin formulations, both of which control the strength of attraction and repulsion forces in the embedding space.

While most approaches simply fix the temperature parameter during training [2, 3, 16] or consider it a constant hyperparameter, recent works also started to explore how to vary or adapt it to improve contrastive training itself [21, 23, 47]. Among them, temperature schedules [23] (TS) explore the effect of controlling temperature to improve contrastive self-supervised learning for imbalanced distributions. Low temperatures amplify pushing forces of negatives, leading to stronger discrimination between individual instances, while a higher temperature exerts less force on negative samples thus enhancing grouping based on semantic structure. By dynamically scheduling the temperature during training, the model continuously learns different semantic features: lower temperatures enhance instance discrimination, while higher temperatures lead to group-wise structure formation. Training under this regime has shown to be valuable as this is especially important for single-modality imbalanced distributions [23].

In this work we propose a Mutli-Modal Temperature and Margin Schedules (MM-TS) framework for multi-modal contrastive learning on long-tail data. Our MM-TS leverages the idea of temperature schedules and adapts it for the multi-modal scenario while at the same time including in-

[1]The code is publicly available at https://github.com/SergShel/MM-TS

formation about the local distribution of the multi-modal data. To this end, first, we employ a dynamic temperature scheduling for both modalities that follows a cosine schedule. Continuous changes of the embedding space allow the model to learn different semantic features and improve its alignment compared to a fixed temperature. Second, we leverage the availability of the two modalities and approximate the distribution of the visual training data by using the text modality. We propose to control the amplitude of the temperature parameter based on the distributions of the training data. Namely, when the corresponding narration contains an object or a concept that frequently occurs in the dataset, we assign it a higher value. This allows the creation of larger clusters in the representation space for those samples, so called group-wise discrimination effect, whereas pairs with rare and unique narrations would repel with a higher strength from all other pairs enforcing instance discrimination.

Finally, we extend the temperature scheduling formulation from the original InfoNCE loss function to the max-margin contrastive loss setting. We observe similar behavior that indicates the importance of the repulsion forces of negatives as the max-margin loss does not explicitly control strength of positive pulling.

We evaluate the influence of the proposed MM-TS framework for the case of long-tail image- and video-language learning, considering image-text settings by pre-training on CC3M with the zero-shot retrieval on Flickr30k and COCO, and video-text datasets such as the fine-grained long-tail EPIC-KITCHENS-100 dataset [8], and YouCook2 dataset [60] . To this end, we extend one of the best performing contrastive methods of each framework and retrain it with the proposed MM-TS extensions.

Our contributions can be summarized as following:

- We propose a new multi-modal framework for contrastive learning on long-tail data that controls temperature parameters by combining cosine temperature schedule with individual adjustments for each sample based on the estimated distributions;
- We generalize temperature scheduling beyond the conventional InfoNCE framework, extending it to the widely adopted max-margin loss. This extension is particularly impactful within communities such as egocentric video analysis, where max-margin loss remains a prevalent choice and datasets are long-tailed.
- We evaluate the proposed apporach across various naturally distributed long-tail image and video datasets such as CC3M, YouCook2 and EPIC-KITCHENS-100.

2. Related work

Uni-modal Contrastive Learning (CL) is a self-supervised learning approach that aims to learn semantically meaningful representations from a single modality, e.g. images, text, or audio, by leveraging an objective that maximizes the similarity between positive pairs, usually constructed by augmenting the original data sample, and minimizes the similarity between negative pairs. Building on the importance of negative samples in contrastive learning [17, 20, 53, 56], MoCo [3, 16] introduces CL with momentum encoder and memory bank in the form of a FIFO queue, used to store generated features from the momentum encoder which are later used to construct negative pairs for CL. SimCLR [2] refines MoCo's approach by eliminating the momentum and dependency on the memory bank by increasing the batch size and introducing a projection head.

Multi-Modal Contrastive Learning. The contrastive learning methods have also been used to learn semantically rich representations in the context of multimodal data. CLIP [42] constructs a joint representation space by employing the InfoNCE objective in multi-modal embedding and learning an alignment for image-text pairs. This simple yet effective approach enables zero-shot transfer to many tasks and datasets. Different methods have aimed to improve CLIP's performance by applying architectural changes [31], refining the optimization process [54], or redefining the objective [7, 24, 34, 35, 57]. SogCLR [54] introduces the idea of a global contrastive objective that aims to contrast a positive pair against all negative pairs in the dataset. CyCLIP [11] addresses the issue of inconsistent image and text representations in CLIP-like models by introducing an extension of the classical InfoNCE loss and proposing two additional objectives to ensure geometric consistency within and between modalities. SigCLIP [55] replaces the classical softmax-based InfoNCE objective with a pairwise sigmoid objective. The sigmoid objective is introduced for image-text pairs, thus eliminating the need for normalization across an entire mini-batch, resulting in less computation, higher efficiency, and scalability.

Temperature in Contrastive Learning. Recent works [21, 23, 25, 41, 47, 48, 57] have shown that temperature τ has a great impact on CL, influencing both the uniformity and the alignment of the embedding space. [47] showed that the contrastive objective is hardness-aware. At small temperatures, the objective penalizes the hard negatives more, compared to soft negatives. This behavior results in instance discrimination, making the embedding space more uniform at the cost of a less aligned embedding space. At high temperatures, the disparity in penalties between soft and hard negatives is reduced, therefore all negatives are penalized more uniformly. This encourages the formation of clusters and trades-off uniformity for alignment [23]. TempNet [41] proposed using an additional temperature prediction network to assign a unique temperature per sample. Even though the idea of using a network to predict a per-sample temperature was considered before [25, 48, 57],

TempNet introduces a novel design based on Distributionally Robust Optimization (DRO) and variational analysis. [21] proposed a hyperparameter-free method that replaces the temperature on the CL with an inverse hyperbolic tangent function. In contrast to the above-mentioned methods, our MM-TS method takes a much simpler approach. Instead of altering the objective, architecture, or the optimization process, our method relies on a standard CLIP objective and modifies only the temperature parameter during the training process.

Imbalanced Self-Supervised Learning aims to tackle the challenging problem of learning semantically meaningful representations without labeled data under the constraints imposed by an imbalanced dataset. Previous works addressed this problem through e.g. resampling [10, 14] or reweighing [6, 37]. Alternatively, [4] argued that using the number of samples as a proxy to calculate the entity weight is not ideal and propose a distribution-based analysis which also accounts for relations between samples, whereas recent work on CL provides another approach to dealing with imbalanced data. [23] shows that using a simple temperature scheduler in uni-modal CL can result in a more uniform embedding space and can address the problem of underrepresented classes. On one hand, under a low-temperature regime, the model shows a good instance-discrimination behavior, from which tail classes can benefit. On the other hand, under a high-temperature regime, the model "switches" to cluster creation behavior, from which highly-represented classes can benefit. Moreover, recent studies, i.a. [18, 28, 52], show that compared to fully supervised methods, self-supervised methods can learn more robust feature representations.

3. Method

In this section, we first review the principles of uni-modal contrastive learning. In particular, we discuss InfoNCE and Max-Margin losses and the impact of the temperature parameter (τ), showing its important role in controlling semantic relationships among data-points. Then, we revisit mutli-modal contrastive loss. Finally, inspired by uni-modal finding on contrastive learning, we propose our method that refines the structure of the multi-modal embeddings and improve representations for long-tailed data.

3.1. Background

Contrastive learning (CL) is one of the most important frameworks to learn expressive representations that are used in various downstream tasks [2, 3]. Given a set of inputs $\{x_1, ..., x_N\}$ and visual encoder f_v, the similarity score between the two inputs can be computed as:

$$s_{ij} = f_v(x_i) f_v(x_j)^T. \tag{1}$$

InfoNCE Loss [36] is a commonly used objective, formulated as:

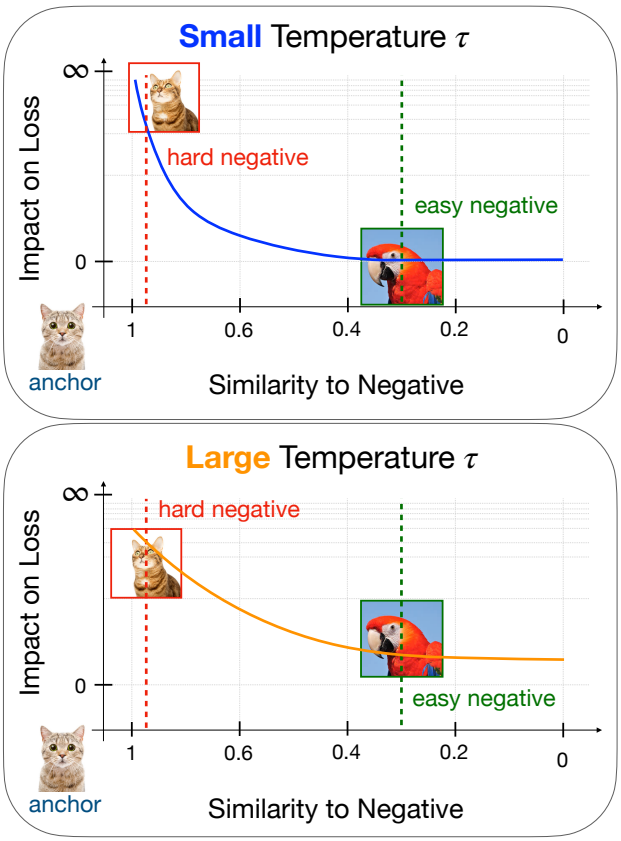


Figure 1. Individual impact of negatives to the loss depends on the temperature parameter. With the small temperature, only hard negatives impact the loss, whereas impact of easy negatives is negligible. With the large temperature impact of easy negatives to the loss increases.

$$\mathcal{L}_{\text{InfoNCE}}(s) = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{N} \exp(s_{ij}/\tau)}, \tag{2}$$

where s_{ii} represents cosine similarity between representations of input x_i and its random augmentation $\hat{x}_i$ (*e.g.* different crops of the same image), and s_{ij} is similarity between representations of x_i and x_j inputs. Further, we refer to x_i as an anchor, $\hat{x}_i$ as a positive, and x_j for all $j \neq i$ as negatives. τ denotes the temperature parameter. Such formulation of the loss encourages pulling together the positive samples, whereas negative samples are pushed away.

Max-Margin Loss [45] is one of the most popular contrastive loss alternatives to InfoNCE loss. The objective of the max-margin loss is to maximize the distance between positive and negative pairs and is formulated as:

$$\mathcal{L}_{\text{max-margin}} = \max(0,\ s_{ij} - s_{ii} + m) \tag{3}$$

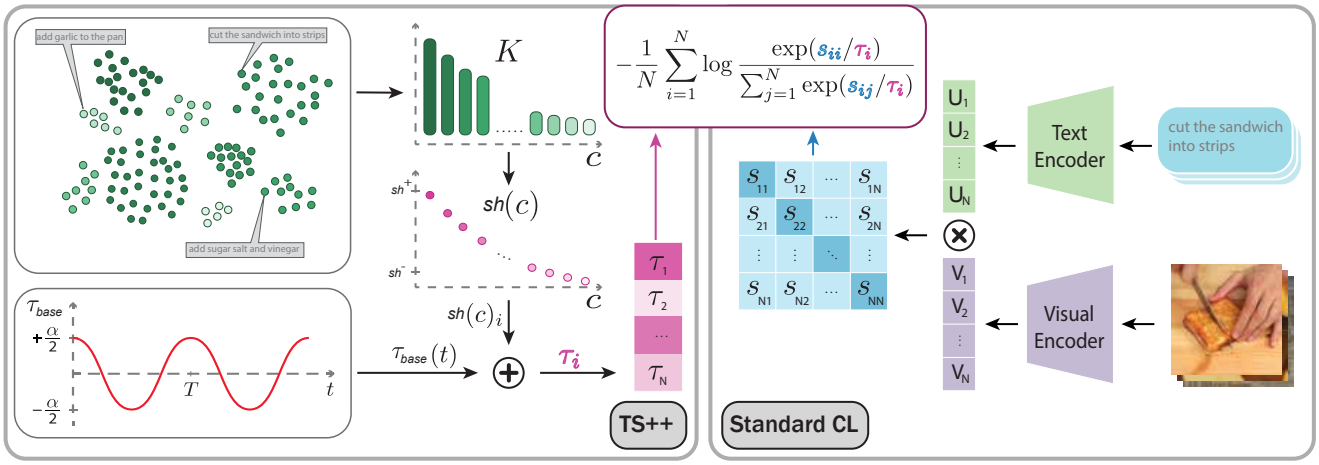


Figure 2. Visualization of our MM-TS approach for computing the InfoNCE loss. First, we cluster text annotations to estimate the distribution of the training data. For each cluster, we assign the respective cluster-based shift ($sh(c)$), where to the larger cluster we assign a larger shift. The base temperature τ_{base} follows the cosine schedule. Next, we adjust the temperature for each individual sample based on the estimate cluster-based shifts resulting in τ_i. The updated individual temperature τ_i is used in a standard InfoNCE loss.

with m being the margin which defines the minimum distance required between the positive and negative pairs.

Influence of the temperature parameter τ As shown i.a. in [23, 47], the influence of τ can be important for the learning of representations. Specifically in [23], the authors discuss impact of the temperature parameter on the learning of the semantic structure in the representation space. A smaller τ amplifies the impact of the nearest negatives (hard negatives) while the impact of more distant (easy) negatives is negligible, see Fig. 1 top. Therefore, only the closest, hardest negatives contribute to the loss. Compared to that, with a larger τ individual contributions of each negatives are on the same order of magnitude, see Fig. 1 bottom. Therefore, a wider range of neighbors is included in the loss, and the model can significantly decrease the loss by maximizing distances to many relatively 'easy negatives', negatives that are easily distinguishable from the anchor, *e.g.* cat vs. parrot. By changing the temperature, the contribution of different negatives to the loss also changes, which leads to different semantic information being learned in the representation space. With small τ, each individual sample maximizes the distance to its closest neighbor that results in a uniform space where the model learns instance specific features and each sample can be clearly distinguised from all others. With large τ, only 'easy negatives' are repelled from each other and the model becomes agnostic to instance-specific features, which allows the creation of semantic clusters, such that a 'cat' cluster will be repelled from a 'parrot' cluster.

We follow [23] and refer to learning with large τ as group-wise discrimination and learning with small τ as instance discrimination. Notably, allowing different types of

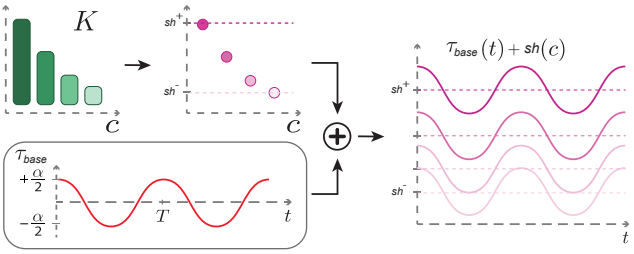


Figure 3. Detailed visualization of temperature calculation for every cluster c on every training iteration t based on the cluster distribution. Given the cosine schedule for the base temperature τ_{base} amplitude α, oscillation period T and cluster-based shifts $sh(c)$, we calculate temperature for each cluster.

discrimination is especially helpful in the case of long-tail distributions. In particular, tail classes need instance discrimination and benefit from local features, whereas head classes need group-wise discrimination and benefit from learned clusters.

Multi-Modal CL includes at least two modalities, *e.g.* images and their captions, where the objective is to increase similarity between the corresponding image-text pairs while repelling from all other instances of both modalities. In particular, there are two aligned sets of inputs, visual inputs $\{v_1, ..., v_N\}$, *e.g.* images or videos, and the corresponding textual descriptions $\{t_1, ..., t_N\}$. Standard multi-modal contrastive learning includes two terms, from text to visual loss and from visual to text loss. Using visual f_v and text f_t encoders the multi-modal similarity scores are:

$$s_{v \to t} = f_v(v) f_t(t)^T, \quad s_{t \to v} = f_t(t) f_v(v)^T. \tag{4}$$

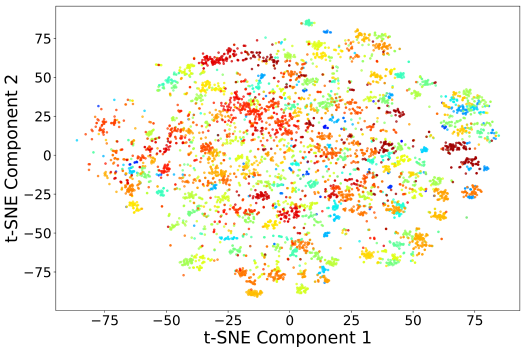


Figure 4. Visualization of the annotation embeddings in the YouCook2 dataset using tSNE. Each point represents a video annotation, and colors indicate the assigned clusters. Number of clusters is 200.

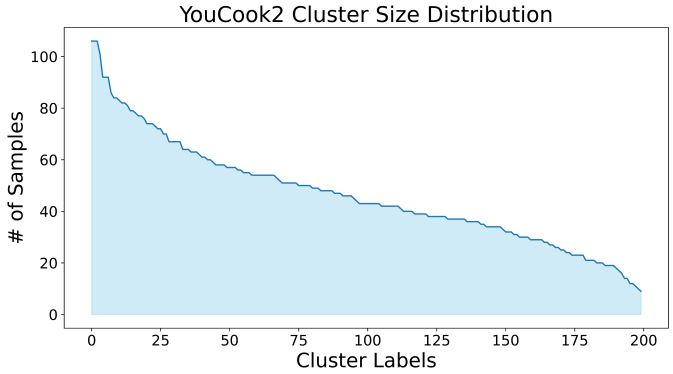


Figure 5. Visualization of long-tail annotations distribution of YouCook2 dataset. Annotations distribution is calculated based on k-mean clustering (200 clusters) of the annotation embeddings. Annotation embeddings are generated using SentenceBERT model [43].

Multi-modal contrastive loss then is an average of the two InfoNCE losses (from vision to text and from text to vision):

$$\frac{1}{2}\left(\mathcal{L}_{\text{InfoNCE}}(s_{v\to t}) + \mathcal{L}_{\text{InfoNCE}}(s_{t\to v})\right). \tag{5}$$

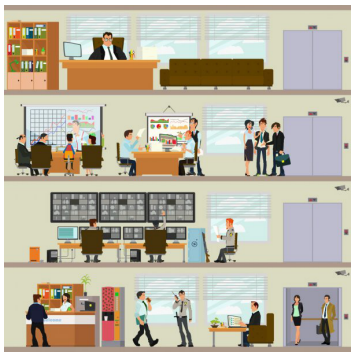
The key distinction between uni-modal and multi-modal contrastive learning lies in how samples interact within the embedding space. In uni-modal contrastive learning, each sample is compared against all other samples within the same modality. In contrast, multi-modal contrastive learning typically involves aligning samples across different modalities, where each instance is contrasted against all available samples from the other modality. Despite these fundamental differences, we demonstrate that insights from uni-modal contrastive learning can be effectively leveraged to enhance multi-modal contrastive learning. Building on this, we propose a novel multi-modal contrastive learning framework that integrates key findings from uni-modal contrastive analysis.

3.2. Multi-Modal Temperature Schedules

Following our observations of the influence of τ on different modalities and findings in [23], we introduce Mutli-Modal Temperature Schedules, MM-TS, for long-tail multi-modal data distributions. MM-TS combines two components: the first is the unsupervised temperature modulation, where we change temperature τ according to a cosine schedule similar to [23], and the second is the individual temperature regulation. To estimate the ranges of τ for individual data points, we propose to leverage the aligned text and vision modalities in order to estimate the distribution of the training data. In particular, we propose to leverage pretrained language models to approximate the distribution of the visual data using the corresponding text information. We show the pipeline to estimate the temperature in Fig. 2. We then use this approximation to estimate individual temperatures to enhance representation learning via a contrastive objective, ultimately improving model performance.

Individual temperature regulation. In [23], the authors discuss a toy experiment on long-tail unimodal data: overrepresented classes are assigned a large τ to encourage group-wise discrimination and learning of the larger clusters, whereas for underrepresented classes a small τ is assigned. The small τ results in instance discrimination and enforces underrepresented samples to push away from similar overrepresented classes improving the learned representations especially for these underrepresented classes. However, for unimodal settings with imbalanced data estimating the data distribution is a challenging. To overcome this problem, we leverage the multi-modal nature of the setting, where the text and visual data are pre-aligned, and extend the toy example from [23] to the real data.

Since the model is trained with contrastive learning to align text and visual modalities, the distribution of one modality can be used to approximate that of the other. In this work, we choose to approximate the distribution using text data, as text embeddings offer more robust semantic representations and are less affected by visual noise. Results of experiments using the visual modality as the distribution source are provided in the Supplementary Materials. We extract all N text embeddings of the corresponding text descriptions with the Bert [43] model and cluster these embeddings into K clusters using K-Means. The size of the resulting clusters is used as a quantized frequency estimate for each corresponding data point. Larger sizes of the text clusters correspond to larger values of τ used during training to induce larger semantic clusters in the multi-modal representation space. Whereas smaller text clusters indicate that some data is semantically underrepresented and could benefit from the explicit separation from other data points, therefore we assign a smaller τ. More specifically, for a cluster c with K_c samples in the corresponding cluster, the



(a) Scenes of people working in the office.



(b) A calm morning before sunrise.



(c) Sketches of cupcakes drawn on a blackboard.

Figure 6. Qualitative samples from different estimated clusters for CC3M. The clusters are estimated using K-Means clustering. The corresponding cluster sizes from left to right are: 69669, 13708, 2904. With yellow we highlight manually identified topic for the corresponding cluster. For more text cluster examples see supmat.

individual cluster shift of the respective temperature is calculated as:

$$sh(c) = \left(\frac{K_c - \min(K)}{\max(K) - \min(K)}\right) \cdot (sh^+ - sh^-) + sh^-, \quad (6)$$

where sh^+ and sh^- define shifts for the largest and smallest clusters, respectively. Clustering and per-sample temperature adjustments introduce negligible overhead, as clustering occurs only once before training. Inspired by temperature schedules [23], we employ a similar intuition and define a dynamically changing schedule for the temperature with respect to the current iteration t as follows:

$$\tau_{base}(t) = \frac{\alpha(1 + \cos(2\pi t/T)) - \alpha}{2}, \quad (7)$$

where α defines the range of changes of the base temperature. And the final temperature for every sample i in the batch is calculated as a sum of the base temperature and the corresponding cluster shift $sh(c_i)$ for the current training iteration:

$$\tau_i = \tau_{base}(t) + sh(c_i), \quad (8)$$

where c_i is cluster of the corresponding sample i, see visualization in Fig. 3. Finally, we employ τ_i in the multi-modal contrastive loss Eq. (9).

Max-Margin Loss. We propose to extend our MM-TS idea to the max-margin loss. In particular, we modify the margin m in Eq. (3) and substitute it with the modulated temperature defined as in Eq. (8). Therefore, we adjust margin in the loss for each sample individually and change it dynamicly over the training. We note that both contrastive losses operate using similar intuition behind them: pulling together similar samples while repelling dissimilar ones. Changes of the margin implies that with a small margin only the strongest negatives are pushed away, whereas with large margin those "hard" negatives do not impact the loss. However, in the max-margin loss the margin does not have influence on the 'strength' of the positive relations in contrast to InfoNCE. Therefore, changing the margin affects only which negatives are pushed away. As we observe similar behavior of both losses, that further confirms that by controlling hardness of the negative samples it is possible to change the semantic of the embedding space.

4. Experiments

In this section, we begin with a discussion of the datasets, the evaluation and the implementation details for the MM-TS method. Then, we discuss the distribution of each training dataset. Lastly, we show that MM-TS improves the performance of various contrastive methods and ablate its components.

4.1. Datasets

CC3M (Conceptual Caption 3M)[46] is a large-scale, internet-crawled image-text dataset. We use it for pre-training. The training split consists of approximately 3.3M image-text pairs representing various styles.

MS COCO (Microsoft Common Objects in Context) is a large-scale dataset used for various tasks, including object detection and image captioning. It consists of 2.5M labels across 328K images, spanning 91 classes. Inspired by previous works [11, 54], we use the Karpathy split [19] for our evaluation. It consists of 5K images for validation and another 5K for test. Retrieval at ranks 1, 5, and 10 is used for evaluation [27].

Flickr30k includes 158K captions across 30K images in different environments. Similar to MS COCO, we use the Karpathy split with 1K images for validation and another 1K for testing. Retrieval at ranks 1, 5, and 10 is used for evaluation [40].

Method	Back-bone	Flickr30K			MSCOCO		
		IR@1	TR@1	Avg.	IR@1	TR@1	Avg.
CLIP [42]	RN50	40.9	50.9	45.9	**21.3**	26.9	24.1
CLIP + Ours	RN50	**41.5**	**54.3**	**47.9**	21.2	**28.4**	**24.8**

Table 1. Performance comparison of pretraining on CC3M and zero-shot evaluation on MSCOCO and Flickr30K. We combine CLIP with our method.

EPIC-KITCHENS-100 (EK-100) is a large-scale video dataset, containing egocentric videos with kitchen-based activities. In our work, we focus on the Multi-Instance Retrieval (MIR) task, which is a video-text retrieval extended with sematic relevancy between narrations. The dataset contains 97 verb and 300 noun classes. The train split contains 67.2K annotated segments, validation split - 9.6K. [8], **YouCook2** is a large-scale video dataset focusing on cooking activities. The dataset contains mainly the 3rd person view videos. Training split contains 10K annotated video segments, validation split contains 3.5K annotated segments. [60]

4.2. Evaluation

Our work addresses two closely related tasks: Image/Video Retrieval and MIR. Multi-Instance Retrieval further necessitates a relevancy matrix to capture the relationships between samples in the dataset. Consequently, we apply two distinct sets of evaluation metrics: Mean Average Precision (mAP) and normalized Discounted Cumulative Gain (nDCG) for MIR, and Recall@1,5,10 for Text-Image/Video Retrieval. Evaluations are conducted bidirectionally, assessing both text-to-image/video and image/video-to-text retrieval performance. For EK-100 and YouCook2, we fine-tune and evaluate on the respective datasets. Whereas, for CC3M experiments, we first fine-tune on CC3M and then perform zero-shot retrieval on MS COCO and Flick30K.

4.3. Implementations Details

For experiments on CC3M, we use standard CLIP [42] framework and conduct experiments with ResNet50 architecture. We initialize it with ImageNet-1K pretrained model. For the CLIP baseline, we use fixed temperature 0.01 as it performed the best.

We conduct all experiments on the EK-100 dataset following the AVION [58] protocol for fine-tuning on the EK-100 MIR task.

For the YouCook2 dataset, we use the fine-tuning configuration of the VAST [1] model for the retrieval task. For the calculation of the cluster distribution we first calculate embeddings of the text annotations using SentenceBert model [43] and then find the clusters with K-Means ($k = 200$).

Regarding hyperparameters (α, sh^-, sh^+), our experiments demonstrate optimal results when parameters oscillate around default temperature or margin values. Specifically, we chose α to keep temperatures positive ($\tau_{\text{default}} - \alpha/2 > 0$), and set sh^- and sh^+ within $[\tau_{\text{default}} - \alpha/2, \tau_{\text{default}} + \alpha/2]$, preventing negative values for tail samples. The values can be found in Sec. 7 (supmat).

Method	Back-bone	mAP			nDCG		
		V→T	T→V	Avg.	V→T	T→V	Avg.
MME [50]	TBN	43.0	34.0	38.5	50.1	46.9	48.5
JPoSE [50]	TBN	49.9	38.1	44.1	55.5	51.6	53.5
EgoVLP [26]	TSF-B	49.9	40.5	45.0	60.9	57.9	59.4
LaViLa [59]	TSF-B	55.2	45.7	50.5	66.5	63.4	65.0
AVION [58]	ViT-B	55.7	48.2	52.0	67.8	65.3	66.5
AVION + Ours	ViT-B	**58.8**	**48.9**	**53.9**	**68.9**	**65.8**	**67.3**

Table 2. Performance comparison of MIR on EPIC-KITCHENS-100 dataset. We combine AVION with our method.

4.4. Estimating the Distribution

The annotations of the EK-100 dataset include extracted verbs and nouns, the dataset naturally follows a long-tail distribution. We visualize verb, noun, and verb-noun combination distributions in Fig. 9 and use these three variants as class distribution sources in our experiments.

For the YouCook2 and CC3M datasets, we derive the distribution of the training data from the K-Means cluster distribution of annotation embeddings. For this purpose, we utilize the Sentence-BERT [43] and distilled version of BERT [44], respectively. With K-Means we estimate 200 distinct clusters, we visualize the estimated distributions in Figs. 5 and 8. In Fig. 6, we show random qualitative samples from clusters of different sizes from CC3M dataset. As these datasets are collected without following any specified class-distribution, we observe that the formed clusters are indeed follow long-tail distribution. In Figs. 4 and 7, we plot tSNE visualization of the clusters for the YouCook2 and CC3M datasets. We observe that most of the clusters has easily identifiable topic, such as *hummus* or *grill*, see Tabs. 10 to 12 (supmat).

4.5. Results

We compare the MM-TS method with current state-of-the-art methods. First, in Tab. 1, we compare our MM-TS method with a standard CLIP objective. We observe that our dynamic individualized temperatures are especially helpful for text-to-image retrieval improving by 3.4% on zero-shot evaluation on Flickr30K and 1.5% on MSCOCO. Second, in Tab. 2, we evaluate the performance on EPIC-KITCHENS-100 dataset using max-margin loss, as it was shown [26, 59] that max-margin loss performs significantly better than InfoNCE when the model is fine-tuned on EK-100. We extend AVION framework based on max-margin loss with the MM-TS framework for dynamic margin estimation and observe improvments across all metrics achiev-

Method	Text-to-Video		
	R@1	R@5	R@10
UniVL [32]	28.9	57.6	70.0
MELTR [22]	33.7	63.1	74.8
VLM [51]	27.1	56.9	69.4
VAST [1]	50.4	74.3	80.8
VAST + Ours	**53.0**	**77.1**	**84.5**

Table 3. Performance comparison on the YouCook2 dataset. We combine VAST with our method. The evaluation is done with Text-to-Video Recall@1,5,10. All benchmarks are multi-modal, containing both audio and subtitles. Methods that also incorporate audio and subtitle modalities (in addition to vision for video representation) are highlighted with a gray background.

ing new state-of-the-art on EK-100. Notably, mAP $V \rightarrow T$ improves more than by 3%. Third, in Tab. 3, we present a comparison of text-to-video retrieval performance on the YouCook2 dataset based on the VAST [1] method. Our method achieves significant performance improvements, surpassing the original VAST results by 2.2-4% and achieving new sota on YouCook2 dataset retrieval task as well.

4.6. Ablations

Impact of MM-TS. In Tab. 4, we analyze the impact of the two main components of our method: temperature (margin) schedules (TS) and individual cluster shifts (ICS), on EK-100 based on MIR task. In particular, TS is the dynamic change of the temperature that follows cosine schedule, whereas ICS are shifts that we employ for the base temperature to adjust for each individual sample. In Tab. 4, we evaluate both, the modified max-margin (MI-MM) and the InfoNce loss (CLIP). We observe that TS is especially effective on the CLIP loss as it improves by 4.3% and 4.0% on avg. mAP and avg. nDCG, correspondingly, whereas ICS is beneficial for MI-MM (max-margin) loss improving by 0.9% on avg. nDCG. However, the combination of the two components brings the best improvements on average for both losses.

Distribution estimation. For EK-100 class annotations such as nouns, verbs and actions are already available (see the corresponding distributions in Fig. 9 (supmat)), with an action being a combination of a respective noun and verb. In Tab. 5, we evaluate influence of the different distributions on the individual cluster shift estimation. We observe that distribution estimated based on nouns results in the best performing model improving both mAP and nDCG results. Interestingly, we manually check formed clusters for both CC3M and YouCook2 (see in Tabs. 10, 11, 12 supmat) and notice that the clusters' topics are biased towards objects and do not reflect distribution of actions (for YouCook2).

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/ TS	58.4	48.6	53.6	**69.4**	64.8	66.9
w/ ICS	56.4	48.4	52.4	68.2	**66.0**	67.1
w/ TS&ICS	**58.8**	**48.9**	**53.9**	68.9	65.8	**67.3**
CLIP	47.0	39.8	43.1	64.4	61.2	62.8
w/ TS	**55.0**	41.7	48.4	**69.1**	64.6	**66.8**
w/ ICS	52.2	43.6	47.8	68.0	**64.8**	66.4
w/ TS&ICS	54.6	**43.9**	**49.2**	68.6	64.4	66.4

Table 4. Performance comparison of AVION on EK-100 dataset (MIR) with modified max-margin- (MI-MM) and CLIP- (InfoNCE) losses. TS and ICS temperature (or margin) schedule and individual cluster shifts, respectively.

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. Performance comparison of AVION with margin distribution method based on the distribution of uniq verbs+nouns combinations, verbs and nouns

5. Conclusion

In this work, we proposed Multi-Modal Temperature and Margin Schedules (MM-TS), a framework designed to enhance multi-modal contrastive learning on long-tail data. Inspired by uni-modal temperature schedules [23], we extended this concept to the multi-modal setting. First, we introduced a dynamic temperature schedule that follows a cosine schedule for both modalities. This approach enables the model to capture different types of semantic information: higher temperatures facilitate the formation of semantic groups in the embedding space, while lower temperatures enhance instance discrimination, helping to distinguish similar samples based on local semantics. To further amplify this effect, we proposed incorporating distribution-dependent shift values. Specifically, we estimated the data distribution either using available class labels or by applying K-Means clustering to pre-extracted text embeddings. Notably, we observed that the resulting clusters encode object-centric semantic structures for both image- and video-language datasets, improving the alignment between modalities. We evaluated our approach on image- and video-language datasets using both InfoNCE and Max-Margin losses. Our results demonstrated consistent improvements across all datasets, evaluation metrics, and losses, highlighting the effectiveness of our method in multi-modal contrastive learning on long-tail data.

References

[1] 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, 8, 1, 2

[2] 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. 1, 2, 3

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1, 2, 3

[4] Xiaohua Chen, Yucan Zhou, Dayan Wu, Chule Yang, Bo Li, Qinghua Hu, and Weiping Wang. Area: adaptive reweighting via effective area for long-tailed classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19277–19287, 2023. 3

[5] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2023. 1

[6] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019. 3

[7] Yufeng Cui, Lichen Zhao, Feng Liang, Yangguang Li, and Jing Shao. Democratizing contrastive language-image pretraining: A clip benchmark of data, model, and supervision. *arXiv preprint arXiv:2203.05796*, 2022. 2

[8] 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. 2, 7

[9] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin F. Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Patrick Collier, Alexey Gritsenko, Vighnesh Birodkar, Cristina Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetić, Dustin Tran, Thomas Kipf, Mario Lučić, Xiaohua Zhai, Daniel Keysers, Jeremiah Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters, 2023. 1

[10] Chris Drummond, Robert C Holte, et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, 2003. 3

[11] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:6704–6719, 2022. 2, 6

[12] 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. 1

[13] Qingpei Guo, Furong Xu, Hanxiao Zhang, Wang Ren, Ziping Ma, Lin Ju, Jian Wang, Jingdong Chen, and Ming Yang. M2-encoder: Advancing bilingual image-text understanding by large-scale efficient pretraining, 2024. 1

[14] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005. 3

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. arxiv e-prints, art. *arXiv preprint arXiv:1911.05722*, 6(7), 2019. 1

[16] 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. 1, 2

[17] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In *Advances in Neural Information Processing Systems*, pages 21798–21809. Curran Associates, Inc., 2020. 2

[18] 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. 3

[19] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 6

[20] Bulat Khaertdinov, Stylianos Asteriadis, and Esam Ghaleb. Dynamic temperature scaling in contrastive self-supervised learning for sensor-based human activity recognition. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(4):498–507, 2022. Publisher Copyright: © 2019 IEEE. 2

[21] Bum Jun Kim and Sang Woo Kim. Temperature-free loss function for contrastive learning. *arXiv preprint arXiv:2501.17683*, 2025. 1, 2, 3

[22] 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. 8

[23] 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. 1, 2, 3, 4, 5, 6, 8

[24] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 2

[25] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1504–1512, 2023. 2

[26] 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. 7

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 6

[28] 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. 3

[29] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 1

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1

[31] Ziqian Lu, Fengli Shen, Mushui Liu, Yunlong Yu, and Xi Li. Improving zero-shot generalization for clip with variational adapter. In *European Conference on Computer Vision*, pages 328–344. Springer, 2024. 2

[32] 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. 8

[33] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *ACL 2024*, 2024. 1

[34] Norman Mu, Alexander Kirillov, David A Wagner, and Saining Xie. Slip: self-supervision meets language-image pre-training. corr abs/2112.12750 (2021). *arXiv preprint arXiv:2112.12750*, 2021. 2

[35] Muhammad Ferjad Naeem, Yongqin Xian, Xiaohua Zhai, Lukas Hoyer, Luc Van Gool, and Federico Tombari. Silc: Improving vision language pretraining with self-distillation. In *European Conference on Computer Vision*, 2024. 2

[36] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[37] Seulki Park, Jongin Lim, Younghan Jeon, and Jin Young Choi. Influence-balanced loss for imbalanced visual classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 735–744, 2021. 3

[38] 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. 1

[39] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, Mingxing Tan, and Quoc V. Le. Combined scaling for zero-shot transfer learning, 2023. 1

[40] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 6

[41] Zi-Hao Qiu, Siqi Guo, Mao Xu, Tuo Zhao, Lijun Zhang, and Tianbao Yang. To cool or not to cool? temperature network meets large foundation models via dro. *arXiv preprint arXiv:2404.04575*, 2024. 2

[42] 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. 1, 2, 7

[43] 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, 2019. 5, 7

[44] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 7, 3

[45] 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. 3

[46] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018. 6

[47] 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. 1, 2, 4

[48] Pei-Hsin Wang, Sheng-Iou Hsieh, Shih-Chieh Chang, Yu-Ting Chen, Jia-Yu Pan, Wei Wei, and Da-Chang Juan. Contextual temperature for language modeling. *arXiv preprint arXiv:2012.13575*, 2020. 2

[49] Shijie Wang, Qi Zhao, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Vamos: Versatile action models for video understanding. In *ECCV*, 2024. 1

[50] 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 in-*

ternational conference on computer vision, pages 450–459, 2019. 7

[51] 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 *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4227–4239, Online, 2021. Association for Computational Linguistics. 8

[52] 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. 3

[53] Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. arxiv e-prints, art. *arXiv preprint arXiv:2110.06848*, 2021. 2

[54] Zhuoning Yuan, Yuexin Wu, Zi-Hao Qiu, Xianzhi Du, Lijun Zhang, Denny Zhou, and Tianbao Yang. Provable stochastic optimization for global contrastive learning: Small batch does not harm performance. In *International Conference on Machine Learning*, pages 25760–25782. PMLR, 2022. 2, 6

[55] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 2

[56] Chaoning Zhang, Kang Zhang, Trung X Pham, Axi Niu, Zhinan Qiao, Chang D Yoo, and In So Kweon. Dual temperature helps contrastive learning without many negative samples: Towards understanding and simplifying moco. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14441–14450, 2022. 1, 2

[57] Oliver Zhang, Mike Wu, Jasmine Bayrooti, and Noah Goodman. Temperature as uncertainty in contrastive learning. *arXiv preprint arXiv:2110.04403*, 2021. 2

[58] Yue Zhao and Philipp Krähenbühl. Training a large video model on a single machine in a day. *arXiv preprint arXiv:2309.16669*, 2023. 1, 7

[59] 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. 1, 7

[60] 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*, 2018. 2, 7