Exploring the Impact of Learnable Softmax Temperature in Contrastive Visual-Textual Alignment Systems: Benefits, Drawbacks, and Alternative Approaches

Anonymous authors
Paper under double-blind review

Abstract

This work does NOT propose *something* that performs better than sota results. Instead, we conduct an empirical analysis of the learnable softmax temperature parameter. This parameter is imperative for optimal system performance, in the practical training of a contrastive visual-textual alignment learning model (commonly referred to as the "CLIP" model). This study addresses three key aspects: 1. The necessity of this temperature parameter; 2. Exploring alternative solutions; 3. Proposing a solution by leveraging the structure of ViTs. Our argument centers around the pivotal role of the softmax temperature in handling noisy training data. This parameter serves as a scaling factor that amplifies the distance range (e.g., [-1, 1] for cosine similarity). However, a large softmax temperature would also result in possible unstable learning dynamics. Subsequently, we re-examine the properties of contrastive learning to figure out alternative approaches to mitigate the problem. Finally, we capitalize on multiple class tokens embedded within the transformer architecture to offer a concise solution. This configuration significantly boosts zero-shot classification performance, enhancing baseline CLIP models pretrained on large-scale datasets by an average of 6.1%. The codes and learned weights are provided in `https://github.com/{Anonymous_authors}`.

1 Introduction

Learning visual and textual feature representations that are semantically aligned in their embedding space is an ordinary problem in the vision-language cross-modal tasks Frome et al. (2013); Karpathy & Fei-Fei (2015); Romera-Paredes & Torr (2015); Wang et al. (2016); Faghri et al. (2017); Xian et al. (2016). In early works that employ feature representations from deep neural networks, *e.g.* Frome et al. (2013), the alignment is often achieved by a fundamental metric learning approach with the hinge rank loss. That is, the similarity between a visual feature vector $\boldsymbol{u}$ and a textual feature vector $\boldsymbol{v}$ is calculated as $\boldsymbol{u}^T W \boldsymbol{v}$, where W are the learnable weight parameters. Thanks to the revolutionary advances in computational power, we can now achieve this in a more effective and practical approach termed contrastive learning, where we align quantities of positive samples and push their negative samples away simultaneously in a large mini-batch using the InfoNCE loss Radford et al. (2021); Singh et al. (2022); Jia et al. (2021); Pham et al. (2021); Yuan et al. (2021).

Given a set of semantically related image-text pairs $\mathcal{S} = \{(\boldsymbol{U}_1, \boldsymbol{V}_1), (\boldsymbol{U}_2, \boldsymbol{V}_2), \ldots, (\boldsymbol{U}_K, \boldsymbol{V}_K)\}$, where $\boldsymbol{U}$ is an image, $\boldsymbol{V}$ is a tokenized text. The goal is to learn a pair of encoders, simultaneously: $f : \boldsymbol{U} \to \boldsymbol{u}, g : \boldsymbol{V} \to \boldsymbol{v}$ to map the image and text into an embedding space, $\boldsymbol{u}, \boldsymbol{v}$ are called embedding vectors of samples. Following the definition in Oord et al. (2018); Wang & Isola (2020); Chen et al. (2021a); Radford et al. (2021), we formulate the contrastive loss as

$$\mathcal{L}_c(f, g; \tau, \mathcal{S}) := \mathop{\mathbb{E}}_{\substack{\boldsymbol{U}, \boldsymbol{V} \sim \mathcal{S} \\ \boldsymbol{U}_i^- \neq \boldsymbol{U} \\ \boldsymbol{V}_j^- \neq \boldsymbol{V}}} \left[-\log \frac{e^{-\tau d(f(\boldsymbol{U}), g(\boldsymbol{V}))}}{N} \right], \tag{1}$$

where τ is the temperature term, we write it as a multiplier for simplicity. $d(\cdot, \cdot)$ is the distance function between two points, and

$$N = \sum_{j \in [M]} e^{-\tau d(f(\boldsymbol{U}), g(\boldsymbol{V}_j^-))} + \sum_{i \in [M]} e^{-\tau d(f(\boldsymbol{U}_i^-), g(\boldsymbol{V}))},$$

is the negative term, with $M \in \mathbb{Z}^+$ denotes a fixed number of negative samples. Intuitively, optimizing this loss term minimizes the distance between positive image-text pairs and maximizes the distance between negative image-text pairs. It is worth mentioning that, in recent studies Radford et al. (2021); Chen et al. (2021b), the contrastive loss is commonly implemented as the cross-entropy between one-hot labels and the class probability obtained by `softmax` within a mini-batch $\mathcal{S}_M$. We also employ this implementation in this work as shown in Section 3.

The standard choice of the distance measure between an image-text pair for the contrastive learning algorithm is the `Cosine Similarity` (in both uni-modal Chen et al. (2020a); Caron et al. (2020); Chen et al. (2020b) and cross-modal Radford et al. (2021); Jia et al. (2021); Singh et al. (2022) scenarios). Mathematically, the `Cosine Similarity` computes the inner product value between normalized feature representation vectors, resulting in a similarity ranging from -1 to 1. While this mathematical framework provides a solid foundation for feature representation vectors from different sources, it is widely acknowledged that training such a configuration is challenging in the absence of a learnable softmax temperature. This learnable temperature is prepended and continuously updated through gradient descent, along with the training progress in practice Wu et al. (2018); Radford et al. (2021).

In this study, our primary contribution lies in addressing three key aspects of this configuration.

1. The benefits and drawbacks of this learnable temperature parameter. The learnable softmax temperature is the key mechanism for contrastive learning on noisy training data. The temperature is essentially a scaling factor for the distance range (e.g., [-1, 1] for cosine similarity), such that the contrastive loss system could reach a numerical "equilibrium". When the system has its gradients from noisy samples neutralized between attraction (the alignment) and repulsion (the uniformity) (Section 2.1). We reveal that the learned softmax temperature is an indicator of the noise level of the dataset (mini-batch) during the contrastive visual-textual alignment. On the other hand, we also discovered that a biased initialization or a higher learned temperature provides inferior learning dynamics(Section 2.4).

2. Alternative configurations from the topological perspective. Although it seems not difficult to build systems that provide good numerical properties to address the distance range limitations. The complexity arises from the need to fulfill two additional properties inherent to contrastive learning. Specifically, not only a broader distance range for the distance function is required, but also a well-defined uniformity of the embedding space along with a "relaxed" triangular inequality for the distance function (Section 2.2). Given these constraints, we suggest the utilization of a product spherical embedding space with the inner product distance as a viable alternative(Section 2.3).

3. Building a practically working system. We additionally explore the utilization of the vision transformer's structure. Specifically, we introduce multiple class tokens to create a product spherical embedding space, resulting in a substantial performance boost for the system (Section 3). In a comprehensive large-scale experiment, we trained a ViT-B/16-based CLIP model that surpassed the baseline model by an average of 6.1% in zero-shot classification tasks (Section 4).

2 Rethinking the Contrastive Alignment

In this section, we discuss our motivation in detail. We first provide a visualized equilibrium. Then, we discuss the conditions (inherent properties) for the contrastive loss to reach the equilibrium. Next, we design a toy experiment under controlled configurations to demonstrate the effects of the properties. To generalize the problem, we conceptualize the embedding vectors as points within specified typologies. We use the term "distance between the points" instead of similarity. We also say a contrastively "well-optimized" visual-textual model should yield a shorter distance between semantically related image-text pairs than the non-related counterparts, regardless of how the pairs are labeled.

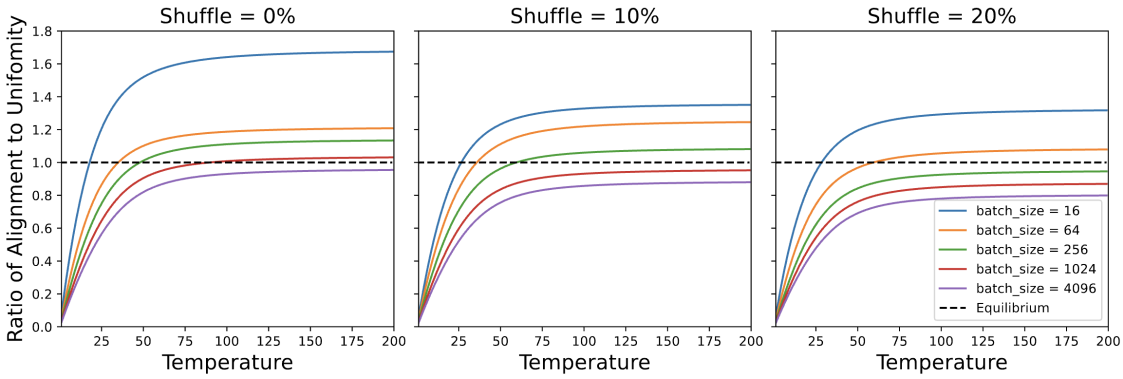


Figure 1: The ratio of alignment loss to uniformity loss under different temperatures. The "equilibrium" is presented as the ratio equals 1.0, when the log loss in Eq.(1) becomes 0.0. Shuffle = P% standards for P percent of the labels are shuffled.

2.1 The visualized equilibrium

In Figure 1, we give an intuitionistic presentation of the equilibrium with our learned reference model in Section 4. As explained by Wang & Isola (2020), contrastive loss is a combination of two objects: a) alignment of features from positive sample pairs and b) the distribution of the features encouraged to match. To demonstrate this, we randomly select a mini-batch of samples from our training dataset of different sizes. Then, We manually add noise to the mini-batch by shuffling a small percentage of the images' labels (the paired texts). Finally, we compute the ratio of alignment loss to uniformity loss (numerator and denominator in Equation (1) under different temperature situations (averaged through all the positive pairs in the mini-batch). It can be seen that a noisier mini-batch or a larger mini-batch size demands a higher temperature to reach the loss equilibrium. Meanwhile, when the system is working under a lower temperature, it tends to push negative pairs away rather than pull positive pairs closer. In the next subsection, we discuss the mechanism behind this observation from three (often) neglected conditions.

2.2 Three conditions to achieve the equilibrium

i. Proper definition of uniformity of the embedding space: Naturally, with the loss form defined in Equation (1), the distribution object will result in a uniform distribution on the sphere. Although the distribution of samples doesn't have to be exactly "uniform" as discovered by Chen et al. (2021a), it is necessary to define a proper prior distribution for samples to match via optimal transport algorithms (*e.g.* sliced Wasserstein distance), which is undoubtedly a computational burden. Consequently, the spherical embedding space is deemed the most suitable topology for contrastive alignment, as it exhibits a proper uniform distribution defined by the surface area measure. In contrast, the commonly adopted unbounded Euclidean space lacks this property.

ii. "Relaxed" triangular inequality: Assume we have a model that is "well-optimized" that is trained using a noisy dataset (even using human-labeled datasets, see Chun et al. (2022)). For this model, we have the following properties: For a positive pair $(\boldsymbol{U}, \boldsymbol{V})^+$, their distance $d(\boldsymbol{u}^*, \boldsymbol{v}^*)$ is upper-bounded by a small ϵ^+, and the distance for negative pairs $(\boldsymbol{U}, \boldsymbol{V})^-$ is lower-bounded by a large ϵ^-. Now, let us consider a set of two pairs of its training samples $\mathcal{S}_\pm = \{(\boldsymbol{U}_1, \boldsymbol{V}_1), (\boldsymbol{U}_2, \boldsymbol{V}_2)\}$. *Accidentally, the pair* $(\boldsymbol{U}_1, \boldsymbol{V}_2)^-$ *is also semantically correlated, despite being recognized as a negative sample* (It is pervasive to have negative pairs of image and text that match each other equally well as the positive ones in the noisy dataset). Consequently, this "well-optimized" model will predict a distance upper-bounded by ϵ^+ for this pair instead of a larger value than ϵ^-. If the distance function d is a *metric*, then according to the triangle inequality axiom of metric, we have the following inequality (see Fig. 2 for an intuitive illustration),

$$\epsilon^- \leq d(\boldsymbol{u}_2^*, \boldsymbol{v}_1^*) \leq d(\boldsymbol{u}_1^*, \boldsymbol{v}_2^*) + d(\boldsymbol{u}_2^*, \boldsymbol{v}_2^*) + d(\boldsymbol{u}_1^*, \boldsymbol{v}_1^*) \leq 3\epsilon^+ \tag{2}$$

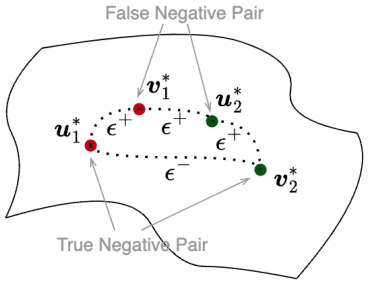


Figure 2: Illustration of the bounded negative distance between true negative pair of samples.

From this simple derivation, in this "well-optimized" model, the negative distance ϵ^- is bounded by ϵ^+, meaning that the system would not learn numerical separated distance ranges for positive and negative pairs to minimize the contrastive loss (See supplementary materials). The solution is to have a "relaxed" triangular inequality to alleviate the ambiguity of the positive/negative pairs. In practice, we have the following observations: i) The positive pairs of samples usually have a much larger distance than that in the perfect alignment condition, ii) $d(\boldsymbol{u}_1^*, \boldsymbol{u}_2^*)$ may become exceptionally small since none of the loss terms regularize it, resulting in a further tightened bound of the negative distance; iii) Although the inner product distance is not a *metric*, it still obeys an "relaxed" triangular inequality because we can yield a metric on the sphere by `ArcCos` (the geodesic), see Schubert (2021) for more information.

`iii.` Broad distance range: We finally discuss the mechanism behind the learnable temperature trick. As discussed in Section 2.1, the contrastive loss is minimized through the gradients of both alignment and uniformity losses. In a noisy dataset inducing "semantic ambiguity", the false negative samples are pushed away from each other (repulsion), while the false positive samples are pulled together (attraction). Ideally, the system could gradually find a numerical "equilibrium" (the stabilized state) when the noisy samples' gradients for attraction and repulsion are equal. For instance, if there is a reasonable amount of false negative samples, the model would learn a smaller distance between the pairs labeled as "negative", such that the uniformity loss won't be too high when encountering false negative samples in mini-batches. On the contrary, the model would learn a larger distance between the pairs labeled as "positive" when the amount of false positive samples is inneglectable. In other words, the model reaches the equilibrium by learning *compromised positive and negative distances* under semantic ambiguity.

However, the required numerical value of the compromised distances is out of the range that the cosine similarity could provide. This is generally because the InfoNCE loss computes the sum of exponential functions. To reach this equilibrium, we need to expand the distance range to $[-\tau, \tau]$, and since we don't know the noisy level of the dataset, we set the temperature learnable through gradients. For instance, the officially released CLIP model Radford et al. (2021) has a glancing similarity of $0.3 \sim 0.5$ and $0.1 \sim 0.3$ for positive and negative pairs of samples, respectively. The learned temperature is approaching 100.0, indicating the value of distances for equilibrium are $30 \sim 50$ and $10 \sim 30$ for positive and negative pairs of samples, respectively.

2.3 The toy experiment

Experiment settings: We design a toy experiment to demonstrate how the properties mentioned above influence the performance of contrastive learning. We employ the 15M subset Cui et al. (2022) of the YFCC100M dataset Thomee et al. (2016) as the training dataset, which contains roughly 15.3 million internet collected weakly related image-text pairs. We evaluate the learned models with the zero-shot retrieval performance on Flickr30K, and Zero-Shot/`Linear Probe` classification performance on ImageNet for reference. We employ the original ViT-S/16 architecture for our image encoders Dosovitskiy et al. (2020), with an input image resolution of 224, resulting in 196 image tokens.

We evaluate four types of configurations for comparison, each featuring a distinct combination of topology and distance function. The configurations are summarized in Table 1. Specifically, we consider: i) the sphere

Topology	Sphere $\mathbb{S}^{d-1}$	Euclidean $\mathbb{R}^d$	PS$(d/m, m)$	PS$(d/m, m)$
Distance	$-\boldsymbol{u}^T\boldsymbol{v}$	$\|\boldsymbol{u}-\boldsymbol{v}\|_2$	Geo$(\boldsymbol{u}, \boldsymbol{v})$	$-\text{tr}(\boldsymbol{u}^T\boldsymbol{v})$
Uniformity	surface measure	undefined	surface measure	surface measure
Inequality	relaxed	restricted	restricted	relaxed
Distance Range	$[-1, 1]$	$[0, +\infty)$	$[0, m\pi]$	$[-m, m]$

Table 1: Summary of different topologies endowed with different distances. The total dimension of the embedding vector is denoted as d. The mini-batch size is denoted as b. Green box stands for the properties that are *favored* for contrastive learning. Red box stands for the properties that are *unfavored* for contrastive learning.

Topology	**Distance**	**Zero-Shot I2T R@1**	**Zero-Shot T2I R@1**	**Zero-Shot Cls. Acc.**	**Linear PrPSe Cls. Acc.**
Temperature: init=1.0, gradient=True					
Sphere	$-\boldsymbol{u}^T\boldsymbol{v}$	49.0	30.33	28.59	59.56
Euc	$\|\boldsymbol{u}-\boldsymbol{v}\|_2$	47.4	30.71	29.85	60.09
PS(64, 8)	Geo$(\boldsymbol{u}, \boldsymbol{v})$	49.9	32.49	30.21	60.61
PS(64, 8)	$-\text{tr}(\boldsymbol{u}^T\boldsymbol{v})$	50.9	32.71	30.50	60.66
Temperature: init=1.0, gradient=False (No temperature)					
Sphere	$-\boldsymbol{u}^T\boldsymbol{v}$	5.1	3.461	4.04	45.37
Euc	$\|\boldsymbol{u}-\boldsymbol{v}\|_2$	47.6	30.43	29.51	59.20
PS(64, 8)	Geo$(\boldsymbol{u}, \boldsymbol{v})$	4.1	2.921	3.10	21.67
PS(64, 8)	$-\text{tr}(\boldsymbol{u}^T\boldsymbol{v})$	30.3	18.48	21.20	57.93

Table 2: The retrieval and classification performance. "gradient={True/False}" donates if the temperature is learnable. **The configurations are chosen intendedly to support the conclusions in Table 1. The results shown in this table do not claim some configurations are better than others.**

$\mathbb{S}^{d-1}$ with the inner product as distance, which is the commonly used `cosine similarity`; ii) the Euclidean space $\mathbb{R}^d$ with ℓ_2 distance; iii) a product spherical embedding space PS(n, m) with the minimizing geodesic as distance, which is denoted as Geo$(\boldsymbol{u}, \boldsymbol{v}) = \text{tr}^{\frac{1}{2}}(\arccos^2(\boldsymbol{u}^T\boldsymbol{v}))$; and iv) The same product spherical embedding space PS(n, m) with the inner product as distance.

Here, we explain how we implement the product sphere embedding space. The product sphere (PS(n, m)) can be defined as $\underbrace{\mathbb{S}^{n-1} \times \cdots \times \mathbb{S}^{n-1}}_{m \text{ copies}}$, where $\mathbb{S}^{n-1}$ is the sphere embedded in $\mathbb{R}^n$. Intuitively, it is a vector composed of m chunks of $n-$dimensional normalized sub-parts. Our implementation follows this intuition; we reshape the original feature vector into a matrix of shape $m \times n$, then ℓ_2-normalize the columns.

We provide further clarification on the distance functions employed in these configurations. As previously discussed, the `cosine similarity` calculates the (negative) inner product for vectors on the unit sphere as their distance. Accordingly, we adopt this design for the product sphere. Hence, the distance could be computed as the negative value of the trace of the matrix product, *i.e.* $d(\boldsymbol{u}, \boldsymbol{v}) = -\text{tr}(\boldsymbol{u}^T\boldsymbol{v})$. This distance is clearly not a restricted metric. Therefore, for reference, we also consider the minimizing geodesic as distance, which is a restricted metric (obeys the triangular inequality).

Hyperparameters and results: For the product spherical embedding space, we employed a structure of $n = 64$, $m = 8$, denoted as PS(64,8). Concerning the temperature parameter, we initialize it with `exp`(0.0) (equivalent to 1.0) and conduct two sets of experiments, one with gradients (learnable) and the other without gradients (not learnable). It's important to note that if the parameter is initialized to 1.0 and is not learnable, it is essentially equivalent to having no such parameter. All the results are presented within Table 2, and

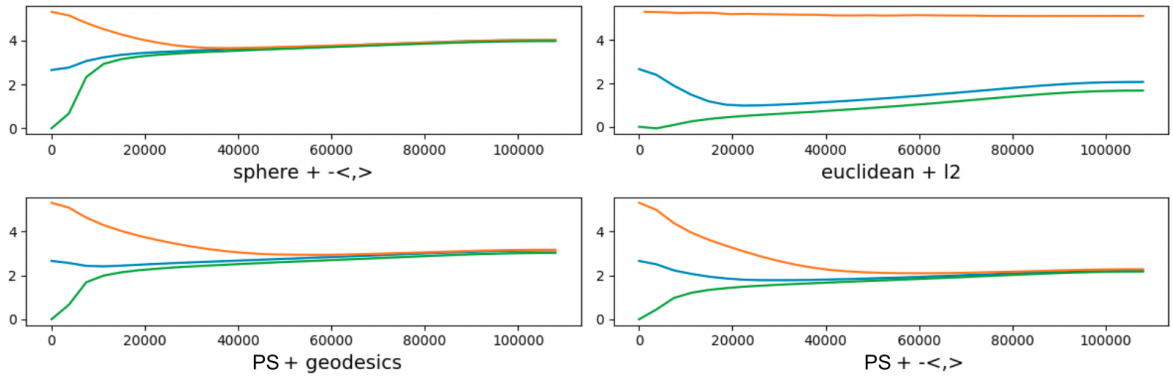


Figure 3: The learning curve of the temperature. -<,>, l2 and geodesics denote the negative inner product, ℓ_2, and geodesic distance, respectively. The orange, blue, and green curves denote the initialization of $\texttt{exp}(5.3)$, $\texttt{exp}(2.64)$, and 1.0, respectively.

subsequent sections will elucidate the reader on how to interpret this tabular data. Results using other initialization (Figure 3) are given in the supplementary materials.

Effects of the uniformity: To examine the effects of uniformity, we compare the performance between the Euclidean with ℓ_2 distance and the product sphere with geodesic distance. These configurations have restricted triangular inequality in common. Meanwhile, if the temperature is learnable, the product sphere also owns a practically unbounded distance range similar to the Euclidean. Then, the only difference is that uniformity can be defined on the product sphere (surface area measure). We observe that the performance of the product sphere with geodesic configuration performs better than the Euclidean with ℓ_2 configuration. This result indicates the importance of properly defined uniformity.

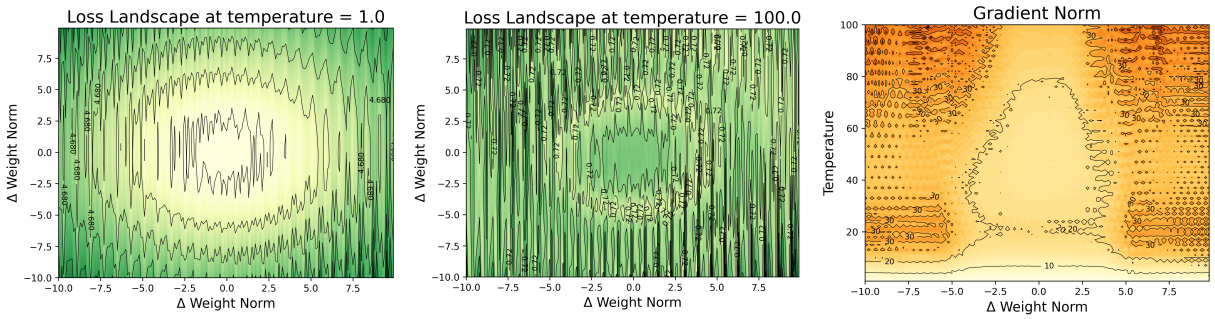
Effects of the Tri-angular Inequality: Next, we examine the effects of the triangular inequality using the product sphere topologies endowed with different distance functions, that is, the geodesic distance and the inner product distance. The results are visually depicted in Table 2, specifically between the 3rd and 4th lines of each data block presented in the table. Upon observation, it becomes evident that the inner product distance outperforms the geodesic distance on average in both retrieval and linear probe tasks. Moreover, when the temperature is unlearnable, the inner product distance still provides the model trainability, showing the advantage of removing the restriction of tri-angular inequality.

Effects of the Distance Range: Finally, we present the effects of the distance range by comparing the proposed product sphere topology with inner product distance and the baseline spherical topology. Notably, when the temperature is learnable, the product sphere demonstrates a reasonable improvement in the top-1 recall and classification accuracy. It is worth mentioning that this implementation does not bring extra computational complexity, and the only difference is the shape of the embedding space. This suggests that a larger distance range helps the alignment of the features from different modalities. There is one more piece of evidence that lies in the last block: the Euclidean with ℓ_2 distance configuration obtains consistent performance regardless of the temperature parameter, as it operates with an unrestricted distance range.

2.4 Drawbacks in the learning dynamics

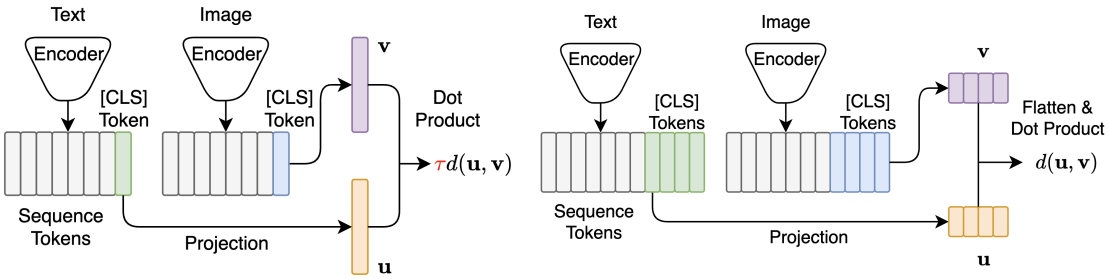
Biased initialization may degrade training: We now examine how the initialization of the temperature impacts the training progress. Specifically, we additionally run experiments with the temperature initialized at $\texttt{exp}(2.64)$, which is the default value in the official CLIP implementation, and a notably higher value of $\texttt{exp}(5.3)$ ($\sim$200). The changes in the temperature during the training process are illustrated in Figure 3[1]. The figures confirm that: The temperature converges to an equilibrium value irrespective of initialization when the uniformity is properly defined. This value reflects the noise level present in the datasets, configurations with

[1]The detailed retrieval and classification results are provided in the supplementary material. The final performance is not largely impacted by the initialization though.



(a) (b) (c)

Figure 4: The impact of temperature on the loss landscape. (a) and (b) depict the contours with two orthogonal weights Δ in random directions at temperatures of 1.0 and 100.0, respectively. (c) presents the contour gradient norm with weight Δ in a random direction across various temperatures (dark color stands for a larger norm).



(a) Vanilla CLIP (b) CLIP with Multiple Class Tokens

Figure 5: A sketch of the contrastive visual-textual alignment (CLIP) system.

"relaxed" triangular inequality and a broader distance range have lower equilibrium values. These findings suggest that, if the training is insufficient or the dataset is too noisy, then with a biased initialization of the temperature parameter, the training progress might be degraded.

Higher temperature hampers the optimization: We next examine the impact of temperature on the loss landscape and the norm of gradient, using the approach proposed in Li et al. (2018). We employ the learned reference model in Section 4 to draw the figures, where the final learned temperature is 100.0 (due to the "clamp" function). We randomly select a mini-batch of size 1024 from the training dataset. The results are shown in Figure 4. From figures (a) and (b) we can observe that the model with a lower temperature has a much smoother loss landscape. Furthermore, in Figure (c) we find that the norm of the gradient might become greater as the temperature grows. Therefore, in practice, we need to employ techniques such as gradient norm clipping to stabilize the training.

3 The Multiple Class Tokens Solution

We first review the importance of the class token. In the design of both the textual transformer (BERT, Kenton & Toutanova (2019)) and the visual transformer (ViT, Dosovitskiy et al. (2020)), *one* learnable embedding is used to represent global information, termed as class token. Different from the sequence (patch) tokens, the class token is a key component of the transformer encoder. It is randomly initialized and updated through gradient descent during the optimization. Furthermore, the class token holds a fixed position embedding, avoiding the influence of the positional information. Therefore, the class token is considered to participate in the computation of global attention.

```
# d      - dimensions of the hidden embedding
# U, V   - mini-batch of images/texts token, [n, p, d] / [n, l, d]
# PS_m   - the dimension of each sub-sphere
# PS_n   - the number of [CLS] tokens attached
# cls_U, cls_V  - class tokens for images/texts, [n, PS_n, d] / [n, PS_n, d]
# t      - learned temperature parameter

# concatenate cls_tokens and extract features
U_, V_ = concatenate([cls_U, U], axis=1), concatenate([cls_V, V], axis=1)
u_bar = visual_transformer(U_) #[n, PS_n + p, d]
v_bar = textual_transformer(V_) #[n, PS_n + l, d]

# map features onto PS(n,m) and calculate distance
u = projection_u(u_bar[:PS_n]).l2_normalize(axis=-1) # [n, PS_n, PS_m]
v = projection_v(v_bar[:PS_n]).l2_normalize(axis=-1) # [n, PS_n, PS_m]
# [n, PS_n, PS_m], [n, PS_n, PS_m] -> [n, n]
neg_distances = einsum('inm,jnm->ij', u, v) * t.exp()

# symmetric loss function
labels = arange(n) # 0, 1, ..., n-1
loss = (CE_loss(neg_distances, labels, axis=0) + CE_loss(neg_distances, labels, axis=1)) / 2
```

Figure 6: Python-like pseudo-code of the proposed approach.

Motivated by the properties of the class token, we propose to employ multiple class tokens to build the product spherical embedding space, with each class token being a sub-sphere of $\mathbb{S}^{n-1}$. For the visual encoder, these class tokens are randomly initialized to break symmetry, while for the textual encoder, we use different absolute positional embeddings for each class token. We present a sketch of the system in Figure 5[2] and a pseudo-code in Section 3, to facilitate a better understanding. Given the fact that the dimension for the embedding space could be a critical factor to the performance of the system Gu et al. (2021), we select the dimension n with a conservative strategy. Specifically, we anchor the dimension of Euclidean spaces to be the same as the reference model, then vary the value of n, m such that $n \times m = l$. We denote this implementation as Multi(n, m). The sub-spheres could benefit from the global attention operation and provide more representative feature embeddings. On the contrary, the multi-token implementation requires more computational resources in the backbone since the class tokens are involved in the computation of global attention.

4 Large-Scale Experimental Results

4.1 Experimental Settings

Please note that our objective is not to produce the best publicly available model; rather, we solely conduct experiments under controlled and restricted conditions. We compare the performance of the proposed method using the configuration, which matches the publicly released ones in teams of dataset samples, model sizes, and training progress. We also re-implement the naive CLIP model as the reference, which holds a similar performance as the publicly released ones.

[2]It is needed to clarify that in the multiple class tokens system, we also employ the learnable temperature practically. Since the finally learned temperature is significantly smaller than that in the vanilla CLIP system (~ 4.0 versus ~ 100.0), we omit it to highlight the difference between the systems.

Method *baseline*[*impl.*]	IN	INV2	IN-A	IN-R	Flickr30K Zero-shot			MSCOCO* Zero-shot		
	ZS cls. Acc@1	**ZS cls.** Acc@1	**ZS cls.** Acc@1	**ZS cls.** Acc@1	**I2T** R@1	**T2I** R@1	**Mean** R@1/5/10	**I2T** R@1	**T2I** R@1	**Mean** R@1/5/10
ViT-B/16-224 as visual bone.										
CLIP[openAI†]	68.7	61.9	50.1	77.7	81.9	62.1	86.1	55.4	38.4	66.3
CLIP[openCLIP‡]	67.0	59.6	33.2	**77.9**	83.2	65.5	87.6	52.4	38.4	62.4
CLIP[our-impl.]	69.5	61.4	49.5	70.6	84.2	61.7	86.4	**64.1**	**43.9**	**72.4**
CLIP[Multi(32,16)]	**76.4**	**68.0**	**55.8**	75.2	**85.2**	**66.3**	**88.3**	63.8	42.9	**72.4**
ViT-L/14-224 as visual bone for reference.										
CLIP[openAI†]	75.5	69.7	70.7	87.9	85.0	65.2	87.7	56.3	36.5	65.2
CLIP[openCLIP‡]	72.7	65.6	46.6	84.8	87.6	70.3	90.1	59.7	43.0	70.0

Table 3: Comparsion of large scale contrastive visual-textual pre-train model on benchmark datasets. † and ‡ denote the implementation from Radford et al. (2021) and Ilharco et al. (2021), respectively. The metric **Mean** stands for the average value of R@1/5/10 of I2T/T2I retrieval performance. * denotes the Karpathy test split Karpathy & Fei-Fei (2015).

Datasets: For the experimental analysis in Section 4.2, we collect data from publicly available datasets Schuhmann et al. (2021); Changpinyo et al. (2021); Sharma et al. (2018); Chen et al. (2015); Krishna et al. (2017); Plummer et al. (2015); Russakovsky et al. (2015); Desai et al. (2021); Kuznetsova et al. (2020); Li et al. (2017)[3], resulting in a total of 420 million individual images and roughly 500 million image-text pairs. This dataset is comparable to the one employed in the official CLIP paper Radford et al. (2021) and another open source re-implementation Ilharco et al. (2021).

Models: Specifically, we employ the ViT-B/16 as our image encoders. For our text encoders, we employ Ernie-2.0-en-base Sun et al. (2020), which is literally a Bert model Devlin et al. (2018) of 12 layers and 512 hidden neuron sizes with a customized vocabulary of 30,522 tokens, and the maximum context length is set to be 77. We project the feature representation (class token) from the top layer of transformers to a (sum of) 512-dimensional embedding space. All the parameters except the temperature are optimized from random initialization. The default initialization of the project matrix employs the Gaussian initializer of zero mean, and standard deviation equal reversed square root of the input size (*a.k.a.* Kaiming initialization). The details of the hyperparameters are provided in supplementary materials.

Evaluation: We first evaluate the proposed methods with two types of vision tasks: i) `Zero-Shot` image-to-text and text-to-image retrieval on Flickr30k Plummer et al. (2015) and MSCOCO Lin et al. (2014) ii) `Zero-Shot` classification on ImageNet-1K Russakovsky et al. (2015), ImageNet-V2 Recht et al. (2019), ImageNet-R Hendrycks et al. (2021a) and ImageNet-A Hendrycks et al. (2021b). For zero-shot retrieval on Flickr30K and MSCOCO, we employ the logits (distance) computed by the distance function and report the image-text pairs with the top-k shortest distance as the retrieval results. For zero-shot classification on ImageNet. We employ multiple prompt templates described in Radford et al. (2021), while we first compute the distances between image and text embeddings, then average the distances. For linear probe classification on ImageNet, we remove the learned projection head (no topological structure is preserved), then attach a random initialized linear projector to map the feature representation to the 1,000 class logits.

Besides the tasks as mentioned above, we provide more results using the ECCV dataset Chun et al. (2022). The dataset is proposed for eliminating the false negative samples in the validation set of the original MSCOCO dataset. Instead of the commonly used Recall@K (R@K) metric, the datasets provide a new ranking-based metric, mAP@R. The authors of the ECCV dataset have shown that the mAP@R metric is more aligned to humans than Recall@k. Therefore, the performance of a model evaluated by mAP@R would be less occasional than the R@1. This metric is deemed more precise for evaluating the performance of models in the presence of noise.

[3]The availability of LAION400M is about 90%.

Method *baseline*[*impl.*]	COCO 1K		COCO 5K		CxC		ECCV Caption					
	I2T	**T2I**	**I2T**	**T2I**	**I2T**	**T2I**	**I2T**			**T2I**		
	R@1	R@1	R@1	R@1	R@1	R@1	mAP@R	R-P	R@1	mAP@R	R-P	R@1
ViT-B/16-224 as visual bone.												
CLIP[openAI†]	71.7	52.5	52.5	33.1	54.0	34.7	23.7	34.0	68.8	34.8	44.0	73.4
CLIP[openCLIP‡]	74.0	57.6	55.4	38.3	57.3	40.0	26.2	36.6	70.3	36.9	46.4	77.5
CLIP[our-impl.]	80.8	63.0	**64.2**	**43.1**	**65.3**	**44.9**	30.5	41.0	**78.6**	40.5	49.9	81.2
CLIP[Multi(32,16)]	**81.1**	**63.1**	63.8	42.9	**65.3**	44.8	**30.9**	**41.7**	76.3	**41.7**	**50.5**	**84.1**
ViT-L/14-224 as visual bone for reference.												
CLIP[openAI†]	74.3	55.4	56.4	36.6	58.0	38.3	24.0	33.8	71.3	32.0	41.8	73.0
CLIP[openCLIP‡]	77.2	61.4	59.7	43.0	61.1	44.8	28.1	38.3	73.0	38.7	47.9	81.2

Table 4: Comparsion of large scale contrastive visual-textual pre-train model on benchmark datasets.

Image-Text Pair | Full Tokens | Token id=01 | Token id=15

Figure 7: Visualization of the importance map using the Grad-CAM algorithm. The columns from left to right stand for: the input image-text pair; the importance map computed based on the final matching score; and the importance maps based on the matching scores of two individual tokens and the involved token IDs. Additational results are provided in the supplementary material (including the failed cases).

4.2 Experimental results

The evaluations of the learned models on the commonly employed image classification and image-text retrieval tasks are reported in Table 3, with the ViT-B/16 as the visual backbone. It can be seen that the re-implemented CLIP holds a similar performance as the publicly released ones in most cases. On the other hand, our proposed model with the multi-token implementation of (32,16) significantly outperforms the other ViT-B/16 models in general, with less than 8% more computational costs. The only exception is the top-1 retrieval performance on the MSCOCO datasets. The reason could be two-fold. Firstly, we observe a mild "semantic decoupling" between the embedding of tokens through the visualization (see Section 4.3); that is, some of the individual class tokens focus on specified objects and provide a high alignment confidence. This may cause confusion in understanding the given scene as a whole; hence, the recall@top-1 performance is degraded. Secondly, the most suitable temperature during training for aligning object-level and scene-level concepts might differ. In our experiment, we decrease the upper limitation of the temperature to 6.25 (100 / 16 [tokens]) since the product spherical topology owns the border distance range. The scene-level concept alignment might require a larger temperature for "ambiguity" to achieve better retrieval performance.

For the ECCV caption dataset, we utilize the officially released evaluation tool and present a summary of the models' performance in Table 4. It is evident that our proposed multi-token product sphere topology outperforms others in terms of mAP@R, signifying the model's enhanced robustness in handling "semantic ambiguity" samples.

4.3 Visualization on Tokens Attention Regions

In this section, we provide a commonly adopted neural network explanation method to visualize the influence of inputs on the final outcome. Specifically, we employ the Grad-CAM Selvaraju et al. (2017) algorithm to highlight the interested parts by the model of both images and their corresponding texts. Notably, the original design of the Grad-CAM algorithm precludes its direct application to textual data. Therefore, we enhance its capabilities such that it also highlights the contributing parts of texts in a token-wise style. We employ examples from the evaluation set of the Flickr30K and MSCOCO datasets for visualization. The results are shown in Figure 7. Through our observations, we have noted that certain pairs of class tokens exhibit independent alignment, thereby reflecting a distinct concept or idea embedded within the image. We attribute the improved classification performance to this phenomenon, as it enables a more effective representation and understanding of the underlying content by leveraging the distinctive alignment of the class token pairs. It is noteworthy that the intrinsic decoupling phenomenon observed within the embeddings of class tokens is *NOT* universally present across all image-text pairs or within every token of an image-text pair. This is due to the inherent challenges faced by visualization algorithms in achieving precise correspondence in the importance maps for intricate semantic representations.

5 Related Works

Momentum distillation: In recent works such as Cheng et al. (2021); Li et al. (2021a), the momentum (self-)distillation is introduced to mitigate the semantic noise in the sample pairs. That is, a momentum version of the model is updated by the moving average of the model's historical parameters. Then, the cross entropy between the softmax logits computed by the model and its momentum version is used as an additional loss for supervision. The authors claim that the pseudo-targets of the momentum (self-)distillation will not penalize the model for matching negative samples that are reasonably similar. Here, we consider that the pseudo-targets do "relax" the triangular inequality restriction implicitly by letting the distance of alignment be reasonably large. Hence, it could be much easier for the optimizer to find the equilibrium discussed in Section 2.

Other implemention of non-metric distance: In Yao et al. (2021), the authors proposed a so-called fine-grained contrastive learning scheme that matches all the visual and textual tokens using a maximum-average operator. Concretely, for each visual token, it finds the textual token with maximum similarity, then takes the average over the visual tokens as the similarity of the image to a text and vice versa. Using our framework, this work can be explained as embedding samples onto the product manifold $\mathbb{S}^{d-1} \times \cdots \times \mathbb{S}^{d-1}$ endowed with the maximum-average distance, which is a non-metric distance. At the same time, the authors employ the sub-manifold $\mathbb{S}^{d-1}$ to represent local information.

The effects of softmax temperature: In Wang & Liu (2021), the authors draw the uniformity of the embedding distribution and the tolerance to semantically similar samples of learned models under different temperatures. From the observations, the authors claim that "a good choice of temperature can compromise these two properties properly to both learn separable features and tolerant to semantically similar samples, improving the feature qualities and the downstream performances". Unlike our work, this work is done under uni-modal contrastive learning, where the semantic correlation of the negative samples is not a property of the datasets but rather a drawback of the larger mini-batch size.

Uni-modal side tasks: In works such as Mu et al. (2021); Li et al. (2021b); Yang et al. (2022), authors combine cross-modal contrastive loss with other uni-modal tasks, for instance, visual/textual self-supervised contrastive learning, masked image/language modeling. These combined methods empirically demonstrate superior performance in downstream tasks such as zero-shot classification. Although these works do not overlap with this one, we find that the uni-modal tasks provide reasonable uniformity within the visual/textual feature embedding, contrary to the cross-modal contrastive shown in Section 2. Therefore, the model could obtain a more "numerically relaxed" triangular inequality when dealing with noisy pairs of samples.

6 Conclusion

Summary: This work discusses the essential properties of the feature embedding space for contrastive alignment. We show that the most commonly adopted cosine similarity has disadvantages in dealing with noisy data and training stability. Therefore, we propose to combine the product sphere with the negative inner product distance to tackle these problems. We employ multiple class tokens to implement the approach, which performs better in various zero-shot classification and image-text retrieval tasks practically.

Limitation: First, given significantly constrained computational resources (and time), we acknowledge our inability to conduct experiments on a larger scale regarding batch size, training data, and neural network parameters. However, the reported results are robust enough to substantiate our claims. Second, in recent studies, besides the contrastive alignment, more pre-training tasks are appended to the head of the model using the non-normalized full token embedding. Such as image-text matching Li et al. (2021a); Yang et al. (2022), image captioning Yu et al. (2022), or masked modeling that do not employ the contrastive alignment Wang et al. (2022). The performance improvement resulting from a better contrastive alignment could be marginal in these configurations. Hence leave future work on designing the model of the full token embedding.

References

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3558–3568, 2021.

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

Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. *Advances in Neural Information Processing Systems*, 34:11834–11845, 2021a.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

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

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*, pp. 9640–9649, 2021b.

Ruizhe Cheng, Bichen Wu, Peizhao Zhang, Peter Vajda, and Joseph E Gonzalez. Data-efficient language-supervised zero-shot learning with self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3119–3124, 2021.

Sanghyuk Chun. Improved probabilistic image-text representations. *arXiv preprint arXiv:2305.18171*, 2023.

Sanghyuk Chun, Wonjae Kim, Song Park, Minsuk Chang Chang, and Seong Joon Oh. Eccv caption: Correcting false negatives by collecting machine-and-human-verified image-caption associations for ms-coco. In *European Conference on Computer Vision (ECCV)*, 2022.

Yufeng Cui, Lichen Zhao, Feng Liang, Yangguang Li, and Jing Shao. Democratizing contrastive language-image pre-training: A clip benchmark of data, model, and supervision, 2022.

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013.

Weiwei Gu, Aditya Tandon, Yong-Yeol Ahn, and Filippo Radicchi. Principled approach to the selection of the embedding dimension of networks. *Nature Communications*, 12(1):3772, 2021.

Haochen Han, Kaiyao Miao, Qinghua Zheng, and Minnan Luo. Noisy correspondence learning with meta similarity correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7517–7526, 2023.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021b.

Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34: 29406–29419, 2021.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip. *Zenodo*, July 2021. doi: 10.5281/zenodo.5143773. URL `https://doi.org/10.5281/zenodo.5143773`.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

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*, pp. 3128–3137, 2015.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, pp. 4171–4186, 2019.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021a.

Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017.

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*, 2021b.

Zheng Li, Caili Guo, Zerun Feng, Jenq-Neng Hwang, and Zhongtian Du. Integrating language guidance into image-text matching for correcting false negatives. *IEEE Transactions on Multimedia*, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021.

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

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V Le. Combined scaling for zero-shot transfer learning. *arXiv preprint arXiv:2111.10050*, 2021.

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*, pp. 2641–2649, 2015.

Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 4948–4956, 2022.

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*, pp. 8748–8763. PMLR, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International conference on machine learning*, pp. 2152–2161. PMLR, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Erich Schubert. A triangle inequality for cosine similarity. In *International Conference on Similarity Search and Applications*, pp. 32–44. Springer, 2021.

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

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

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 the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15638–15650, 2022.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8968–8975, 2020.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59 (2):64–73, 2016.

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

Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5005–5013, 2016.

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

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

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*, pp. 3733–3742, 2018.

Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 69–77, 2016.

Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2022.

Shuo Yang, Zhaopan Xu, Kai Wang, Yang You, Hongxun Yao, Tongliang Liu, and Min Xu. Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19883–19892, 2023.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

A Detailed Training Hyper-parameters Used in Experiments in the Main Manuscripts

Hyperparameters	Value for Naive CLIP	Value for CLIP-Multi(32,16)	Value for Ablation Table 7	Value for Ablation Table 6
Batch size	32,768	32,768	2,048	2,048
Vocabulary size	30,522	30,522	30,522	30,522
Training epochs	32	32	15	15
Number [CLS] Tokens	1	16	1/8	2/8/32/128
Projection dims	512	32	512/64	256/64/16/4
Maximum temperature	100.0	3.95	100.0	100.0
Weight decay	0.2	0.2	0.5	0.5
Warm-up iterations	2,000	2,000	5,000	5,000
Peak Learning Rate	0.0005	0.0005	0.0005	0.0005
Adam β_1	0.9	0.9	0.9	0.9
Adam β_2	0.998	0.998	0.98	0.98
Adam ϵ	10^{-8}	10^{-8}	10^{-8}	10^{-8}
Gradient global norm	1.0	1.0	1.0	1.0
GPUs	128×A100	128×A100	32×V100	32×V100
Train Time	∼5 days	∼5 days	∼1 day	∼1 day

Table 5: Detailed hyper-parameters used for in the experimental analysis.

We provide the hyper-parameters employed in the experiments in Table 5. We follow most of the hyper-parameters employed in the original CLIP Radford et al. (2021) paper for both Naive CLIP re-implementation and our multi-token and single-token product sphere implementation. We provide the details of the hyper-parameters for large-scale and ablation experiments below.

Large-scale Experiments: We train with a batch size of 32,768 and the AdamW optimizer Loshchilov & Hutter (2017) in all the large-scale experiments. We apply the standard training scheme of the original CLIP model, which contains 32 epochs of training. We did not employ mixed precision to reduce the possible overflow introduced by randomness for a stable reproduction. We set the $\beta_1 = 0.9$, $\beta_2 = 0.998$, $\epsilon = 1e\text{-}8$ in AdamW, and weight decay = 0.2 to further improve the stability. We use the cosine learning rate decay scheme of peak learning rate equal to $5e\text{-}4$, combined with a warmup period of 2,000 iterations. For data augmentation, we only apply the `RandomResizedCrop` with a scale range of $[0.8, 1.0]$. Finally, in our multi-token product sphere implementation, we reduced the maximum temperature to 3.95 due to its border distance range. This is a value obtained from the ablation study from Appendix E.

Ablation Experiments: We train with a batch size of 2,048 and the AdamW optimizer Loshchilov & Hutter (2017) in all the ablation experiments. We apply a compact training scheme that updates the model for 108,000 iterations, which is roughly equal to training the model for 15 epochs of the dataset. Since this is a fast training scheme, we set the $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 1e\text{-}8$ in AdamW, and weight decay = 0.5, such that the training could converge faster in a stable approach. We use the cosine learning rate decay scheme of peak learning rate equal to $5e\text{-}4$, combined with a warmup period of 5,000 iterations. In the linear probe evaluation, the hyperparameters follow the setup of MoCo v3 Chen et al. (2021b). Concretely, we use SGD without momentum and no weight decay. The learning rate is schemed by cosine decay with a peak learning rate equal to 1.0, combined with a warmup period of 5 epochs. We train for 100 epochs and augment the image using the `RandomResizedCrop` with a scale range of $[0.75, 1.0]$ and `AutoAugment` with the code `rand-m9-mstd0.5-inc1`. In the ablation experiments, we do not change the maximum temperature clip value, leaving it the same for all topology configurations.

B Choices on structures and multi-token implementation

In Table 6, we modify the structure of the product sphere manifold under fixed total dimensions. It can be seen that a higher m value (*i.e.* the number of product sub-spheres) is more likely to obtain a better zero-shot classification accuracy and text-to-image retrieval recall. We, therefore, conjecture that the broader distance

Topology	Distance	Zero-Shot I2T R@1	Zero-Shot T2I R@1	Zero-Shot Cls. Acc.	Linear PrPSe Cls. Acc.
Temperature, init=$e^{2.64}$, gradient=True					
Sphere(512)	$-\boldsymbol{u}^T\boldsymbol{v}$	48.3	31.45	30.62	60.38
PS(256, 2)	$-\text{tr}(\boldsymbol{u}^T\boldsymbol{v})$	48.0	32.25	30.33	60.52
PS(64, 8)	$-\text{tr}(\boldsymbol{u}^T\boldsymbol{v})$	**52.3**	32.89	30.70	60.32
PS(16, 32)	$-\text{tr}(\boldsymbol{u}^T\boldsymbol{v})$	50.4	**33.01**	**30.93**	**60.79**
PS(4, 128)	$-\text{tr}(\boldsymbol{u}^T\boldsymbol{v})$	48.2	32.91	30.57	59.99
Multi(256, 2)	$-\text{tr}(\boldsymbol{u}^T\boldsymbol{v})$	49.2	32.29	30.04	61.59
Multi(64, 8)	$-\text{tr}(\boldsymbol{u}^T\boldsymbol{v})$	54.0	**34.27**	**31.93**	62.41
Multi(16, 32)	$-\text{tr}(\boldsymbol{u}^T\boldsymbol{v})$	**54.0**	33.43	30.88	**63.71**

Table 6: The retrieval and classification performance of the proposed approach using different PSlique manifold structures and the multi-token implementation. "gradient={True/False}" donates if the temperature is learnable.

range helps the system reach equilibrium faster. However, an over-complicated structure such as PS(4,128) could ruin the performance. The possible reason is that each sub-sphere $\mathbb{S}^{d-1}$ that is embedded in $\mathbb{R}^d$ has one less effective dimension. Therefore, the product sphere structure with large numbers of sub-spheres may perform worse.

However, an over-complicated structure such as PS(4,128) could ruin the performance. We conjecture that, since the sphere has one redundant dimension, the larger number of product sub-spheres reduces the representation capacity of the topology. And because we employ a textual encoder that is gently larger than the visual one, therefore the moderate reduction of the capacity helps overcome the overfitting on the textual side.

We also provide the ablation results of different multi-tokens product sphere structure implementations in the table's lower half, denoted as Multi$(\cdot,\cdot)$. We concatenate all the representations together for the linear probe before projecting them to 1,000 class logit. It can be seen that the multi-token product sphere implementations consistently outperform their single-token versions. Notably, since the increased number of parameters for the class tokens ($n \times d$) is negligible compared to that of the overall system, we consider the participants of class tokens in global attention as the primary reason for the performance boost.

C Detailed performance of configurations under different temperature initilization

In Table 7, we provide the detailed experimental results of Figure 2 in the main manuscripts. We further provide the final temperature at the end of training and at what step the temperature converges (changes less than 2% for an epoch). It can be seen that the performance of the Euclidean topology is only slightly affected by the initialization of the temperatures, and even though the temperature is detached from learning, it still performs reasonably well because of the unlimited distance range. At the same time, the spherical and product sphere topologies are affected by how the temperature is initialized. However, a rough trend can be seen that the faster the temperature converges, the better performance the model achieves, which means the learnable temperature delays the learning of the methods. The model needs first to find a proper temperature and then begin to learn representations well.

D Distribution of Learned Distance

We depict the distribution of distance for pairs of samples in Figure 8. We further employ the RedCaps Desai et al. (2021) dataset as the out-domain data for visualizing the distributions of sample distances. As argued in Section 3.2 of the main manuscripts, since the cross-modal contrastive loss does not handle the uni-modal data distributions, the distance between negative pairs of images and texts could be much smaller than that of a positive image-text pair, resulting in a tighter distance bound. Also, we can see this phenomenon

Temp. Init.	Temp. Final	Converge Step	Zero-Shot I2T R@1	Zero-Shot T2I R@1	Zero-Shot Cls. Acc.	Linear Cls. Acc.
Topology: Sphere, Distance: $-\boldsymbol{u}^T\boldsymbol{v}$						
2.659	4.033	18k	48.3	31.45	30.62	60.38
5.310	4.021	39k	46.8	31.13	29.60	59.82
1.000	3.976	22k	49.0	30.33	28.59	59.56
1.000	1.000	Detach	5.1	3.461	4.04	45.37
Topology: Euclidean, Distance: $\|\boldsymbol{u}-\boldsymbol{v}\|_2$						
2.659	2.067	21k	47.9	32.29	30.36	59.90
5.310	5.107	1k	48.5	32.69	30.68	59.74
1.000	1.668	25k	47.4	30.71	29.85	60.09
1.000	1.000	Detach	47.6	30.43	29.51	59.20
Topology: PS(64, 8), Distance: Geo$(\boldsymbol{u}, \boldsymbol{v})$						
2.659	3.135	20k	50.7	32.35	**30.79**	60.43
5.310	3.168	55k	50.7	31.59	30.34	59.60
1.000	3.024	42k	49.9	32.49	30.21	60.61
1.000	1.000	Detach	4.1	2.921	3.10	21.67
Topology: PS(64, 8), Distance: $-\text{tr}(\boldsymbol{u}^T\boldsymbol{v})$						
2.659	2.231	24k	**52.3**	32.89	30.70	60.32
5.310	2.280	57k	50.3	**33.37**	30.23	59.76
1.000	2.174	36k	50.9	32.71	30.50	**60.66**
1.000	1.000	Detach	30.3	18.48	21.20	57.93

Table 7: The retrieval and classification performance of different configurations under different temperature initialization conditions. "Temp. Init." denotes the values for initializing temperature; "Temp. Final" denotes the final temperature at the end of training; "Converge Step" denotes the number of steps for temperature starts to converge (changes less than 2% for an epoch.)

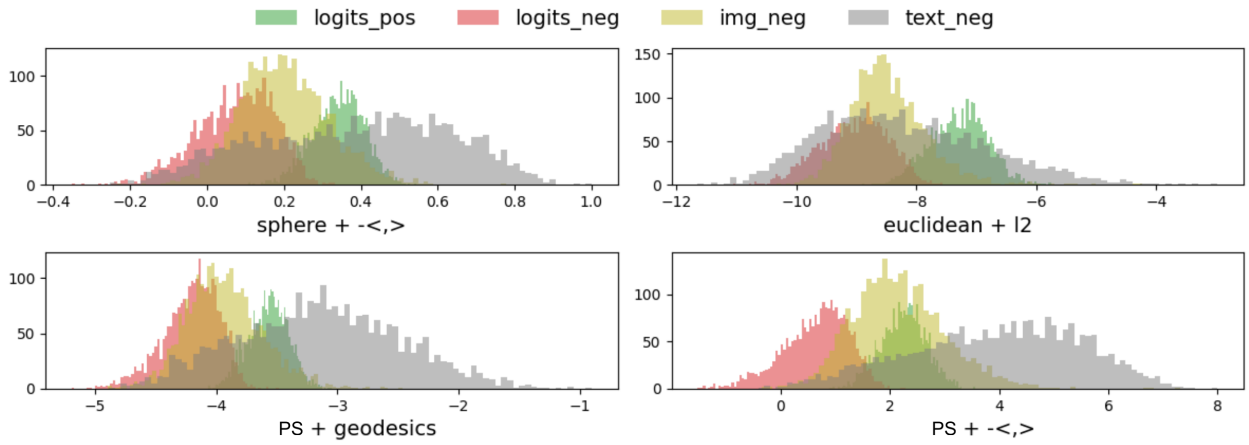
Topology	Distance	Zero-Shot I2T R@1	Zero-Shot T2I R@1	Zero-Shot Cls. Acc.	Linear Probe Cls. Acc.
Temperature, init=$e^{2.64}$, gradient=True					
Sphere(512)	$-\boldsymbol{u}^T\boldsymbol{v}$	48.3	31.45	30.62	60.38
Sphere(1024)	$-\boldsymbol{u}^T\boldsymbol{v}$	50.7	32.05	29.60	**60.53**
PS(128, 8)	$-\text{tr}(\boldsymbol{u}^T\boldsymbol{v})$	49.4	32.85	30.55	60.12
PS(64, 16)	$-\text{tr}(\boldsymbol{u}^T\boldsymbol{v})$	50.3	33.25	30.34	60.16
PS(32, 32)	$-\text{tr}(\boldsymbol{u}^T\boldsymbol{v})$	**52.3**	**33.47**	**30.62**	60.32

Table 8: The retrieval and classification performance of the proposed approach using different oblique manifold structures and the multi-token implementation. "gradient={True/False}" donates if the temperature is learnable.

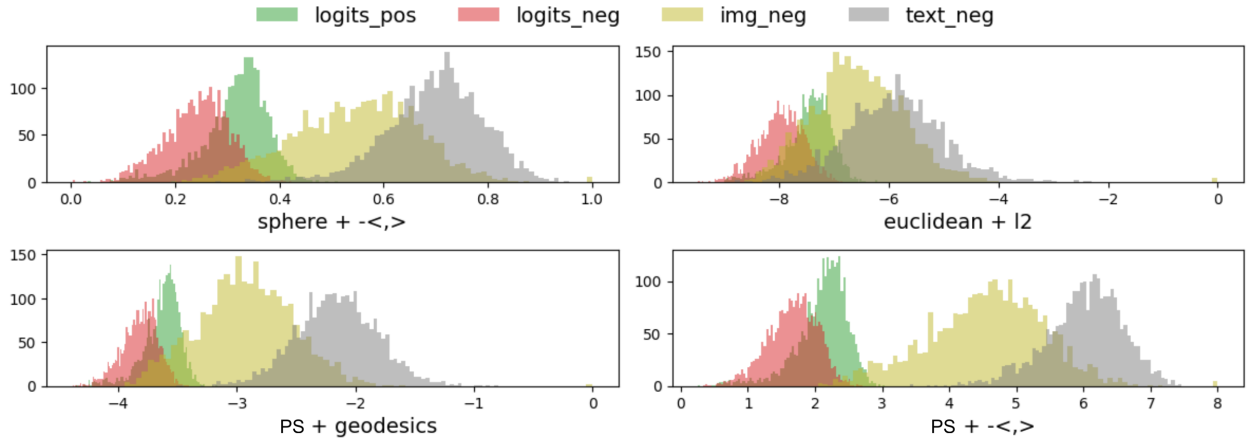
is much more severe in out-domain data, which could reduce the transferability of the feature embeddings to downstream tasks. It is also notable that, the product sphere with the negative inner product as the distance function learns similar distributions compared to the sphere reference, while the numerical values of distances between samples are inherently larger without having multiplied with temperature.

E Additional Ablation on product sphere Structure

We provide more ablation results regarding the structure of the product sphere manifold under fixed total dimensions in Table 8. We can observe that the PS(32, 32) configuration performs the best in general, while the sphere with more 1024-dimensional embedding has slightly better linear probe performance. We also notice that a more complicated structure provides better text-to-image retrieval results.



(a)



(b)

Figure 8: Visualization of the distribution of distances between samples. The `logits_pos` and `logits_neg` denote the distances between positive and negative image-text pairs, respectively. The `img_neg` and `text_neg` denote the distances between negative image-image and text-text pairs, respectively. The models are trained using the Yfcc datasets, (a) and (b) depict the distribution of in-domain data (Yfcc) and out-domain data (RedCaps), respectively.

F Additional Results Using the TCL Framework

We combine our proposed method with the TCL model Yang et al. (2022), which is one of the state-of-the-art vision-language retrieval models that employ contrastive visual-textual alignment in its earlier stage. During the pre-training, the TCL induces a mixture of in-modal and cross-modal contrastive losses, while conducting the masked language modeling (MLM) and image-text matching tasks simultaneously. During the testing, the cross-modal contrastive alignment head first lists sample pairs with high similarity scores, and then these pairs are fed into the matching head to obtain the final matching scores. We alternate the topologies of all the embedding spaces with PS(128,2). For the experimental analysis in this subsection, we follow the configurations of the reference models, employ a collection of CC3M Sharma et al. (2018), MSCOCO Captions Chen et al. (2015), Visual genome Krishna et al. (2017) and SBU Ordonez et al. (2011) as the pre-training dataset, which contains roughly 4 million annotated image-text pairs. The models are then evaluated using Flickr30k Plummer et al. (2015) and MSCOCO Captions Chen et al. (2015).

The results are shown in Table 9. Since our method does not affect the matching head, we also report the performance of the contrastive alignment head. In general, our method improves the average recall

Method *baseline*[*impl.*]	Flickr **I2T** R@1	Flickr **T2I** R@1	Flickr **Recall** mean	Coco **I2T** R@1	Coco **T2I** R@1	Coco **Recall** mean
Zero-shot performance.						
TCL[official]	93.00	79.60	93.97	71.40	53.50	79.49
	(84.20)	(67.10)	(88.45)	(55.40)	(40.80)	(69.92)
TCL[our-impl.]	91.00	78.28	93.25	70.16	53.05	79.07
	(83.30)	(68.40)	(88.73)	(57.34)	(43.21)	(71.31)
TCL[PS(128,2)]	91.20	78.14	**93.29**	70.14	53.35	**79.14**
	(84.80)	(67.86)	(88.84)	(57.10)	(43.13)	(71.32)
Fine-tuned performance.						
TCL[official]	94.90	84.00	95.57	75.60	59.00	82.87
	(87.90)	(71.38)	(90.92)	(65.34)	(48.94)	(76.53)
TCL[our-impl.]	93.80	83.06	95.17	73.56	57.74	82.06
	(88.30)	(72.94)	(91.27)	(66.98)	(50.34)	(77.43)
TCL[PS(128,2)]	93.80	82.90	**95.18**	74.78	57.72	**82.13**
	(88.60)	(73.26)	(91.39)	(65.60)	(49.83)	(76.86)

Table 9: Retrieval performance on Flickr30K and MSCOCO of our implemented TCL model and the variant using our proposed method. The numbers in brackets are the performance obtained using the contrastive alignment head.

#Tokens	**1**	**2**	**4**	**8**
Top1 Acc.	13.0±9.7	21.5±14.3	27.7±20.4	62.1 ±4.4

Table 10: ImageNet zero-shot classification performance of CLIP[Multi(32,16)] model using a randomly selected subset of [CLS] tokens.

performance, but the improvement is not significant. We consider the reasons as i) The method (or recent similar methods) employs pre-trained vision and language models, as well as a matching head and an MLM head; hence it is less sensitive to the gradients from the contrastive alignment; ii) The datasets employed for training contain less noise, while the training is scheduled with an overlength scheme (the zero-shot performance does not increase in the last 5 epochs).

Additional Notes on TCL We also provide the comparison results with officially released checkpoints. It can be seen that our implementation performs 0.5-1.0% worse than the official checkpoints. On the other hand, our implementation has better alignment head performance. Since we are employing the codes released in the official repository, the reason might be the following: i) Datasets difference, that we have ∼3000 fewer images in the SBU dataset while owning 5000 more images in the CC3M dataset; ii) We resize the CC3M dataset to short edge 500 pixels, while the official repository does not clearly provide the pre-processing approach; iii) We implicitly have a short training time or smaller matching loss weight than the official checkpoints due to the difference in the framework.

G Test of Mixture-of-Expert Hypothesis:

We investigate the mixture-of-expert hypothesis of the proposed method. Since the class token is considered to encode the global representation of the sample, the employment of multiple class tokens may function in a mixture-of-expert style. That is, after training, each sub-sphere (or a subset of sub-spheres) in the product sphere structure is capable of alignment. Then, the system functions as a mixture of weak alignment models (experts). To test this hypothesis, we calculate the zero-shot classification performance of the CLIP[Multi(32,16)] model with randomly selected subsets of sub-spheres. From Table 10, we find that the drop in performance is reasonably small (∼12%) with half of the alignment tokens. This result reveals a

possible mechanism of the product sphere structure during optimization, where a subset of sub-spheres is priorly aligned.

H More Visualization using GradCAM

In Figure 9, we provide more visualization results using GradCAM. In Figure 10, we show some failure cases when the attention in the textual mode is focused on non-object words.

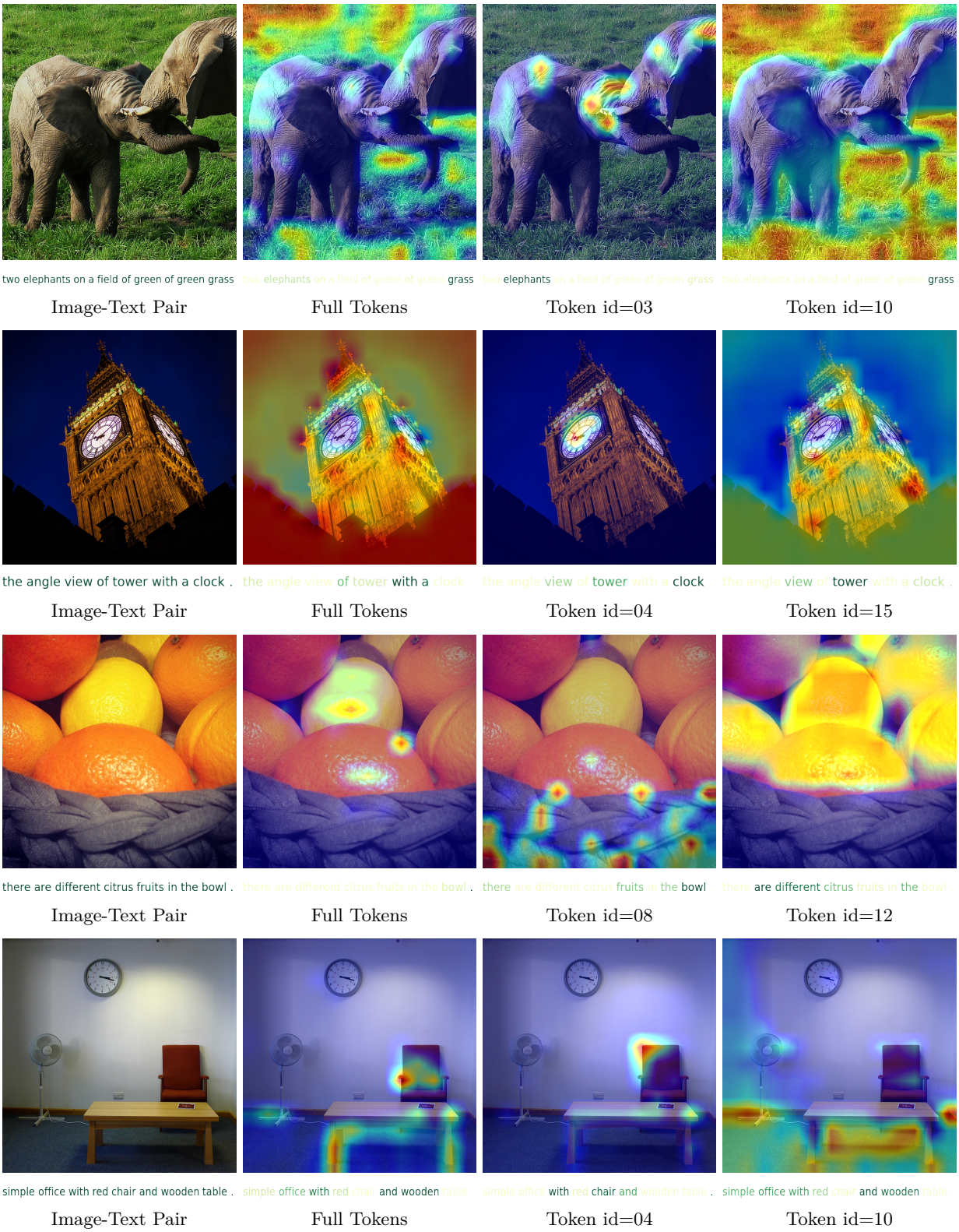


Figure 9: More visualization of the importance map using the Grad-CAM algorithm.

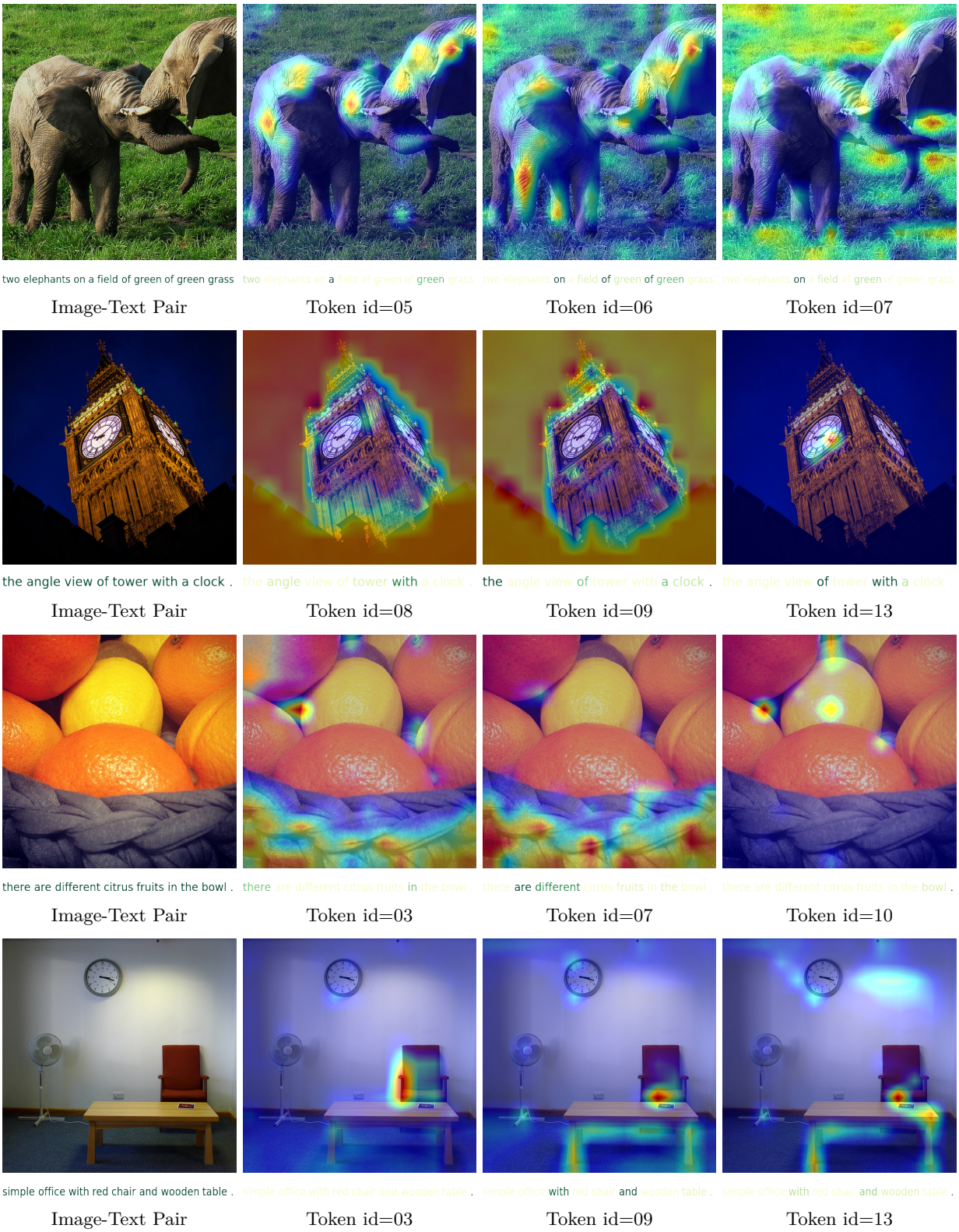


Figure 10: Failure cases of the importance map using the Grad-CAM algorithm.

I A note on the recent advance in noisy image-text matching

Recently, many pieces of research have been made to tackle the noisy image-text matching problem in contrast learning. Below, we provide a concise survey of these works, for the readers who want to know more about this topic. Although these works may not be comparable with our proposed method, they still support that noisy visual-textual correspondences are an important research topic in this field.

Chun et al. (2022): This paper argues that existing ITM benchmarks have a significant limitation of many missing correspondences. Then, it proposes a new dataset, ECCV Caption, to correct the massive false negatives and a new metric, mAP@R, to evaluate VL models.

Li et al. (2023): This paper proposes a method to correct false negatives by integrating language guidance into the ITM framework. This framework corrects the locations of false negatives in the embedding space.

Chun (2023): This paper also argues that the image-text matching task suffers from ambiguity due to multiplicity and imperfect annotations. Then, this paper proposes an improved probabilistic ITM approach that introduces a new probabilistic distance with a closed-form solution.

Huang et al. (2021): This paper points out that the training data may contain mismatched pairs. To learn the noisy correspondence, the authors divide the data into clean and noisy partitions and then rectify the correspondence via an adaptive prediction model.

Qin et al. (2022): This paper considers the major challenge in cross-modal retrieval is the noisy correspondence in training data. This refers to the fact that some of the training pairs may not be correctly aligned, *i.e.*, the image and text do not actually correspond to each other. They propose a framework to address this challenge by integrating two novel techniques: Cross-modal Evidential Learning and Robust Dynamic Hinge.

Yang et al. (2023): This paper proposes a general framework for cross-modal matching that can be easily integrated into existing models and improve their robustness against noisy data. This framework estimates soft labels for noisy data pairs by exploiting the consistency of cross-modal similarities.

Han et al. (2023): The paper proposes a Meta Similarity Correction Network to provide reliable similarity scores for cross-modal retrieval. The method learns to distinguish between positive and negative pairs of data using meta-data, and can be used to remove noisy samples from the training dataset.