On the Generalization of Multi-modal Contrastive Learning

Qi Zhang[*1] **Yifei Wang**[*2] **Yisen Wang**[1 3]

Abstract

Multi-modal contrastive learning (MMCL) has recently garnered considerable interest due to its superior performance in visual tasks, achieved by embedding multi-modal data, such as visual-language pairs. However, there still lack theoretical understandings of how MMCL extracts useful visual representation from multi-modal pairs, and particularly, how MMCL outperforms previous approaches like self-supervised contrastive learning (SSCL). In this paper, by drawing an intrinsic connection between MMCL and asymmetric matrix factorization, we establish the first generalization guarantees of MMCL for visual downstream tasks. Based on this framework, we further unify MMCL and SSCL by showing that MMCL implicitly performs SSCL with (pseudo) positive pairs induced by text pairs. Through this unified perspective, we characterize the advantage of MMCL by showing that text pairs induce more semantically consistent and diverse positive pairs, which, according to our analysis, provably benefit downstream generalization. Inspired by this finding, we propose several methods to significantly improve the downstream performance of SSCL on ImageNet by leveraging multi-modal information. Code is available at `https://github.com/PKU-ML/CLIP-Help-SimCLR`.

1. Introduction

Recently, multi-modal contrastive learning (MMCL), including CLIP (Radford et al., 2021) and its variants (Li et al., 2022b; Mu et al., 2022; Yao et al., 2022), has achieved impressive performance for visual representation learning, and transfer well to various downstream tasks like zero-shot and few-shot image classification. The core idea of MMCL is rather simple, which aligns the samples of the same image-text pairs together while pushing away other unrelated samples in the latent feature space. However, it remains not fully clear to us why matching multi-modal pairs would benefit visual representation learning, and what are the key factors that affect its downstream performance.

Meanwhile, another popular scenario for contrastive learning is self-supervised learning, which also obtains competitive performance recently (Chen et al., 2020; He et al., 2020; Wang et al., 2021). Nevertheless, recent MMCL methods (like CLIP) have shown significant advantages over its self-supervised contrastive learning (SSCL) counterparts like SimCLR (Chen et al., 2020). Existing theories of SSCL (Saunshi et al., 2019; HaoChen et al., 2021; Wang & Isola, 2020) only establish the optimality of self-supervised representations on downstream tasks, and fail to characterize why MMCL could outperform SSCL. Another major obstacle is the generation process of data pairs. In particular, positive pairs in SSCL are visual-only samples generated by random data augmentations of the raw image; Instead, the positive pairs in MMCL are multi-modal (*e.g.,* visual-language) pairs directly provided by the dataset. Since existing SSCL theories rely crucially on the assumption that data augmentations produce overlap between visual samples (Wang et al., 2022; Saunshi et al., 2022), they cannot be directly applied to MMCL that relies on multi-modal data pairs.

In this paper, we propose the first theoretical analysis on the generalization ability of MMCL. To achieve this, we establish an equivalence between the MMCL objective and the asymmetric matrix factorization (AMF) of the multi-modal co-occurrence matrix. Built upon this connection, we characterize the ideal pretrained representations of MMCL and its generalization bounds on visual and language downstream tasks, where the bounds are influenced by the properties of the multi-modal co-occurrence matrix, for example, its singular value.

The established theoretical framework also allows us to characterize the difference between MMCL and SSCL under a unified perspective. To be specific, we first formally unify MMCL and SSL under the framework of uni-modal similarity graphs, where language pairs in MMCL can be regarded as a special kind of data augmentation for generating pos-

[*] Equal contribution [1] National Key Lab of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University [2] School of Mathematical Sciences, Peking University [3] Institute for Artificial Intelligence, Peking University. Correspondence to: Yisen Wang <yisen.wang@pku.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

itive visual pairs. Based on this perspective, we compare MMCL and SSCL on real-world data and show that text-induced positive pairs have better semantic consistency and diversity than augmentation-based ones in SSCL, which explains the superiority of MMCL on downstream tasks. Besides the empirical comparisons, we theoretically analyze this difference by modeling the data generation process with the hierarchical random graph (Clauset et al., 2008). Based on this understanding, we further leverage multi-modal information in CLIP to assist the self-supervised visual learning with SimCLR on ImageNet and achieve significant improvements, which validates our understanding of the superiority of multi-modal positive pairs.

We summarize our contributions as follows:

- We establish the first generalization theoretical guarantee for multi-modal contrastive learning (MMCL). We provide a new perspective of the multi-modal contrastive loss by connecting it with an asymmetric matrix decomposition objective.
- We provide a unified perspective for understanding the connections and differences between multi-modal and self-supervised contrastive learning. Based on this perspective, we examine their differences on real-world data, and find that multi-modal information induces better positive visual pairs than self-supervision (with better semantic consistency and diversity), which explains the superiority of MMCL.
- As a verification of our understanding above, we further investigate a new scenario where we leverage multi-modal information in pretrained models (like CLIP) to assist self-supervised learning like SimCLR. We propose four different techniques and they both bring improvements (as much as 6.2%) on ImageNet.

2. Related Work

Multi-modal Pretraining Applications. Traditional single-stream models (Lu et al., 2019; Li et al., 2019) have been widely discussed and shown the impressive performance in various multi-modal tasks. However, as they do not have independent encoders for different modals, the transferability of these frameworks is usually limited. On contrast, multi-modal contrastive learning paradigms represented by CLIP (Radford et al., 2021) have recently obtained the promising performance in multi-modal downstream tasks including zero-shot learning, finetuning and linear-probing. Inspired by CLIP, various variants are proposed to improve the efficiency and performance of multi-modal pretraining. SLIP (Mu et al., 2022) and DeCLIP (Li et al., 2022b) combine the self-supervised and multi-modal contrastive learning to accelerate the training process. FILIP (Yao et al., 2022) propose fine-grained multi-modal contrastive objective to make the encoder focus more on the local features.

Theory of Contrasative Learning. Motivated by the empirical success of the contrastive objective, many researchers try to theoretically analyze how it works. Wang & Isola (2020) understand the contrastive loss from two terms in it: the alignment of the positive samples and the uniformity of the negative samples. Hjelm et al. (2019) analyze the objective from the mutual information theory. Saunshi et al. (2019) establish the theoretical guarantee between the pretraining contrastive loss and the downstream classification performance. HaoChen et al. (2021) revisit the contrastive objective from a spectral graph perspective, which explains the relationship between the augmented samples and the downstream performance of contrastive learning. Wang et al. (2022; 2023) provide a theoretical understanding for contrastive learning from the perspective of augmentation overlap and message passing respectively. As these prior theoretical works mainly focus on the single-modal contrastive learning, the theoretical analysis on the multi-modal contrastive learning is still quite limited. In this work, we theoretically analyze the relationship between the design of the multi-modal contrastive paradigms and its generalization ability on downstream tasks.

Theory of Multi-modal Learning. For the theoretical analysis of multi-modal learning, there are few related works. Sun et al. (2020) propose a information-theoretic framework and prove that their method can learn ground-truth Bayesian posterior classifier for each modality and the Bayesian posterior aggregator for all modalities. Huang et al. (2021) proves that the multi-modal models can learn better representations than single-modal models in certain conditions. However, both of their analysis do not focus on the multi-modal *contrastive* paradigm and can not explain why the contrastive methods can achieve such an impressive performance.

3. Generalization Theory of Multi-Modal Contrastive Learning

3.1. Mathematical Formulation

We start by introducing the basic mathematical formulation for multi-modal contrastive learning. Without loss of generality, taking CLIP (Radford et al., 2021) for an example, we have the paired data (x_v, x_l) from the visual domain (x_v denotes an image) and the language domain (x_l denotes a corresponding text description of the image). Each x_v or x_l belongs to one of r classes. We use $\mathcal{X}_V$ to denote the set of all visual data with distribution $\mathcal{P}_V$, and $\mathcal{X}_L$ to denote the set of all language data with distribution $\mathcal{P}_L$. Their joint multi-modal distribution is $\mathcal{P}_M$. For ease of exposition, we assume $\mathcal{X}_V, \mathcal{X}_L$ to be finite but exponentially large sets[1], and denote $N_V = |\mathcal{X}_V|$ and $N_L = |\mathcal{X}_L|$. The goal of multi-modal contrastive learning is to obtain a joint embedding of the visual data $\mathcal{X}_V$ and language data $\mathcal{X}_L$ in the k-dimensional latent space $\mathcal{Z} \in \mathcal{R}^k$ by learning a visual encoder $f_V : \mathcal{X}_V \to \mathcal{Z}$ and a language encoder $f_L : \mathcal{X}_L \to \mathcal{Z}$, such that semantically similar samples (either image-image, text-text or image-text pairs) have close representations, and different samples are apart. A recent work (Tschannen et al., 2022) also explores a Siamese network, *i.e.,* $f_V = f_L$. Here we consider the general case with two different encoders.

For multi-modal positive and negative pairs, we define an image-text pair drawn from the paired visual-language data, *i.e.,* $(x_v, x_l) \sim \mathcal{P}_M$, as positive pairs, and draw independent samples from each domain, $x_v^- \sim \mathcal{P}_V, x_l^- \sim \mathcal{P}_L$, and treat (x_v, x_l^-), (x_v^-, x_l) and (x_v^-, x_l^-) as negative pairs, because samples in these pairs are independent of each other.

Given positive and negative pairs (x_v, x_l, x_v^-, x_l^-), one popular learning objective is the symmetric cross entropy (SCE) loss (adopted in CLIP) calculated over similarity scores:

$$\begin{aligned}\mathcal{L}_{\rm SCE}(f_V, f_L) = &- \mathbb{E}_{x_v, x_l} \log \frac{\exp\left(f_V(x_v)^\top f_L(x_l)\right)}{\mathbb{E}_{x_l^-} \exp(f_V(x_v)^\top f_L(x_l^-))} \\ &- \mathbb{E}_{x_v, x_l} \log \frac{\exp\left(f_V(x_v)^\top f_L(x_l)\right)}{\mathbb{E}_{x_v^-} \exp(f_V(x_v^-)^\top f_L(x_l))}.\end{aligned} \tag{1}$$

This objective can be seen as an extension of the popular InfoNCE loss (Oord et al., 2018) to the multi-modal scenario (Zhang et al., 2020). During the learning process, positive pairs (x_v, x_l) are pulled together in the latent space while negative pairs (x_v, x_l^-) and (x_v^-, x_l) are pushed apart. Following the same spirit, we consider a similar multi-modal spectral loss for the ease of theoretical analysis,

$$\begin{aligned}&\mathcal{L}_{\rm SCL}(f_V, f_L) \\ =& - 2\mathbb{E}_{x_v, x_l} f_V(x_v)^\top f_L(x_l) + \mathbb{E}_{x_v^-, x_l^-}(f_V(x_v^-)^\top f_L(x_l^-))^2.\end{aligned} \tag{2}$$

Comparing Eq. 1 and Eq. 2, we can easily see that the two objectives have the same loss for positive pairs, and only differ at the specific loss function used for pushing negative pairs apart (logsumexp loss in Eq. 1 v.s. ℓ_2 loss in Eq. 2). The multi-modal spectral loss can be regarded as an extension of the visual spectral contrastive loss that achieves comparable performance to the InfoNCE loss in visual tasks (HaoChen et al., 2021). Nevertheless, their analysis can only be applied to self-supervised contrastive learning where positive and negative pairs come from the same domain.

After pretraining, we evaluate the learned representations by applying them to downstream tasks. Taking the visual linear probing task as an example, we train a linear classifier to predict class labels $y \in \mathcal{Y}$ from the output features of f_V by $g_{f,B_V}(x_v) = \arg\max_{i \in [r]} (f_V(x_v)^\top B_V)_i$, where $B_V \in \mathbb{R}^{k \times r}$ denotes the weight matrix. The linear probing error of f_V is defined as the error of the optimal linear classifier on the encoded features, *i.e.,*

$$\mathcal{E}(f_V) = \min_{B_V} \mathbb{E}_{x_v \sim \mathcal{P}_V} \mathbb{1}[g_{f,B_V}(x_v) \neq y(x_v)], \tag{3}$$

where $y(x_v)$ denotes the label of x_v. Likewise, we can define the linear probing error $\mathcal{E}(f_L)$ for the text classification.

3.2. An Asymmetric Matrix Factorization View of Multi-modal Contrastive Learning

With its samplewise pretraining objective (Eqs. 1 & 2), multi-modal contrastive learning (MMCL) is usually understood as an instance-level feature matching task between visual and language domains (Radford et al., 2021). However, little is known about the overall distribution of the learned features, which hinders us from understanding how its instance-level pretraining benefits downstream applications. In this section, with a reformulation of the MMCL objective, we show that MMCL is essentially equivalent to the asymmetric matrix factorization (AMF) of the joint data distribution $\mathcal{P}_M(x_v, x_l)$. AMF is an important class of methods in classical machine learning with inherent connections to PCA, K-means, and spectral clustering (Ding et al., 2005), and is widely adopted in unsupervised learning scenarios like Latent Semantic Analysis (Deerwester et al., 1990) and word embedding (Pennington et al., 2014). Generally speaking, AMF can extract low-frequency components that underline the common structure of the joint distribution, which is helpful for MMCL analysis.

We start by formulating the joint distribution $\mathcal{P}_M(x_v, x_l)$ as a *co-occurrence matrix* $P_M \in \mathbb{R}^{N_V \times N_L}$ between all visual-language data pairs, where

$$(P_M)_{x_v, x_l} = \mathcal{P}_M(x_v, x_l) \geq 0, \ \forall \, x_v \in [N_V], x_l \in [N_L]. \tag{4}$$

We can see that P_M is a non-negative asymmetric matrix that can be exponentially large. A canonical assumption of representation learning is that high-dimensional data (like images and text) lie in a low-dimensional manifold. Then, we consider the following low-rank matrix factorization for the *normalized co-occurrence matrix* $\tilde{P}_M$:

$$\mathcal{L}_{\rm AMF}(F_V, F_L) = \|\tilde{P}_M - F_V F_L^\top\|^2, \tag{5}$$

where $F_V \in \mathcal{R}^{N_V \times k}, F_L \in \mathcal{R}^{N_L \times k}$ are factorized low-rank components ($k \ll \min(N_V, N_L)$) of the visual and language domains, respectively. To obtain the normalized co-occurrence matrix $\tilde{P}_M$, we adopt two-side normalization

$$(\tilde{P}_M)_{x_v, x_l} = \frac{\mathcal{P}_M(x_v, x_l)}{\sqrt{\mathcal{P}_V(x_v)\mathcal{P}_L(x_l)}}, \tag{6}$$

[1] With some non-essential nuances as in HaoChen et al. (2021), our analysis can also be extended to the infinite data setting.

where $\mathcal{P}_V(x_v) = \sum_{x_l} \mathcal{P}_M(x_v, x_l)$ denotes the marginal probability of x_v, and $\mathcal{P}_L(x_l) = \sum_{x_v} \mathcal{P}_M(x_v, x_l)$ denotes the marginal probability of x_l. Based on this formulation, we are ready to establish the key result of this paper.

Theorem 3.1 (Equivalence)**.** *Let the x_v-row of F_V and the x_l-row of F_L represent the corresponding encoded features of these samples in the following form,*

$$(F_V)_{x_v} = \sqrt{\mathcal{P}_V(x_v)} f_V(x_v)^\top, \tag{7a}$$

$$(F_L)_{x_l} = \sqrt{\mathcal{P}_L(x_l)} f_L(x_l)^\top. \tag{7b}$$

Then low-rank asymmetric matrix factorization loss (Eq. 5) is equivalent to the multi-modal contrastive loss (Eq. 2) up to a constant,

$$\mathcal{L}_{\mathrm{AMF}}(F_V, F_L) = \mathcal{L}_{\mathrm{SCL}}(f_V, f_L) + const. \tag{8}$$

Proof. Taking the definition of F_V and F_L in Eq. 7 into the decomposition loss $\mathcal{L}_{\mathrm{AMF}}(F_V, F_L)$, and combing with the definition of $\tilde{P}_M$ in Eq. 6, we have

$$\begin{aligned}
&\mathcal{L}_{\mathrm{AMF}}(F_V, F_L) \\
=&\|\tilde{P}_M - F_V F_L^\top\|^2 \\
=&\sum_{x_v, x_l} \Bigg(\frac{\mathcal{P}_M(x_v, x_l)}{\sqrt{\mathcal{P}_V(x_v)\mathcal{P}_L(x_l)}} \\
&\qquad - \sqrt{\mathcal{P}_V(x_v)} f_V(x_v)^\top \sqrt{\mathcal{P}_L(x_l)} f_L(x_l) \Bigg)^2 \\
=&\sum_{x_v, x_l} \Bigg(\frac{\mathcal{P}_M(x_v, x_l)^2}{\mathcal{P}_V(x_v)\mathcal{P}_L(x_l)} - 2\mathcal{P}_M(x_v, x_l) f_V(x_v)^\top f_L(x_L) \\
&\qquad + \mathcal{P}_V(x_v)\mathcal{P}_L(x_l) \left(f_V(x_v)^\top f_L(x_L) \right)^2 \Bigg) \\
=&\underbrace{\sum_{x_v, x_l} \left(\frac{\mathcal{P}_M(x_v, x_l)^2}{\mathcal{P}_V(x_v)\mathcal{P}_L(x_l)} \right)}_{const} - 2\mathbb{E}_{x_v, x_l} f_V(x_v)^\top f_L(x_l) \\
&\qquad + \mathbb{E}_{x_v^-, x_l^-} \left(f_V(x_v^-)^\top f_L(x_l^-) \right)^2 \\
=&\mathcal{L}_{\mathrm{SCL}}(f_V, f_L) + const,
\end{aligned}$$

which completes the proof. □

Theorem 3.1 reveals a crucial fact that multi-modal contrastive learning essentially learns the low-rank factorization of the co-occurrence matrix. Meanwhile, we notice that the original factorization loss is actually intractable to directly solve because of the exponentially large size of the co-occurrence matrix P_M, while multi-modal contrastive learning avoids this problem by transforming it into a tractable and scalable objective that simply requires samples from the joint probability $\mathcal{P}_M$. But theoretically, this equivalence allows us to characterize the overall distribution of multi-modal contrastive learning, and provides guarantees on downstream tasks for its ideal representations in the following part.

3.3. Characterizing Ideal Representations of Multi-modal Contrastive Learning

In multi-modal contrastive learning (MMCL) like CLIP (Radford et al., 2021), a common pipeline is to apply the pretrained representations to downstream visual tasks like image classification. Therefore, in order to characterize the pretraining and downstream behaviors of MMCL, it matters for us to understand the properties of the optimally pretrained representations, and how they generalize to downstream tasks.

Ideal Representations. First, we characterize the *general solution* to the multi-modal pretraining loss, under the ideal assumption that the neural networks are expressive enough.

Theorem 3.2. *Let $\tilde{P}_M = U \Sigma V^\top$ is the singular value decomposition (SVD) of the normalized co-occurrence matrix $\tilde{P}_M$ (Eq. 6), where $U \in \mathbb{R}^{N_V \times r}, V \in \mathbb{R}^{r \times N_L}$ are unitary matrices, and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$ contains descending singular values $\sigma_1 \geq \ldots \sigma_r \geq 0, , r = \min(N_V, N_L)$. Assume the neural networks are expressive enough for any features. The multi-modal contrastive loss (Eq. 2) attains its optimum when $\forall\, x_v \in \mathcal{X}_V, x_l \in \mathcal{X}_L$,*

$$f_V^*(x_v) = \frac{1}{\sqrt{\mathcal{P}_V(x_v)}} \left(U_{x_v}^k D R \right)^\top, \tag{9a}$$

$$f_L^*(x_l) = \frac{1}{\sqrt{\mathcal{P}_L(x_l)}} \left(V_{x_l}^k \,\mathrm{diag}(\sigma_1, \ldots, \sigma_k) D^{-1} R \right)^\top, \tag{9b}$$

where U_x takes the x-th row of U, and U^k, V^k denote the submatrices containing the first k columns of U, V, respectively; $D \in \mathcal{R}^{k \times k}$ is an arbitrary invertible diagonal matrix; and $R \in \mathbb{R}^{k \times k}$ is an arbitrary unitary matrix.

Theorem 3.2 shows that the ideal representations of MMCL are largely determined by the k leading eigenvectors, up to some affine transformations (scaling D and rotation R). Although the optimal solution is not unique, when we apply this representation to the linear probing task, the linear classifier can absorb the differences in affine transformations and yield the same classification error for different variants at the optimum. Built upon these optimal representations, we are ready to establish formal guarantees for the generalization of multi-modal contrastive learning on the downstream linear probing tasks in both the visual and language domains.

Theorem 3.3. *Given a specific joint data distribution $\mathcal{P}_M$, we define the labeling error α as the average label agreement among the visual-language positive pairs*

$(x_v, x_l) \sim \mathcal{P}_M$, *i.e.,*

$$\alpha = \mathbb{E}_{x_v,x_l} \mathbb{1}[y(x_v) \neq y(x_l)], \tag{10}$$

where $y(\cdot)$ returns the ground-truth label of the operand. Denote the empirical estimate of the visual and text encoders from n pretraining examples as $\hat{f}_V^*, \hat{f}_L^*$, respectively. With probability $1-\delta$, the visual linear probing error $\mathcal{E}(\hat{f}_V^*)$ and text linear probing error $\mathcal{E}(\hat{f}_L^*)$ can be upper-bounded by

$$\{\mathcal{E}(\hat{f}_V^*),\mathcal{E}(\hat{f}_L^*)\} \lesssim \frac{\alpha}{1-\sigma_{k+1}^2} + \underbrace{\frac{ck}{\Delta_\sigma^2}\left(\widehat{\mathcal{R}}_{n/3}(\mathcal{F}) + \sqrt{\frac{\log 2/\delta}{2n/3}} + \delta\right)}_{\textit{finite-sample generalization terms}} \tag{11}$$

where $\lesssim$ omits some constant terms, σ_{k+1} (c.f. Theorem 3.2) is the $(k+1)$-th largest singular value of the normalized co-occurrence matrix $\tilde{P}_M$. In the finite-sample generalization terms, $\hat{\mathcal{R}}_{n/3}(\mathcal{F})$ denotes a Rademacher complexity of the model class $\mathcal{F}$ with $n/3$ samples, k is the representation dimension, $\Delta_\sigma = \sigma_{\lfloor 3k/4 \rfloor}^2 - \sigma_k^2$, and $c \lesssim (k\kappa + 2k\kappa^2 + 1)^2$ with κ upper bounding $\|f_V(x)\|_\infty$ and $\|f_L(x)\|_\infty$.

In the upper bound of Eq. 11, aside from the canonical generalization terms relating to the number of samples and neural network complexity, there are two important factors reflecting the influence of the multi-modal pretraining task, the labeling error α and the singular value σ_{k+1}.

Labeling error α accounts for the label mismatch between the constructed visual-language pairs, which may differ in practice depending on how the dataset is constructed. For example, the MS-COCO dataset contains human-provided captions for 120K images using Amazon Mechanical Turk (Lin et al., 2014), while the large-scale YFCC dataset (Thomee et al., 2016) contains 99M Flickr images along with their posted titles as captions without filtering or post-processing, which could be quite noisy. A recent work (Santurkar et al., 2022) empirically finds that a single MS-COCO image-caption pair is worth five YFCC captions for CLIP training. These findings can be justified by our theory that the written captions in MS-COCO induce a smaller labeling error α.

Singular value σ_{k+1} is a spectral property of the co-occurrence matrix P_M. One way to understand its role is from a graph perspective. Specifically, we can regard P_M as a (partial) adjacency matrix of a bipartite graph[2] established between the visual set $\mathcal{X}_V$ and the language set $\mathcal{X}_L$. According to the spectral graph theory (Chung, 1997), the singular values generally represent the connectivity of the bipartite graph (*e.g.,* how many disjoint sub-graphs), and smaller leading singular values correspond to better connectivity (*e.g.,* fewer sub-graphs). Therefore, Theorem 3.3 shows that better connectivity (by creating diverse connections between samples) with a smaller σ_{k+1} could bring smaller downstream errors. In fact, several recent works can be understood as increasing the diversity of multi-modal pairs by data augmentations. For example, FLIP (Li et al., 2022a) introduces patch masking to the images input, and Santurkar et al. (2022) rewrite text captions using a GPT model. Our generalization bound provides a theoretical justification for the effectiveness of these approaches.

To warp up, our generalization bounds in Theorem 3.3 provide not only guarantees but also principled guidelines for multi-modal contrastive learning: 1) we should create high-quality multi-modal pairs by human writing or automatic filtering to reduce the labeling error α, and 2) we should create better multi-modal diversity by data augmentations in both domains to ensure a smaller singular value σ_{k+1}.

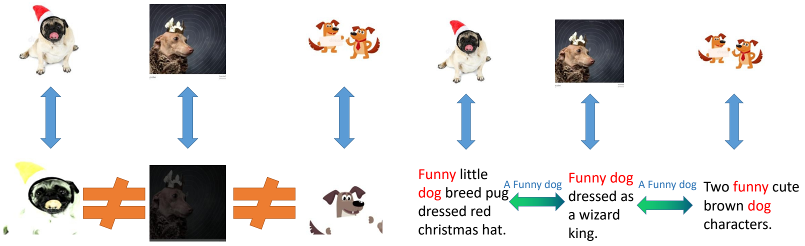
3.4. Discussion

In this section, we establish the first comprehensive study on the theoretical guarantees of multi-modal contrastive learning in terms of two aspects: optimal representations and downstream guarantees. A closely related work is HaoChen et al. (2021) that establishes theoretical guarantees for self-supervised contrastive learning. Our analysis extends their theory to the multi-modal setting, with the following key differences:

1) Data generation. Their analysis only applies to positive pairs (x, x^+) that are both augmented samples from the same domain $\mathcal{X}$, while the multi-modal pair (x_v, x_l) are directly given by data samples and are asymmetric ones from *different domains* $\mathcal{X}_V, \mathcal{X}_L$. Correspondingly, our analysis deals with the multi-modal co-occurrence matrix $\tilde{P}_M$ instead of the aggregated augmentation graph $\tilde{A}$ defined over $\mathcal{X}$ in HaoChen et al. (2021) as the approximation target.

2) Learning objective. Their analysis only applies to the uni-model spectral contrastive loss using a Siamese architecture, which corresponds to *symmetric* matrix factorization. Instead, in multi-modal learning, the positive pairs are not symmetric and require different encoders in general. Correspondingly, we propose the multi-modal spectral contrastive loss that corresponds to *asymmetric* matrix factorization, which requires different techniques to analyze and yield different optimal representations and downstream generalization bounds.

[2] For a bipartite graph, only interleaving edges between $\mathcal{X}_V$ and $\mathcal{X}_L$ (represented by P_M) could contain non-zero weights. So we consider P_M for simplicity.


Funny little dog breed pug dressed red christmas hat.
A Funny dog
Funny dog dressed as a wizard king.
A Funny dog
Two funny cute brown dog characters.

(a) SimCLR (b) CLIP

Figure 1. Illustration of raw and augmented samples generated by SimCLR and CLIP on the CC12M dataset (Changpinyo et al., 2021), where the former are generated by manual data augmentations and the latter are induced by visual-language pairs.

4. Formal Comparison between Multi-modal and Self-Supervised Contrastive Learning

In Section 3, we have established a theoretical framework for analyzing multi-modal contrastive learning (MMCL) from the perspective of asymmetric matrix factorization. Meanwhile, we know that MMCL originates from self-supervised contrastive learning (SSCL) like SimCLR (Chen et al., 2020) and MoCo (He et al., 2020), which is self-supervised (usually visual). These two contrastive learning paradigms have a close resemblance by both adopting InfoNCE-like objectives, while they differ mainly on the chosen positive and negative pairs. Take two representative methods in each paradigm, CLIP (MMCL) and SimCLR (SSCL), as an example. CLIP adopts visual-language pairs collected from the Internet, while SimCLR generates positive pairs by visual data augmentations like cropping and color jittering. Despite the similarity in learning objectives, CLIP shows much better performance on zero-shot and few-shot transfer learning tasks than SimCLR (Radford et al., 2021), suggesting that different sources of positive pairs have a crucial impact on the downstream performance of contrastive learning. Nevertheless, there still lack theoretical understanding and characterization of this phenomenon.

In this section, we propose a unified theoretical framework to understand the inherent connections between the two paradigms (Section 4.1). Based on this unified perspective, we compare CLIP and SimCLR on real-world data to understand their differences in downstream tasks (Section 4.1). At last, we theoretically analyze the differences from a data generation perspective (Section 4.2).

4.1. Unified Formulation and Analysis for Multi-modal and Self-Supervised Contrastive Learning

We begin with a brief introduction to self-supervised contrastive learning. Instead of using raw images $x_v \in \mathcal{X}_V$ as in multi-modal contrastive learning, self-supervised contrastive learning like SimCLR (Chen et al., 2020) applies aggressive data augmentation $\mathcal{A}(\cdot|x_v)$ two times and get a pair of augmented samples $x_a, x_a^+ \in \mathcal{X}_A$ as positive pairs to align together. Accordingly, the negative sample is defined as augmented samples x_a^- independently drawn from its marginal distribution. The self-supervised spectral contrastive loss (HaoChen et al., 2021) learns a Siamese visual encoder $f_V : \mathcal{X}_A \to \mathbb{R}^k$ with

$$\begin{aligned}\mathcal{L}^{\text{ss}}_{\text{SCL}}(f_V) = &-2\mathbb{E}_{x_a,x_a^+} f_V(x_a)^\top f_V(x_a^+)\\ &+\mathbb{E}_{x_a,x_a^-}(f_V(x_a)^\top f_V(x_a^-))^2,\end{aligned} \tag{12}$$

where the joint distribution of positive pairs follows

$$\mathcal{P}_A(x_a, x_a^+) = \mathbb{E}_{x_v \sim \mathcal{P}_V} \mathcal{A}(x_a|x_v)\mathcal{A}(x'_a|x_v), \tag{13}$$

which is marginalized over the augmentations of all natural samples. Different from multi-modal learning, the joint distribution is symmetric, *i.e.,* $\mathcal{P}_A(x_a, x_a^+) = \mathcal{P}_A(x_a^+, x_a)$, and HaoChen et al. (2021) show that this self-supervised loss is equivalent to a symmetric matrix factorization (SMF) objective. Nevertheless, there is a noticeable difference between the multi-modal and self-supervised objectives, that the joint distribution $\mathcal{P}_M$ defines connections between two domains $\mathcal{X}_V, \mathcal{X}_L$ while $\mathcal{P}_A$ defines connections only among visual samples in $\mathcal{X}_A$. It thus remains unclear to us how to compare the quality of multi-modal and self-supervised pairs and characterize their influence on downstream tasks.

A key insight here: we notice that CLIP does not only work well for multi-modal tasks like image-text retrieval, but also performs surprisingly well on visual-only tasks like zero-shot image classification, which indicates that it also implicitly aligns semantically similar visual samples together during the joint embedding process. The following theorem characterizes this intuition by establishing an equivalence between multi-modal contrastive learning and a corresponding self-supervised contrastive learning objective among visual-only samples.

Theorem 4.1. *The optimal visual representations of multi-modal contrastive learning (Eq. 9a) are equivalent (up to scaling and rotation) to that of the following uni-modal contrastive learning objective,*

$$\begin{aligned}\mathcal{L}^{\text{uni}}_{\text{SCL}}(f_V) = &-2\mathbb{E}_{x_v,x_v^+} f_V(x_v)^\top f_V(x_v^+)\\ &+\mathbb{E}_{x_v,x_v^-}(f_V(x_v)^\top f_V(x_v^-))^2,\end{aligned} \tag{14}$$

Table 1. Comparison (in the uni-model setting) of estimated labeling error and intra-class connectivity between CLIP and SimCLR.

	CLIP	SimCLR
Labeling Error (↓)	**0.601**	0.846
Intra-class Connectivity (↑)	**1.322**	1.072

where (x_v, x_v^+) are drawn from the text-induced joint distribution over visual samples $\mathcal{P}_T$ that $\forall\, x_v, x_v' \in \mathcal{X}_V$,

$$\mathcal{P}_T(x_v, x_v') = \mathbb{E}_{x_l \sim \mathcal{P}_L} \mathcal{P}_M(x_v|x_l)\mathcal{P}_M(x_v'|x_l), \qquad (15)$$

with $\mathcal{P}_M(x_v|x_l) = \mathcal{P}_M(x_v, x_l)/\mathcal{P}_L(x_l)$, and x_v^- is independently drawn from $\mathcal{P}_V$. Accordingly, the linear probing error $\mathcal{E}(f_V^*)$ of multi-modal learning is also equal to that of the self-supervised learning in Eq. 14.

Theorem 4.1 draws an inherent connection between multi-modal contrastive learning (MMCL) and self-supervised contrastive learning (SSCL) by showing that MMCL also implicitly performs uni-modal contrastive learning among visual samples, just like SSCL. Notably, different from SSCL that relies on manual data augmentations $\mathcal{A}(x_a|x_v)$, MMCL's uni-modal objective (Eq. 14) leverages the multi-modal conditional distribution $\mathcal{P}_M(x_v|x_l)$ to generate positive visual pairs *via languages as a pivot*. In other words, the multi-modal signals serve as a new type of data augmentation such that image pairs x_v, x_v^+ with the same (or similar) text descriptions can serve as positive pairs for uni-modal contrastive learning, as illustrated in Figure 1(b).

This unified perspective enables us to understand the advantage of CLIP over SimCLR for visual representation learning (Radford et al., 2021). Intuitively, compared to SimCLR relying on object-agnostic and low-level manual data augmentations, *e.g.,* color and contrast variation in Figure 1(a), text descriptions contain high-level semantics of images (*e.g.,* "funny", "dog" in Figure 1(b)), and the use of the text-induced augmentation in CLIP can bridge semantically similar images more effectively. Thus, CLIP has two main advantages over SimCLR for downstream tasks according to Theorem 3.3. First, CLIP has a lower labeling error because the text-induced positive pairs usually contain the same object and while manual data augmentations often lose the object. Second, CLIP yields better connectivity among visual samples using high-level semantics. In the following, we provide empirical and theoretical comparisons to characterize the differences between them.

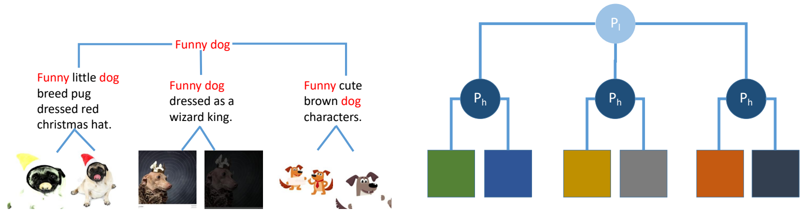
Based on the unified theoretical understanding above, we further investigate the differences between the augmentation-induced joint distribution $\mathcal{P}_A$ (self-supervised, SimCLR) and the text-induced one $\mathcal{P}_T$ (multi-modal, CLIP) on real-world data. For a fair comparison, we pretrain the same backbone ViT-B (Dosovitskiy et al., 2021) on the same dataset, YFCC15M (Thomee et al., 2016; Radford et al., 2021), and evaluate the learned representations on ImageNet (Deng et al., 2009a). For efficiency, we randomly draw 1,000 samples from 10 random classes of the ImageNet validation set. According to the matrix factorization perspective, the learned features approximate the ground-truth distribution (unknown to us). Thus, we can approximately calculate the (uni-modal) labeling error and sample connectivity using learned representations. For an intuitive measure of the desired sample connectivity, we calculate the average feature similarity between intra-class samples as a surrogate metric. See details in Appendix A.1.

From Table 1, we observe that the labeling error of SimCLR is indeed much larger than that of CLIP (0.846 *v.s.* 0.601), suggesting that the text-induced (implicit) positive images have higher semantic consistency than manual image transformations. Meanwhile, we also observe that CLIP has high intra-class connectivity than SimCLR (1.322 *v.s.* 1.072), suggesting that text descriptions can induce better intra-class sample diversity with the high-level semantic relationship.

4.2. A Data Generation Perspective via the Language of Hierarchical Random Graph

As discussed above, the key difference between augmentation and text-induced positive pairs is that they operate on different levels of semantics. This difference can be understood and modeled in a hierarchical structure of data generation. As shown in the examples in Figure 2(a), we can regard that the three images of funny dogs are firstly generated under high-level concepts captured by their text description, and then adding more detailed variations that can be captured by data augmentations. Therefore, the shared text span "funny dog" can draw these images together, but the commonly used data augmentations cannot because they are very different in pose and style.

Inspired by the observation that the joint distribution between positive visual pairs $\mathcal{P}_T(x_v, x_v')$ can be regarded as the adjacency matrix of a graph over all image samples (HaoChen et al., 2021), we model this distribution (graph) with hierarchical random graph (Clauset et al., 2008) designed to model the hidden structure of a given graph. Different from vanilla random graph where each edge is randomly drawn with the same probability, hierarchical random graph assumes that the edges are drawn according to a hierarchical tree, which suits our need to characterize different levels of semantics. In a hierarchical random graph $\mathcal{G}$ shown in Figure 2(b), each internal node s is associated with a probability p_s, each leaf node is a node in the original graph, and the probability of having an edge between two nodes is the probability contained in their lowest common ancestor node. In our case, we assume two hidden layers for simplicity, with p_l modelling the probability high-level connection in


Funny dog
Funny little dog breed pug dressed red christmas hat.
Funny dog dressed as a wizard king.
Funny cute brown dog characters.
P$_l$
P$_h$
P$_h$
P$_h$

(a) an illustration of the hierarchical structure

(b) a hierarchical random graph

Figure 2. Illustrations of the hierarchical structure on real-world datasets CC12M and a hierarchical random graph with two hidden layers. Here, each internal node is associated with a probability that a pair of vertices in the left and right subtrees of that node are connected.

the first layer and p_h modeling the probability of lower-level connection in the second layer. We assume $p_h > p_l$ as there are less high-level interactions between samples.

The following theorem shows that a larger high-level connection probability p_l yields better downstream performance by inducing better graph connectivity (algebraically measured by the singular value σ_t).

Theorem 4.2. *For two three-layer hierarchical random graphs $\mathcal{G}$, $\mathcal{G}'$ with probabilities $(p_l, p_h), (p'_l, p'_h)$, respectively. If $p_h - p_l \leq p'_h - p'_l$, we have*

$$\sigma_t \leq \sigma'_t,$$

where the σ_t, σ'_t are the t-th largest singular values of $\mathcal{G}, \mathcal{G}'$, respectively. According to Theorem 3.3, smaller singular value indicates better downstream performance under the same labeling error α. Therefore, contrastive learning with samples generated according to graph $\mathcal{G}'$ will have better downstream performance.

Theorem 4.2 shows smaller $p_h - p_l$ can bring better downstream generalization[3]. In practice, we can improve p_h by generating positive samples sharing common high-level semantics, as done in CLIP with the text description of the image. Therefore, our hierarchical random graph perspective can help characterize the benefit of CLIP over SimCLR from the kind of information they leverage. This perspective also suggests a way to improve (self-supervised) contrastive learning, that is to add more diverse with better augmentation strategies, such as, using realistic generative models like diffusion models (Ho et al., 2020). In the next section, we provide empirical verification of this understanding by showing how MMCL information can be used to boost augmentation-based SSCL methods like SimCLR.

5. Boosting SimCLR with Guided Positive Selection

Learning from the theoretical and empirical evidence in Section 4, we have known that compared to self-supervision, languages are better at generating positive pairs for visual representation learning due to their advantage of capturing high-level similarities. In this section, we further leverage this advantage to improve self-supervised learning.

Prior to ours, there are several papers exploring the combination of self-supervision and multi-modal supervision, such as SLIP (Mu et al., 2022), DeCLIP (Li et al., 2022b), and FLIP (Li et al., 2022a). Contrary to these methods all focusing on pretraining on multi-modal data, in this work, we focus on utilizing the estimated multi-modal information in a pretrained CLIP model to improve self-supervised contrastive learning (SimCLR) from unlabeled images alone, which, up to our knowledge, is not considered yet. Our experiment is designed as a verification of our analysis above, because if the language information is as helpful for uni-modal contrastive learning as we suppose, the CLIP-assisted SimCLR can obtain better performance on downstream tasks.

5.1. Methods

Following our analysis, we consider four strategies for leveraging CLIP to help self-supervised contrastive learning with SimCLR.

AddNewPositive & **DropFalsePositive**. Because multi-modal contrastive learning is good at generating more diverse and consistent positive pairs (Figure 1(b)), we leverage the pretrained CLIP to generate a new pair of positive samples for training SimCLR. Specifically, in a mini-batch, we find the nearest neighbor of each sample x in the feature space of CLIP, denoted as $\mathcal{N}(x)$, and regard $(x, \mathcal{N}(x))$ as a pair of positive samples. We mix this new positive pair with the original self-supervised one with a tunable ratio. On the other hand, because multi-modal pairs have less labeling error (Table 1), CLIP can also be leveraged to filter out

[3] We note that two quantities p_h, p_l are not independent. Since the total probability sums to one, a higher p_h means a lower p_l, and vice versa.

Table 2. The linear probing accuracy of SimCLR and its CLIP-assisted variants on ImageNet (ViT-B, 100-epoch training).

Method	Baseline (SimCLR)	AddNewPositive	DropFalsePositive	DropFalseNegative	DropEasyNegative
Linear Acc	61.2	**67.4 (+6.2)**	61.8 (+0.6)	61.4 (+0.2)	62.3 (+1.1)

false positive pairs that may contain different objects with a tunable ratio.

DropFalseNegative & **DropEasyNegative**. We can also leverage CLIP to select negative samples. One option is to drop negatives with the largest similarity, which could be false negatives from the same class of positive samples. Another is to drop negative samples with the smallest similarity, with corresponds to easy negative samples that are already pushed apart.

For the CLIP model, we adopt the pretrained ViT-B provided by the official implementation. For SimCLR, following the standard protocol, we pretrain a ResNet-50 (He et al., 2016) on ImageNet for 100 epochs. See details in Appendix A.2.

5.2. Results

From Table 2, we can see that all four techniques can bring benefits over the vanilla SimCLR, suggesting that the multi-modal information in CLIP indeed benefits self-supervised learning in terms of both positive and negative sample selection. Meanwhile, comparing the four strategies, we notice that AddNewPositive with CLIP brings the highest improvement of 6.2% accuracy over the vanilla SimCLR. This successfully verifies our previous analysis that multi-modal learning is better at generating diverse and positive samples than self-supervised learning for better downstream performance. We leave more advanced techniques for leveraging this observation for future work.

6. Conclusion

In this paper, we proposed the first theoretical framework for multi-modal contrastive learning. By drawing the connection to asymmetric matrix factorization, we characterized its optimal representations and established the first guarantees on the downstream generalization of multi-modal contrastive learning. Based on our framework, we provided a unified perspective of multi-modal and self-supervised contrastive learning, characterized their differences on real-world data, and verified our insights by bringing benefits on benchmark datasets. In this way, our theory has established a principled understanding of multi-modal contrastive learning, while delivering practical insights for combining multi-modal and self-supervised learning methods.

Acknowledgement

Yisen Wang is partially supported by the National Key R&D Program of China (2022ZD0160304), the National Natural Science Foundation of China (62006153), Open Research Projects of Zhejiang Lab (No. 2022RC0AB05), and Huawei Technologies Inc.

References

Changpinyo, S., Sharma, P., Ding, N., and Soricut, R. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

Chung, F. R. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

Clauset, A., Moore, C., and Newman, M. E. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009a.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009b.

Ding, C., He, X., and Simon, H. D. On the equivalence of nonnegative matrix factorization and spectral clustering. In *SDM*, 2005.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Eckart, C. and Young, G. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *NeurIPS*, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, 2016.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Huang, Y., Du, C., Xue, Z., Chen, X., Zhao, H., and Huang, L. What makes multi-modal learning better than single (provably). In *NeurIPS*, 2021.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Li, Y., Fan, H., Hu, R., Feichtenhofer, C., and He, K. Scaling language-image pre-training via masking. *arXiv preprint arXiv:2212.00794*, 2022a.

Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., and Yan, J. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *ICLR*, 2022b.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019.

Mu, N., Kirillov, A., Wagner, D., and Xie, S. Slip: Self-supervision meets language-image pre-training. In *ECCV*, 2022.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *EMNLP*, 2014.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Santurkar, S., Dubois, Y., Taori, R., Liang, P., and Hashimoto, T. Is a caption worth a thousand images? a controlled study for representation learning. *arXiv preprint arXiv:2207.07635*, 2022.

Saunshi, N., Plevrakis, O., Arora, S., Khodak, M., and Khandeparkar, H. A theoretical analysis of contrastive unsupervised representation learning. In *ICML*, 2019.

Saunshi, N., Ash, J., Goel, S., Misra, D., Zhang, C., Arora, S., Kakade, S., and Krishnamurthy, A. Understanding contrastive learning requires incorporating inductive biases. In *ICML*, 2022.

Sun, X., Xu, Y., Cao, P., Kong, Y., Hu, L., Zhang, S., and Wang, Y. Tcgm: An information-theoretic framework for semi-supervised multi-modality learning. In *ECCV*, 2020.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

Tschannen, M., Mustafa, B., and Houlsby, N. Image-and-language understanding from pixels only. *arXiv preprint arXiv:2212.08045*, 2022.

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020.

Wang, Y., Geng, Z., Jiang, F., Li, C., Wang, Y., Yang, J., and Lin, Z. Residual relaxation for multi-view representation learning. In *NeurIPS*, 2021.

Wang, Y., Zhang, Q., Wang, Y., Yang, J., and Lin, Z. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. In *ICLR*, 2022.

Wang, Y., Zhang, Q., Du, T., Yang, J., Lin, Z., and Wang, Y. A message passing perspective on learning dynamics of contrastive learning. In *ICLR*, 2023.

Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., and Xu, C. Filip: Fine-grained interactive language-image pre-training. In *ICLR*, 2022.

Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., and Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020.

A. Experimental Details

A.1. Details of Empirical Comparison in Section 4.1

Approximation of Data Probability. Similar to the multi-modal spectral loss, Equation (14) can be rewritten as a matrix decomposition loss, i.e., $\mathcal{L}_{\mathrm{SCL}}^{\mathrm{uni}}(f_V) = \|\tilde{P}_T - F_V F_V^\top\|^2 + const$, where P_T is the co-occurrence matrix of the distribution $\mathcal{P}_T(x_v, x_v')$, $(\tilde{P}_T)_{(x_v,x_v')} = \frac{\mathcal{P}_T(x_v,x_v')}{\sqrt{\mathcal{P}_V(x_v)\mathcal{P}_V(x_v')}}$ and $(F_V F_V^\top)_{(x_v,x_v')} = \frac{f_V(x_v)^\top f_V(x_v')}{\sqrt{P_V(x_v)P_V(x_v')}}$. So $(P_T)_{(x_v,x_v')}$ can be approximated by $f_V(x_v)^\top f_V(x_v')$ when the loss is minimized. Similarly, we can estimate the co-occurrence matrix $(P_A)_{(x_a,x_a^+)}$ of $\mathcal{P}_A(x_a, x_a^+)$ by $f_V(x_a)^\top f_V(x_a^+)$. In practice, we use ViT-Base trained by CLIP (Radford et al., 2021) and SimCLR (Chen et al., 2020) as the encoders.

Setup. We respectively encode the samples from 1000 samples randomly selected from 10 classes of ImageNet (Deng et al., 2009b) with two encoders and construct the embedding matrix $\hat{F}_T \in \mathbb{R}^{1000\times k}$ and $\hat{F}_A \in \mathbb{R}^{1000\times k}$ (k is the output dimension of ViT-Base)[4]. Then we normalize the similarity matrices of the embeddings to and estimate the co-occurrence matrices with them, i.e., $\hat{P}_T = \mathrm{normalize}(\hat{F}_T\hat{F}_T^\top)$, $\hat{P}_A = \mathrm{normalize}(\hat{F}_A\hat{F}_A^\top)$. In the next step, we evaluate the properties of the estimated matrices, e.g., the labeling error, the eigenvalues, etc.

Estimation of Labeling Error. When evaluating the labeling error α in Eq. 11, as ImageNet is a vision dataset, we have no access to the corresponding text data. So we use a surrogate metric α_T, and it is defined as:

$$\alpha_T = \sum_{x_v,x_v'} (P_T)_{x_v,x_v'} \mathbb{1}[y(x_v) \neq y(x_v')], \tag{16}$$

and $y(x_v)$ denotes the ground-truth label of x_v. Note that α_T is lower bounded by the ground-truth labeling error α:

Proposition A.1. *For the surrogate metric α_T, we have*

$$\alpha \geq \frac{1}{2}\alpha_T.$$

Proof. Expanding the estimated labeling error and we obtain

$$\begin{aligned}
\alpha_T &= \sum_{(x_v,x_v')} \mathcal{P}_T(x_v, x_v')\mathbb{1}[y(x_v) \neq y(x_v')] \\
&= \sum_{x_v,x_v'} \mathbb{E}_{x_l}\left[\mathcal{P}_M(x_v|x_l)\mathcal{P}_M(x_v|x_l)\mathbb{1}[y(x_v) \neq y(x_v')]\right] \\
&\leq \sum_{x_v,x_v'} \mathbb{E}_{x_l}\left[\mathcal{P}_M(x_v|x_l)\mathcal{P}_M(x_v|x_l)(\mathbb{1}[y(x_v) \neq y(x_l)] + \mathbb{1}[y(x_v') \neq y(x_l)])\right] \\
&= 2\mathbb{E}_{x_l}[\mathcal{P}_M(x_v|x_l)\mathbb{1}[y(x_v) \neq y(x_l)]] \\
&= 2\mathbb{E}_{x_v,x_l}\mathbb{1}[y(x_v) \neq y(x_l)] \\
&= 2\alpha.
\end{aligned}$$

□

As a result, a large α_T implies a large labeling error α. Then we replace P_T with $\hat{P}_T$, and obtain the estimation $\hat{\alpha}_T = \sum_{x_v,x_v'} (\hat{P}_T)_{x_v,x_v'}\mathbb{1}[y(x_v) \neq y(x_v')]$. Similarly, we define the estimated labeling error of P_A as $\hat{\alpha}_A = \sum_{x_v,x_v^+} (\hat{P}_A)_{x_v,x_v^+}\mathbb{1}[y(x_v) \neq y(x_v^+)]$.

Estimation of Intra-class Connectivity. When evaluating the intra-class connectivity, we respectively select 1000 samples from 10 different classes of ImageNet. Taking the multi-modal pretraining as an example, following the process we construct

[4] As the samples of the P_A are augmented images, we transform the selected samples with the augmentations used in SimCLR when constructing $\hat{F}_A$.

$\hat{P}_T$, we respectively construct ten intra-class feature similarity matrices $\{\hat{P}_{in}^k\}_{k=1}^{10}$. Then we randomly select 1000 samples from the selected samples and construct an inter-class feature similarity matrix $\hat{P}_{out}$. We use the average relative value of the intra-class and inter-class feature similarity matrix to represent the intra-class connectivity. To be specific, we denote the intra-class connectivity as β and evaluate it by:

$$\begin{aligned}(\hat{P}_{re})_{i,j}^k &= (\hat{P}_{in})_{i,j}^k / \underset{i,j}{\operatorname{mean}}(\hat{P}_{out}),\\ \beta_k &= \underset{i,j}{\operatorname{mean}}(\hat{P}_{re})_{i,j}^k,\\ \beta &= \underset{k}{\operatorname{mean}}(\beta_k).\end{aligned} \tag{17}$$

A.2. Details of Verification Experiments in Section 5

We use SimCLR (Chen et al., 2020) as our baseline and adopt the popular backbone ResNet-50. With the default setting of SimCLR, we add a projector MLP following the backbone. During the pretraining process of SimCLR, we train the encoder for 100 epochs on ImageNet with 512 batch size and use the LARS optimizer with a cosine annealed learning rate schedule. When estimating the co-occurrence matrix P_T, we compute the feature similarity matrix with the well-trained ViT-B encoder provided by the official repository of CLIP (Radford et al., 2021). For selecting new positive pairs, we set the ratio between the new regularizer and the original loss to 1. When filtering false positive samples, we throw the 10% positive pairs that are most dissimilar in the feature space encoded by the CLIP encoder. And for selecting better negative samples. we respectively throw 5% samples that have the largest similarity with the positive samples and 10% samples that have the smallest similarity with the positive samples. After the pretraining process, we train a linear classifier following the frozen backbones and optimize the CrossEntropy loss with the SGD optimizer.

B. Proofs

B.1. Proof of Theorem 3.1

Proof. Expanding the decomposition object $\mathcal{L}_{\mathrm{AMF}}$ and we obtain,

$$\begin{aligned}\mathcal{L}_{\mathrm{AMF}}(F_V, F_L) &= \|\tilde{P}_M - F_V F_L^\top\|^2\\ &= \sum_{x_v, x_l} \left((\tilde{P}_M)_{x_v, x_l} - (F_V)_{x_v} (F_L)_{x_l}^\top\right)^2\\ &= \sum_{x_v, x_l} \left(\frac{\mathcal{P}_M(x_v, x_l)}{\sqrt{\mathcal{P}_V(x_v)\mathcal{P}_L(x_l)}} - \sqrt{\mathcal{P}_V(x_v)} f_V(x_v)^\top \sqrt{\mathcal{P}_L(x_l)} f_L(x_l)\right)^2\\ &= \sum_{x_v, x_l} \left(\frac{\mathcal{P}_M(x_v, x_l)^2}{\mathcal{P}_V(x_v)\mathcal{P}_L(x_l)} + \mathcal{P}_V(x_v)\mathcal{P}_L(x_l)\left(f_V(x_v)^\top f_L(x_L)\right)^2 - 2\mathcal{P}_M(x_v, x_l) f_V(x_v)^\top f_L(x_L)\right)\\ &= \sum_{x_v, x_l} \left(\frac{\mathcal{P}_M(x_v, x_l)^2}{\mathcal{P}_V(x_v)\mathcal{P}_L(x_l)}\right) - 2\mathbb{E}_{x_v, x_l} f_V(x_v)^\top f_L(x_l) + \mathbb{E}_{x_v^-, x_l^-}\left(f_V(x_v^-)^\top f_L(x_l^-)\right)^2\\ &= \mathcal{L}_{\mathrm{SCL}}(f_V, f_L) + const.\end{aligned}$$

□

B.2. Proof of Theorem 3.2

Proof. According to Eckart-Young Theorem (Eckart & Young, 1936), the optimal solution $F_V^\star, F_L^\star$ of the decomposition objective $\mathcal{L}_{\mathrm{AMF}}(F_V, F_L) = \|\tilde{P}_M - F_V F_L^\top\|^2$ satisfy:

$$F_V^\star (F_L^\star)^\top = U^k \operatorname{diag}(\sigma_1, ..., \sigma_k)(V^k)^\top,$$

where we denote $\tilde{P}_M = U\Sigma V^\top$ as the singular value decomposition of $\tilde{P}_M$, $(\sigma_1, ..., \sigma_k)$ are the k-largest singular values of $\tilde{P}_M$, the t-th column of $U^k \in \mathbb{R}^{N_V \times k}$ contains the corresponding eigenvectors of the t-th largest singular values and

$V^k \in \mathbb{R}^{N_L \times k}$ is a unitary matrix. Then we respectively represent the optimal solutions $F_V^\star$ and $F_L^\star$:

$$F_V^\star = U^k DR,$$
$$F_L^\star = V^k \operatorname{diag}(\sigma_1, ..., \sigma_k) D^{-1} R,$$

where $R \in \mathbb{R}^{k \times k}$ is a unitary matrix and D is an invertible diagonal matrix. With $(F_V)_{x_v} = (f_V(x_v))^\top \sqrt{P_V(x_v)}$ and $(F_L)_{x_l} = (f_L(x_l))^\top \sqrt{P_L(x_l)}$, we obtain

$$f_V^*(x_v) = \frac{1}{\sqrt{\mathcal{P}_V(x_v)}} (U_{x_v}^k DR)^\top, \tag{18}$$

$$f_L^*(x_l) = \frac{1}{\sqrt{\mathcal{P}_L(x_l)}} (V_{x_l}^k \operatorname{diag}(\sigma_1, \dots, \sigma_k) D^{-1} R)^\top. \tag{19}$$

□

B.3. Proof of Theorem 3.3

We first introduce a lemma in HaoChen et al. (2021):

Lemma B.1 (Theorem 3.8 in HaoChen et al. (2021))**.** *Denote the labeling error as $\alpha = \mathbb{E}_{(x_v, x_l)} \mathbb{1}[y(x_v) \neq y(x_l)]$. Let $f_V'^\star$ be a minimizer of the $\mathcal{L}_{\mathrm{SCL}}^{\mathrm{uni}}(f_V)$, we obtain*

$$\mathcal{E}(f_V'^\star) \leq \frac{2\phi^y}{\sigma'_{k+1}} + 8\alpha,$$

where σ'_{k+1} is the k-smallest eigenvalue of the Laplacian matrix of P_T.

Then we give the proof of Theorem 3.3 in the following.

Proof. We denote $y(x)$ as the label of data x. Then we define the probability that two image samples related to the same text sample have different labels as

$$\phi^y = \sum_{x_v, x'_v} \mathcal{P}_T(x_v, x'_v) \mathbb{1}[y(x_v) \neq y(x'_v)]. \tag{20}$$

We note that

$$\begin{aligned}
\phi^y &= \sum_{(x_v, x'_v)} \mathcal{P}_T(x_v, x'_v) \mathbb{1}[y(x_v) \neq y(x'_v)] \\
&= \sum_{x_v, x'_v} \mathbb{E}_{x_l} \left[\mathcal{P}_M(x_v|x_l) \mathcal{P}_M(x_v|x_l) \mathbb{1}[y(x_v) \neq y(x'_v)] \right] \\
&\leq \sum_{x_v, x'_v} \mathbb{E}_{x_l} \left[\mathcal{P}_M(x_v|x_l) \mathcal{P}_M(x_v|x_l) (\mathbb{1}[y(x_v) \neq y(x_l)] + \mathbb{1}[y(x'_v) \neq y(x_l)]) \right] \\
&= 2\mathbb{E}_{x_l} [\mathcal{P}_M(x_v|x_l) \mathbb{1}[y(x_v) \neq y(x_l)]] \\
&= 2\mathbb{E}_{x_v, x_l} \mathbb{1}[y(x_v) \neq y(x_l)] \\
&= 2\alpha.
\end{aligned}$$

Combined with Lemma B.1, we have $\mathcal{E}(f_V'^\star) \leq \widetilde{O}(\frac{\alpha}{\sigma'_{k+1}})$, where $\widetilde{O}(\cdot)$ is used to hide universal constant factors. We denote the $(k+1)$-largest singular values of $\tilde{P}_M$ as σ_{k+1}. As $\tilde{P}_T = \tilde{P}_M \tilde{P}_M^\top$ and the singular values are positive, the $(k+1)$-largest singular values of $\tilde{P}_T$ is $(\sigma_{k+1})^2$, i.e., $\sigma'_{k+1} = 1 - (\sigma_{k+1}^2)$. Combined with Theorem 4.1 (proofs are provided in the following), for the image encoder $f_V^\star$ that minimizes $\mathcal{L}_{\mathrm{SCL}}$, we obtain

$$\mathcal{E}(f_V^\star) = \mathcal{E}(f_V'^\star) \leq \widetilde{O}(\frac{\alpha}{1 - \sigma_{k+1}^2}). \tag{21}$$

Obviously, the linear probing error of the text encoder $f_L^\star$ that minimizes $\mathcal{L}_{\mathrm{SCL}}$ has the similar results:

$$\mathcal{E}(f_L^\star) \le \widetilde{O}(\frac{\alpha}{1-\sigma_{k+1}^2}). \tag{22}$$

Then we consider the empirical loss with finite samples. We construct a multi-modal dataset $\hat{\mathcal{X}} = \{(z_v^1, z_l^1), ..., (z_v^n, z_l^n)\}$ and the n positive pairs are i.i.d sampled from $\mathcal{P}_M(x_v, x_l)$. We first sample a permutation $\pi : [n] \to [n]$, then we construct the positive pairs and negative pairs as follows:

$$\begin{aligned} x_v^i &= z_v^{\pi(3i-2)}, \\ x_l^i &= z_l^{\pi(3i-2)}, \\ (x_l^i)^- &= z_l^{\pi(3i-1)}, \\ (x_v^i)^- &= z_v^{\pi(3i)}. \end{aligned}$$

and the empirical loss is

$$\mathcal{L}_{\mathrm{emp}}(f_V, f_L) \quad = -\frac{2}{n/3}\sum_{i=1}^{n/3} f_V(x_v^i)^\top f_L(x_l^i) + \frac{1}{n/3}\sum_{i=1}^{n/3}(f_V(x_v^i)^\top f_L\left(x_l^i)^-\right)^2 + \frac{1}{n/3}\sum_{i=1}^{n/3}(f_V((x_v^i)^-)^\top f_L(x_l^i))^2. \tag{23}$$

Considering the expectation of $\mathcal{L}_{\mathrm{emp}}$, we obtain

$$\begin{aligned} \mathbb{E}_{\hat{\mathcal{X}}}\mathcal{L}_{\mathrm{emp}}(f_V, f_L) &= -\frac{2}{n/3}\sum_{i=1}^{n/3} f_V(x_v^i)^\top f_L(x_l^i) + \frac{1}{n/3}\sum_{i=1}^{n/3}(f_V(x_v^i)^\top f_L((x_l^i)^-)^2 + \frac{1}{n/3}\sum_{i=1}^{n/3}(f_V((x_v^i)^-)^\top f_L(x_l^i))^2. \\ &= -2\mathbb{E}_{x_v, x_l} f_V(x_v)^\top f_L(x_l) + \mathbb{E}_{x_v\sim\mathcal{P}_V(x_v), x_l\sim\mathcal{P}_L(x_l)}(f_V(x_v^i)^\top f_L(x_l^j))^2 \\ &= \mathcal{L}_{\mathrm{SCL}}(f_V, f_L). \end{aligned}$$

So the empirical loss is an unbiased estimator. We denote that Rademacher complexity of $\mathcal{F}$ over n data as

$$\hat{\mathcal{R}}_n(\mathcal{F}) = \max_{\{x_1, \dots x_n\}} \mathbb{E}_\sigma \left[\sup_{f\in\mathcal{F}, i} \left(\frac{1}{n}\sum_{j=1}^{n} \rho_j f_i(x_j) \right) \right],$$

where $f_i(x_j)$ denotes the i-th dimension of $f(x_j)$ and ρ is a uniform random vector in $\{-1, 1\}^n$.

Following Theorem 4.2 in HaoChen et al. (2021), when $\mathcal{E}(\hat{f}_V^*)$, $\mathcal{E}(\hat{f}_V^*)$ are the minimizers of $\mathcal{L}_{\mathrm{emp}}(f_V, f_L)$, we obtain

$$\{\mathcal{E}(\hat{f}_V^*), \mathcal{E}(\hat{f}_L^*)\} \lesssim \frac{\alpha}{1-\sigma_{k+1}^2} + \underbrace{\frac{ck}{\Delta_\sigma^2}\left(\widehat{\mathcal{R}}_{n/3}(\mathcal{F}) + \sqrt{\frac{\log 2/\delta}{2n/3}} + \delta\right)}_{\text{finite-sample generalization terms}} \tag{24}$$

where $\lesssim$ omits some constant terms, σ_{k+1} (c.f. Theorem 3.2) is the $(k+1)$-th largest singular value of the normalized co-occurrence matrix $\tilde{P}_M$. In the finite-sample generalization terms, $\hat{\mathcal{R}}_{n/3}(\mathcal{F})$ denotes a Rademacher complexity of the model class $\mathcal{F}$ with $n/3$ samples, k is the representation dimension, $\Delta_\sigma = \sigma_{\lfloor 3k/4 \rfloor}^2 - \sigma_k^2$, and $c \lesssim (k\kappa + 2k\kappa^2 + 1)^2$ with κ upper bounding $\|f_V(x)\|_\infty$ and $\|f_L(x)\|_\infty$.

□

B.4. Proof of Theorem 4.1

We first introduce a lemma that states that multiplying the embedding matrix by an invertible matrix on the right will not influence the linear probing error (HaoChen et al., 2021):

Lemma B.2 (Lemma 3.1 in HaoChen et al. (2021))**.** *For two learned embedding matrices F, $\widetilde{F}$, a diagonal matrix D and an invertible matrix Q, if $F = D\widetilde{F}Q$, they have the equal linear probing error, i.e.,*

$$\mathcal{E}(F) = \mathcal{E}(\widetilde{F}).$$

Then we give the proof of Theorem 4.1 in the following.

Proof. With Theorem 3.2, the optimal solutions $F_V^\star$, $F_L^\star$ of $\mathcal{L}_{\rm AMF}(F_V, F_L) = \|\tilde{P}_M - F_V F_L^\top\|^2$ can be respectively represented as:

$$F_V^\star = U^k D R,$$
$$F_L^\star = V^k D_2 R,$$

where $R \in \mathbb{R}^{k\times k}$ is a unitary matrix and D, D_2 are diagonal matrices that satisfy $D_2 = \mathrm{diag}(\sigma_1, ..., \sigma_k)D^{-1}$. Following the proof of theorem 3.1, the uni-modal contrastive loss is also equivalent to a matrix decomposition loss, i.e., $\mathcal{L}_{\rm SCL}^{\rm uni}(f_V) = \|\tilde{P}_T - F_V F_V^\top\|^2 + const$, where $(\tilde{P}_T)_{(x_v,x_v')} = \frac{\mathcal{P}_T(x_v,x_v')}{\sqrt{\mathcal{P}_V(x_v)\mathcal{P}_V(x_v')}}$ and $(F_V)_{x_v} = \frac{f_V(x_v)^\top}{\sqrt{P_V(x_v)}}$. Then we consider the objective $L_{\rm mf}(F_V) = \|\tilde{P}_T - F_V F_V^\top\|^2$. Similar to the asymmetric decomposition objective, the optimal solution can be represented as:

$$(F_V^\star)' = U_T^k D_T R_T,$$

where $U_T^k \in \mathbb{R}^{N_V\times k}$ contains k corresponding eigenvectors of k largest singular values of $\tilde{P}_T$, $D_T \in \mathbb{R}^{k\times k}$ is an invertible diagonal matrix and $R_T \in \mathbb{R}^{k\times k}$ is a unitary matrix. In the next step, we analyze the relationship between $\tilde{P}_M$ and $\tilde{P}_T$. Considering the (x_v, x_v')-th element of $\tilde{P}_M\tilde{P}_M^\top$, we have

$$\begin{aligned}
(\tilde{P}_M\tilde{P}_M^\top)_{x_v,x_v'} &= \sum_{x_l}(\tilde{P}_M)_{x_v,x_l}(\tilde{P}_M)_{x_v',x_l} \\
&= \sum_{x_l}\frac{\mathcal{P}_M(x_v,x_l)\mathcal{P}_M(x_v',x_l)}{\mathcal{P}_L(x_l)\sqrt{\mathcal{P}_V(x_v)\mathcal{P}_V(x_v')}} \\
&= \frac{1}{\sqrt{\mathcal{P}_V(x_v)\mathcal{P}_V(x_v')}}\sum_{x_l}\mathcal{P}_L(x_l)\mathcal{P}_M(x_v|x_l)\mathcal{P}_M(x_v'|x_l) \qquad (\mathcal{P}_M(x_v,x_l) = \mathcal{P}_M(x_v|x_l)\mathcal{P}_L(x_l)) \\
&= \frac{\mathbb{E}_{x_l}\mathcal{P}_M(x_v|x_l)\mathcal{P}_M(x_v'|x_l)}{\sqrt{\mathcal{P}_V(x_v)\mathcal{P}_V(x_v')}} \\
&= (\tilde{P}_T)_{x_v,x_v'}.
\end{aligned}$$

We know that $\tilde{P}_T = \tilde{P}_M\tilde{P}_M^\top$, so $\tilde{P}_T$ and $\tilde{P}_M$ share the same eigenvectors, i.e., $U^k = U_T^k$. As D, D_2, R, D_T, R_T are invertible matrices and the product of the invertible matrices is still invertible, we obtain

$$F_V^\star = (F_V^\star)'T,$$

where $T = (D_T)^{-1}(R_T)^{-1}DR$ is an invertible matrix. With Lemma B.2, we obtain

$$\mathcal{E}(f_V^\star) = \mathcal{E}(f_V'^\star),$$

where $(F_V^\star)_{x_v} = f_V^\star(x_v)^\top, (F_V^\star)'_{x_v} = f_V'^\star(x_v)^\top$. So Theorem 4.1 is proved. □

B.5. Proof of Theorem 4.2

Proof. The co-occurrence matrix of the three-layer hierarchical random graph is:

$$P = \begin{pmatrix} p_h & \cdots & p_h & p_l & \cdots & p_l & \cdots & p_l & \cdots & p_l \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ p_h & \cdots & p_h & p_l & \cdots & p_l & \cdots & p_l & \cdots & p_l \\ p_l & \cdots & p_l & p_h & \cdots & p_h & \cdots & p_l & \cdots & p_l \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ p_l & \cdots & p_l & p_h & \cdots & p_h & \cdots & p_l & \cdots & p_l \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ p_l & \cdots & \cdots & \cdots & \cdots & p_l & \cdots & p_h & \cdots & p_h \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ p_l & \cdots & \cdots & \cdots & \cdots & p_l & \cdots & p_h & \cdots & p_h \end{pmatrix}.$$

Then we consider the process of computing the eigenvalues:

$$|\sigma E - P| = \begin{vmatrix} \sigma - p_h & \cdots & -p_h & -p_l & \cdots & -p_l & \cdots & -p_l & \cdots & -p_l \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ -p_h & \cdots & \sigma - p_h & -p_l & \cdots & -p_l & \cdots & -p_l & \cdots & -p_l \\ -p_l & \cdots & -p_l & \sigma - p_h & \cdots & -p_h & \cdots & -p_l & \cdots & -p_l \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ -p_l & \cdots & -p_l & -p_h & \cdots & \sigma - p_h & \cdots & -p_l & \cdots & -p_l \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ -p_l & \cdots & \cdots & \cdots & \cdots & -p_l & \cdots & \sigma - p_h & \cdots & -p_h \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ -p_l & \cdots & \cdots & \cdots & \cdots & -p_l & \cdots & -p_h & \cdots & \sigma - p_h \end{vmatrix}.$$

We denote that the first layer has s_l branches and the second layer has s_h branches. Add every column to the first column:

$$\begin{vmatrix} \sigma - s_h * p_h - (s_l - 1) * s_h * p_l & \cdots & -p_h & -p_l & \cdots & -p_l & \cdots & -p_l & \cdots & -p_l \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \sigma - s_h * p_h - (s_l - 1) * s_h * p_l & \cdots & \sigma - p_h & -p_l & \cdots & -p_l & \cdots & -p_l & \cdots & -p_l \\ \sigma - s_h * p_h - (s_l - 1) * s_h * p_l & \cdots & -p_l & \sigma - p_h & \cdots & -p_h & \cdots & -p_l & \cdots & -p_l \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \sigma - s_h * p_h - (s_l - 1) * s_h * p_l & \cdots & -p_l & -p_h & \cdots & \sigma - p_h & \cdots & -p_l & \cdots & -p_l \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \sigma - s_h * p_h - (s_l - 1) * s_h * p_l & \cdots & \cdots & \cdots & \cdots & -p_l & \cdots & \sigma - p_h & \cdots & -p_h \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \sigma - s_h * p_h - (s_l - 1) * s_h * p_l & \cdots & \cdots & \cdots & \cdots & -p_l & \cdots & -p_h & \cdots & \sigma - p_h \end{vmatrix}.$$

For the i-row, if i is not divisible by s_h, then minus the row by $(i|s_h * s_h)$-row, and we obtain

$$\begin{vmatrix} \sigma - s_h * p_h - (s_l - 1) * s_h * p_l & \cdots & -p_h & -p_l & \cdots & -p_l & \cdots & -p_l & \cdots & -p_l \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \sigma & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \sigma - s_h * p_h - (s_l - 1) * s_h * p_l & \cdots & -p_l & \sigma - p_h & \cdots & -p_h & \cdots & -p_l & \cdots & -p_l \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & -\sigma & \cdots & \sigma & \cdots & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \sigma - s_h * p_h - (s_l - 1) * s_h * p_l & \cdots & \cdots & \cdots & \cdots & -p_l & \cdots & \sigma - p_h & \cdots & -p_h \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & \cdots & -\sigma & \cdots & \sigma \end{vmatrix}.$$

For the j-column that satisfies j is divisible by s_h and $0 < j \leq (s_l - 1) * (s_h)$, add $\{j + 1, \cdots, j + s_h\}$-columns, and for the j-column that satisfies j is divisible by s_h and $0 < j < (s_l - 1) * (s_h)$, minus $\{j + s_h + 1, \cdots, j + 2 * s_h\}$-columns to the j-column, then we have

$$\begin{vmatrix}
\sigma - s_h * p_h - (s_l - 1) * s_h * p_l & \cdots & -p_h & 0 & \cdots & -p_l & \cdots & -s_h * p_l & \cdots & -p_l \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & \sigma & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\
\sigma - s_h * p_h - (s_l - 1) * s_h * p_l & \cdots & -p_l & \sigma - s_h * (p_h - p_l) & \cdots & -p_h & \cdots & -s_h * p_l & \cdots & -p_l \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & 0 & 0 & \cdots & \sigma & \cdots & 0 & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\sigma - s_h * p_h - (s_l - 1) * s_h * p_l & \cdots & \cdots & \cdots & \cdots & -p_l & \cdots & \sigma - s_h * p_h & \cdots & -p_h \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & \cdots & \cdots & \cdots & 0 & \cdots & 0 & \cdots & \sigma
\end{vmatrix}.$$

When expanding the determinant, the i-row that satisfies i is not divisible by s_h only has one non-zero value σ in i column, so the det is equal to

$$\sigma^{(s_l - 1) * s_h} \begin{vmatrix}
\sigma - s_h * p_h - (s_l - 1) * s_h * p_l & 0 & \cdots & 0 & -s2 * p_l \\
\sigma - s_h * p_h - (s_l - 1) * s_h * p_l & \sigma - s2 * (p_h - p_l) & \cdots & 0 & -s2 * p_l \\
\sigma - s_h * p_h - (s_l - 1) * s_h * p_l & -\sigma + s2 * (p_h - p_l) & \cdots & 0 & -s2 * p_l \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & -\sigma + s2 * (p_h - p_l) & \sigma - s2 * p_h
\end{vmatrix}.$$

The form of the det is easy to expand and we obtain the results:

$$\sigma^{(s_l - 1) * s_h} * (\sigma - s_h * p_h - (s_l - 1) * s_h * p_l) * (\sigma - s_h * (p_h - p_l))^{s_h - 1}. \tag{25}$$

So the eigenvalues are

$$\begin{aligned}
&\sigma_1 = s_h * p_h + (s_l - 1) * s_h * p_l = \frac{1}{s_l * s_h}, \\
&\sigma_2 = \cdots = \sigma_{s_l} = s_h * (p_h - p_l), \\
&\sigma_{s_l + 1} = \cdots = \sigma_{s1 * s2} = 0.
\end{aligned}$$

where $\frac{1}{s_l * s_h}$ and 0 are constants. As the matrix is a real symmetric matrix, the eigenvalues are equal to the singular values. And the row sum of the matrix is a constant, so we can obtain the results of Theorem 4.2. □