When Contrastive Learning Meets Bayesian Modeling: Learning Multi-Modal Representation Alignments from Noisy Data Pairs

Anonymous ACL submission

Abstract

Contrastive learning (CL) stands as a leading paradigm for self-supervised representation learning, achieving state-of-the-art results in multi-modal learning. However, a notable drawback of standard CL is its lack of robustness in the face of noisy (misaligned) data pairs. For instance, not all negative samples are truly negative; within a mini-batch, there can be negative samples that are semantically similar to positive samples. This issue is prevalent in many web-sourced multimodal datasets like CC3M and YFCC, commonly used for CL, due to their inherently noisy nature during dataset crawling. Consequently, dataset noise could significantly undermine the efficacy of CL. On the other hand, Bayesian modeling is renowned for its inherent capability to handle data noise and uncertainty. Is it possible to merge the strengths of both approaches by incorporating Bayesian modeling into CL for noise-robust representation learning? In this paper, we propose a novel solution by reimagining standard CL within a probability framework and introducing learnable random weights to associate with data pairs. Our framework enables automatic inference of the level of noisiness for each data pair through efficient Bayesian sampling, based on a technique borrowed from Bayesian data augmentation. Importantly, our model can be effectively optimized using a novel learning algorithm based on stochastic expectation maximization. We demonstrate the efficacy of our approach on various standard multi-modal CL benchmarks, showcasing significant performance improvements over standard CL methods.

1 Introduction

Contrastive learning has gained increasing popularity in multi-modal representation learning due to its effectiveness in aligning representations from different modalities. In the realm of vision-language representation learning, the objective is to acquire generic representations from images and texts that could enhance multi-modal downstream applications, such as zero-shot image classification and image-text retrieval. Recent advancements (Jia et al., 2021a; Radford et al., 2021; Li et al., 2021; Zhou et al., 2022; Gao et al., 2023; Guo et al., 2023) have scaled up vision-language representation learning by leveraging contrastive loss to pre-train models with a substantial volume of web-sourced paired image-text data, such as Conceptual Caption (Sharma et al., 2018), YFCC (Thomee et al., 2016), and LAION (Schuhmann et al., 2022). While some studies amalgamate the representations of two modalities into a single encoder (Wang et al., 2021a,b, 2022b,c), it is more prevalent to represent the image and text modalities separately using modality-specific encoders, akin to the CLIP framework (Mokady et al., 2021; Shen et al., 2021; Yang et al., 2022; Shukor et al., 2022). Following pre-training, the model can generate general representations of both image and text inputs, showcasing outstanding performance in downstream tasks, such as text-guided generation of natural images (Ramesh et al., 2021; Crowson et al., 2022; Xu et al., 2023; Ruiz et al., 2023; Liu et al., 2023), videos (Kwon et al., 2022; Lin et al., 2022; Rasheed et al., 2023), 3D shapes (Sanghi et al., 2023; Wang et al., 2022a; Sanghi et al., 2022), point clouds (Zhu et al., 2022), and semantic segmentation (Park et al., 2022; Zhou et al., 2023; Liang et al., 2023), among others.

In multi-modal representation learning, the standard contrastive loss aims to maximize the similarity between corresponding image-text pairs (referred to as “positive pairs”) while distinguishing them from all non-matching image-text pairs (referred to as “negative pairs”). This objective aligns true image-text pairs to construct meaningful representations. Despite the effectiveness of contrastive loss in empirical applications for multi-modal representation learning, two open questions have been largely overlooked in previous works. Firstly, the reliability of the ground truth labels "positive" and "negative" from web-sourced datasets warrants scrutiny. Most common web-sourced datasets consider images and their corresponding descriptions as the sole true positive pairs. However, in these datasets, multiple image-text pairs may contain similar contents while being labeled as negative pairs. In other words, due to their large volume and automated collection processes without human labeling, web-sourced datasets naturally contain substantial noisy pairs. For instance, consider Figure 1, where the first image is deemed a true positive match for the text "*man and woman hold hands, walk to the beach*". Both other texts in the same batch would typically be labeled as negative samples to be distinguished from the image representation. However, the second text "*loving couple on a beach*" could also be considered semantically positive in reflecting the content of the first image, despite being labeled as "negative" during training. Moreover, other positive pairs in the dataset may feature dissimilar or vague descriptions, as illustrated in the right example in Figure 1. Such noisy data pairs have the potential to introduce mixed training signals and compromise performance accuracy.

The second open question pertains to whether contrastive learning can effectively handle such noisy pairs. The conventional design of contrastive learning emphasizes the significance of true positive pairs within every mini-batch while uniformly pushing away all negative pairs. Consequently, it may be susceptible to inconsistent training signals. For example, as depicted in Figure 1, even though the second text contains content more similar to the image, it is considered equally "negative" as other texts in the same batch. Without the flexibility to adjust the importance of each data pair, contrastive learning may tend to overfit to the noisy data pairs within web-sourced datasets, thereby resulting in sub-optimal solutions.

To address these limitations, we propose leveraging Bayesian modeling, known for its robustness in handling data noise with uncertainty, to enhance contrastive learning. Our approach involves augmenting the contrastive loss with stochastic weighting, allowing for automatic inference on the degree of noise present in each data pair. This introduces a level of flexibility, enabling the system to better navigate and adapt to the inherent uncertainties within the dataset. By assigning probability weights to data pairs, we ensure that they are treated more accurately based on their posterior of being genuine positive or negative pairs, rather than relying solely on batch-specific determinations, which can be erratic.

To facilitate efficient learning and inference, we first reframe the problem within a probability framework using Bayesian data augmentation techniques. This formulation enables us to infer the weight of each data pair in contrastive learning, thus ensuring the learned representations are robust to noisy training data. Finally, we develop a novel stochastic expectation maximization algorithm to incorporate the inferred random weights into the learning process of model parameters. In summary, our contributions are as follows:

- We identify and address the inherent noise problem in commonly used datasets for contrastive learning, formulating it as contrastive learning with noisy data pairs.
- We propose a principled method to tackle this problem by reformulating it within a probability framework and developing a stochastic expectation maximization algorithm for robust learning while inferring stochastic data-pair weights.
- We demonstrate significant performance improvements through extensive experiments on various public benchmarks for multi-modal contrastive learning.

2 Method

We begin by outlining the foundational setup and notation for contrastive learning. In this framework, a backbone network, parameterized by $\boldsymbol{\theta}$, is employed to generate generalized representations denoted as $\mathbf{z} = \mathsf{enc}(\mathbf{x}; \boldsymbol{\theta})$, where $\mathbf{x}$ denotes the input data. The multi-modal data is organized into positive and negative pairs. Specifically, in a multi-modal dataset $\mathcal{D} \triangleq (\mathbf{x}_i^1, \mathbf{x}_i^2)$, with superscripts indicating different modalities and subscripts indexing individual data samples, each $(\mathbf{x}_i^1, \mathbf{x}_i^2)$ represents a positive pair, while each $(\mathbf{x}_i^1, \mathbf{x}_j^2)$ with $i \neq j$ represents a negative pair. We define $s_{i+} \triangleq \mathsf{sim}(\mathsf{enc}(\mathbf{x}_i^1; \boldsymbol{\theta}), \mathsf{enc}(\mathbf{x}_i^2; \boldsymbol{\theta}))$ as the similarity score between the positive pair $(\mathbf{x}_i^1, \mathbf{x}_i^2)$ after passing through the encoder. Additionally, $s_{ik^-} \triangleq \mathsf{sim}(\mathsf{enc}(\mathbf{x}_i^{m_1}; \boldsymbol{\theta}), \mathsf{enc}(\mathbf{x}_k^{m_2}; \boldsymbol{\theta}))$ denotes the similarity score between the negative pair $(\mathbf{x}_i^{m_1}, \mathbf{x}_k^{m_2})$, where $m_1, m_2 \in 1, 2$, and $\mathsf{sim}(\cdot, \cdot)$ represents a similarity metric yielding positive values. In this paper, we adopt the exponential cosine similarity,

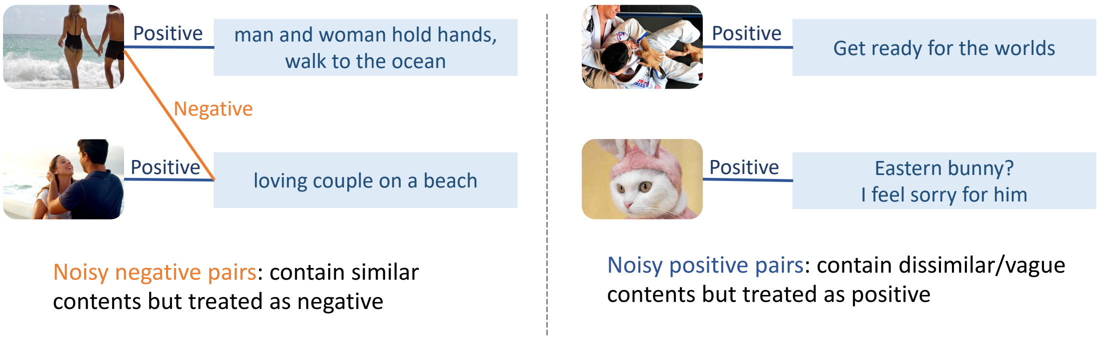

Positive
man and woman hold hands, walk to the ocean
Negative
Positive
loving couple on a beach
Noisy negative pairs: contain similar contents but treated as negative
Positive
Get ready for the worlds
Positive
Eastern bunny? I feel sorry for him
Noisy positive pairs: contain dissimilar/vague contents but treated as positive

Figure 1: Examples from CC3M dataset that contain noisy pairs.

commonly used in contrastive learning methods, defined as $\mathsf{sim}(\mathbf{x}_1, \mathbf{x}_2) \triangleq e^{\mathbf{x}_1^T \mathbf{x}_2}$. Note that the similarity scores depend on the model parameter $\boldsymbol{\theta}$, although we omit explicit reference to it in our notation for simplicity.

Preliminaries on Bayesian Modeling In contrast to the conventional approach to neural network modeling, Bayesian modeling treats parameters of interest as random variables, such as the weighting parameters w_i^+ and w_{ik}^- introduced in equation equation 1 below. Each stochastic parameter is associated with two types of distributions: the prior distribution and the posterior distribution. The prior distribution encapsulates our initial belief about the parameter's distribution before observing any data. For instance, in our case, we define the stochastic weights w_i^+ and w_{ik}^- to follow Gamma distributions, reflecting our hypothesis that each data pair should contribute differently to the loss. On the other hand, the posterior distribution combines our prior belief with the actual observed data, representing the "optimal" distribution in some sense and serving as the target of Bayesian inference. In the subsequent sections, we first reformulate the standard contrastive learning framework into a probabilistic framework and then propose efficient Bayesian inference methods to compute the posterior distribution of our stochastic weights, to compensate potential data noise in learning.

2.1 Probability Weighted Contrastive Learning

As discussed in the Introduction, contrastive learning is tailored for scenarios involving clean pair data. In the standard setup, each data sample comprises one positive pair and K negative pairs. The contrastive loss function is defined as follows:

$$\mathcal{L}_{\text{con}}(\mathcal{D}; \boldsymbol{\theta}) = -\frac{1}{|\mathcal{D}|} \sum_{\mathbf{x}_i \in \mathcal{D}} \log(\mathcal{L}_{\mathbf{x}_i}),$$

$$\text{with } \mathcal{L}_{\mathbf{x}_i} \triangleq \frac{s_{i+}}{s_{i+} + \sum_{k=1}^{K} s_{ik-}} .$$

However, real-world data often contain noisy pairs, making direct application of contrastive learning challenging. Here, we present our fundamental approach to address noisy pair data in contrastive representation learning. Our basic strategy is intuitive: we extend the standard contrastive loss by introducing learnable stochastic weights for all data pairs. Specifically, we incorporate local learnable weights w_i^+, w_{ik}^- associated with the data pairs, defining the noise-robust weighted contrastive loss as follows:

$$\mathcal{L}_{\text{con}}^r(\mathcal{D}; \boldsymbol{\theta}) = -\frac{1}{|\mathcal{D}|} \sum_{\mathbf{x}_i \in \mathcal{D}} \log(\mathcal{L}_{\mathbf{x}_i}^r), \tag{1}$$

$$\text{with } \mathcal{L}_{\mathbf{x}_i}^r \triangleq \frac{w_i^+ s_{i+}}{w_i^+ s_{ij+} + \sum_{k=1}^{K} w_{ik}^- s_{ik-}} , \tag{2}$$

where $\{w_i^+\}$ represents weights for positive pairs, and $\{w_{ik}^-\}$ for negative pairs. Notably, when all weights are set to one, the loss reduces to the standard contrastive loss.

One challenge with such a loss, however, is the quadratic growth of auxiliary random weights concerning the training data size (including augmented data), which becomes impractical to store in the context of continuous data augmentation. To address this challenge, drawing inspiration from the recent probability reformulation of contrastive learning (Chen et al., 2022), we propose a scalable Bayesian learning mechanism to efficiently sample the local weights in each iteration. These weights are then integrated into the contrastive loss to optimize the global model parameter. Specifically, we reframe the problem from a Bayesian modeling perspective by assigning appropriate priors for the weights. We can consider Bernoulli priors to model weights as binary random variables, or Gamma priors to model them as positive values. For simplicity, we adopt Gamma priors, given by:

$$w_i^+ \sim \mathsf{Gamma}(a_+, b_+), \;\; w_{ik}^- \sim \mathsf{Gamma}(a_-, b_-)\ ,$$

where a_+ and a_- are the shape parameters, and b_+ and b_- are the rate parameters. This gives a joint posterior distribution over the global model parameter and local random weight variables w_i^+ and w_{ik}^-:

$$p(\{w_i^+\}, \{w_{ik}^-\}, \boldsymbol{\theta}; \mathcal{D}) \propto \tag{3}$$
$$\textstyle\prod_{\mathbf{x}_i \in \mathcal{D}} \frac{w_i^+ s_{i+}}{w_i^+ s_{ij+} + \sum_{k=1}^K w_{ik}^- s_{ik-}} \cdot p(\{w_i^+\}) p(\{w_{ik}^-\}) p(\boldsymbol{\theta})$$

This probability weighting mechanism serves as a measure of confidence in the pairing, facilitating a more flexible and adaptive learning process. It accommodates variations and possible inconsistencies in the data, enabling the model to better adapt to real-world complexities.

Another challenge, however, arises from the infeasibility of directly performing Bayesian inference on such a posterior distribution due to the non-conjugacy between the priors and likelihood. To address this, we draw inspiration from (Chen et al., 2022) and introduce an augmented random variable u_i associated with each data point. This augmentation yields an augmented joint posterior distribution $p(\boldsymbol{\theta}, \mathbf{u}, \mathbf{w} \,|\mathcal{D})$[*], expressed as:

$$p(\boldsymbol{\theta}, \mathbf{u}, \mathbf{w} \,|\mathcal{D}) \propto \prod_{i:\mathbf{x}_i \in \mathcal{D}} w_i^+ s_{i+} e^{-\mathbf{u}_i w_i^+ s_{i+}} \tag{4}$$
$$\cdot \prod_k e^{-u_i w_{ik}^- s_{ik-}} p(\{w_i^+\}) p(\{w_{ik}^-\}) p(\boldsymbol{\theta})\ ,$$

where $\mathbf{u} \triangleq \{u_1, u_2, \cdots, u_{|\mathcal{D}|}\}$ and $\mathbf{w} \triangleq \{w_i^+\} \cup \{w_{ik}^-\}$. Subsequently, learning and inference can be conducted based on the augmented posterior $p(\boldsymbol{\theta}, \mathbf{u}, \mathbf{w} \,|\mathcal{D})$. In the following, we propose an efficient algorithm based on stochastic expectation maximization (stochastic EM) to alternately infer the local random variables and optimize the global model parameter.

2.2 Efficient Inference and Learning with Stochastic Expectation Maximization

Drawing on the concept outlined in (Chen et al., 2022), we propose a stochastic expectation maximization (EM) algorithm for efficient inference and learning of our model. Stochastic EM, a stochastic variant of the widely used EM algorithm, alternates between inferring local random variables and optimizing global model parameters for a latent variable model (Allassonnière and Chevallier, 2021; Chen et al., 2018; Delyon et al., 1999). The algorithm comprises three main steps: simulation, stochastic approximation, and maximization.

In our context, simulation involves sampling local random variables for a batch of data, denoted as $\mathbf{u}$ and $\mathbf{w}$. Stochastic approximation then employs the sampled auxiliary random variables to update a stochastic objective $Q(\boldsymbol{\theta})$ at each iteration t, given by the recursive formula: $Q_{t+1}(\boldsymbol{\theta}) = Q_t(\boldsymbol{\theta}) + \lambda_t(\log p(\boldsymbol{\theta}, \mathbf{u}, \mathbf{w} \,|\mathcal{D}) - Q_t(\boldsymbol{\theta}))$, where λ_t is a sequence of decreasing weights. Finally, in the maximization step, we optimize the model parameter $\boldsymbol{\theta}$ by maximizing the stochastic objective $Q_{t+1}(\boldsymbol{\theta})$. Further details are provided below.

Simulation Given the joint posterior distribution in equation 4 and the current batch of data, sampling the local random variables $\mathbf{u}$ and $\mathbf{w}$ is straightforward. Specifically, each u_i and w_i^+, w_{ik}^- follows a Gamma distribution:

$$u_i | \{w_i^+, w_{ik}^-, \boldsymbol{\theta}\} \sim \mathsf{Gamma}(a_u, b_u + w_i^+ s_{i+} + \sum_k w_{ik}^- s_{ik-}), \tag{5}$$

$$w_i^+ | \{\mathbf{u}, \boldsymbol{\theta}\} \sim \mathsf{Gamma}(1 + a_+, u_i s_{i+} + b_+) \tag{6}$$
$$w_{ik}^- | \{\mathbf{u}, \boldsymbol{\theta}\} \sim \mathsf{Gamma}(a_-, u_i s_{ik-} + b_-), \forall i, k$$

These sampled random variables for the current batch of data are then used in the subsequent stochastic approximation step. Optionally, for stability, we propose updating u_i's with moving averages after sampling. This involves maintaining u_i in memory and updating them as follows:: $u_i \leftarrow \alpha u_i + (1-\alpha)\tilde{u}_i$, where $\tilde{u}_i \sim \mathsf{Gamma}(a_u, b_u + w_i^+ s_{i+} + \sum_k w_{ik}^- s_{ik-})$ and $\alpha \in [0, 1]$ is a hyperparameter to balance old and new values. This strategy only requires limited storage overhead as we only need extra memory proportional to the training data size, which is considered negligible compared to other parameters.

Stochastic approximation Next, we calculate the stochastic approximation based on the simu-

[*] In the sense that marginalizing over the augmented random variables w_i^+ and w_{ik}^- in $p(\theta, \mathbf{U}, w_i^+, w_{ik}^- | \mathcal{D})$ yields the original $p(w_i^+, w_{ik}^-, \boldsymbol{\theta}; \mathcal{D})$. Thus, learning and inferences on the two forms are equivalent.

Algorithm 1 Noise-Robust Contrastive Learning with Stochastic EM

```
Initialize $\boldsymbol{\theta}$; set $t = 1$
for $\mathbf{x}_1, \mathbf{x}_2$ in loader do    ▷ load a minibatch $(\mathbf{x}_1, \mathbf{x}_2)$ with $B$ samples
    Calculate positive/negative similarity scores $\{s_{i+}\}$ and $\{s_{ik-}\}$
    Initialize all the weights $\{w_i^+\}$ and $\{w_{ik}^-\}$ to be one
    for $k = 1 \cdots$ iter [2 in practice] do
        Sample $\mathbf{u}$ according to equation 5
        Sample $\mathbf{w}$ according to equation 6
    end for
    Calculate the weighted contrastive loss in equation 1 with the sampled $\mathbf{w}$ on the current batch of data
    Update the model parameter by stochastic gradient descent with the calculated weighted contrastive loss
    $t = t + 1$
end for
```

lated local random variables. For simplicity in notation, let $Q_0(\boldsymbol{\theta}) = 0$. We reformulate $Q_{t+1}(\boldsymbol{\theta})$ by decomposing the recursion:

$$Q_{t+1}(\boldsymbol{\theta}) = \sum_{\tau=0}^{t} \tilde{\lambda}_\tau \log p(\boldsymbol{\theta}, \mathbf{u}_\tau, \mathbf{w}_\tau \,|\mathcal{D}_\tau), \tag{7}$$

$$\text{where } \tilde{\lambda}_\tau \triangleq \lambda_\tau \prod_{t'=\tau+1}^{t} (1 - \lambda_{t'}) \,,$$

where τ indexes the minibatch and the corresponding local random variables at the current time τ.

Maximization The stochastic approximation objective in equation 7 provides a convenient form for stochastic optimization over time, akin to online optimization. At each time t, we initialize the parameter $\boldsymbol{\theta}$ from the previous step and update it via stochastic gradient descent computed from the current batch of data. To mitigate variance, we propose optimizing a marginal version of $p(\boldsymbol{\theta}, \mathbf{u}_\tau, \mathbf{w}_\tau \,|\mathcal{D}\tau)$ by integrating out $\mathbf{u}_\tau$, essentially reducing to our original weighted contrastive loss in equation 1.

With these steps, we are prepared to optimize the model using stochastic EM. Algorithm 1 provides further details.

3 Experiments

We concentrate on image-text contrastive learning employing CLIP-based models, which utilize two separate encoders to align features across image and text modalities. Subsequently, we assess their performance on established benchmarks encompassing zero-shot, distribution shift, and linear probing tasks. Additionally, we conduct an ablation study and delve into the analysis of sampling hyperparameters and sampled weights to further illuminate our findings.

Method	Top1 ↑	Top5↑
CyCLIP (Goel et al., 2022)	17.77	36.20
LATENT (Jiang et al., 2023)	20.45	**39.28**
CLIP (Radford et al., 2021)	17.71	35.87
DeCL (Chen et al., 2022)	17.55	36.46
RINCE (Chuang et al., 2022)	17.46	34.61
OURS	**20.96**	38.24

Table 1: Zero-Shot Transfer Learning Classifiction Accuray (%) on ImageNet1K.

3.1 Experiments Setup

For our encoders, we utilize ResNet-50 (He et al., 2016) for the image encoder and BERT (Devlin et al., 2018) for the text encoder within our CLIP model. We adopt the official code from OpenCLIP (Ilharco et al., 2021) and DeCL (Chen et al., 2022) to replicate the baselines and implement our methods. Our reproduced CLIP results align closely with recent works (Mu et al., 2021; Gao et al., 2021; Jiang et al., 2023), albeit slightly lower than reported in the original CLIP paper. This discrepancy may stem from our utilization of fewer GPUs, resulting in a smaller effective batch size. *It is crucial to note that all methods utilize the same OpenCLIP codebase and same pretraining dataset, we also reproduce three baselines (CLIP, DeCL and RINCE) with identical hyper-parameter configurations.*, ensuring a fair comparison. And it is unfair to compare results from some literature that adopts different experimental settings, *e.g.*, (Andonian et al., 2022; Jia et al., 2021b; Hu et al., 2023).

Pre-training: We adhere to standard practice and pre-train our model on the CC3M dataset (Sharma et al., 2018), consisting of 3 million unique images and 4 million image-text pairs.

Evaluation: For zero-shot image classification, we leverage the pre-trained image encoder to derive image representations and utilize the pre-trained text encoder and prompts to formulate class descriptions, thus obtaining class representations. Evaluation is conducted on ImageNet for embedding

Method	ImageNetV2		ImageNetSketch		ImageNet-A		ImageNet-R	
	Top1 ↑	Top5 ↑	Top1 ↑	Top5 ↑	Top1 ↑	Top5 ↑	Top1 ↑	Top5 ↑
CyCLIP (Goel et al., 2022)	15.25	32.15	8.30	20.77	3.27	13.07	19.85	40.35
LATENT (Jiang et al., 2023)	17.37	36.65	10.90	26.11	3.87	16.76	23.85	45.03
CLIP (Radford et al., 2021)	16.44	**34.15**	10.23	24.21	**5.05**	**17.71**	24.75	46.30
DeCL (Chen et al., 2022)	15.58	33.11	10.1	22.57	3.94	15.66	22.68	44.26
RINCE (Chuang et al., 2022)	14.92	31.89	9.7	22.62	2.79	12.11	21.88	41.43
OURS	**17.63**	33.25	**12.36**	**25.76**	4.21	14.76	**25.85**	**46.42**

Table 2: Zero-Shot Natural Distribution Shift Classifiction Accuray (%).

	Caltech101	**SVHN**	**STL10**	**CIFAR10**	**CIFAR100**	**DTD**	**FGVCAircraft**	**OxfordPets**	**SST2**	**Food101**	**GTSRB**	**StanfordCars**	**Flowers102**	**ImageNet1K**	**Average**
CLIP	79.3	45.9	88.7	76.1	54.1	55.9	21.4	57.8	54.2	55.2	68.2	78.1	17.7	51.1	57.4
DeCL	76.5	40.9	89.2	75.3	52.7	56.3	19.8	56.1	53.6	53.0	66.8	73.3	15.4	50.24	55.68
OURS	**81.4**	**49.2**	**89.9**	**77.4**	**55.5**	**58.0**	**23.8**	**62.1**	**56.8**	**59.0**	**73.9**	**80.5**	**19.3**	**52.93**	**60.0**

Table 3: Linear Probing Top1 Classification Accuracy (%) on Vision Benchmarks.

quality and its distribution shifted benchmarks to assess the robustness of our methods. Additionally, we evaluate linear probing performance, wherein the encoders remain fixed, and a single linear layer is trained with supplementary supervision to evaluate the quality of the learned representations.

To underscore the effectiveness of our approach with noisy datasets, we introduce random noise of 10% into the training data by randomly selecting 10% of data pairs within each batch and resampling the positive labels, resulting in 10% of the training data featuring incorrect positive pairs. Three baselines(CLIP, DeCL, RINCE) are trained from scratch using the same codebase, fixed random seed, and consistent hyper-parameters for equitable comparisons. After pre-training, we evaluate the model trained on the final epoch for all baselines and our approach. We also report results from two other baselines (CyCLIP and LATENT) with same pre-training data and network architecture.

3.2 Zero-Shot Transfer Learning Evaluation

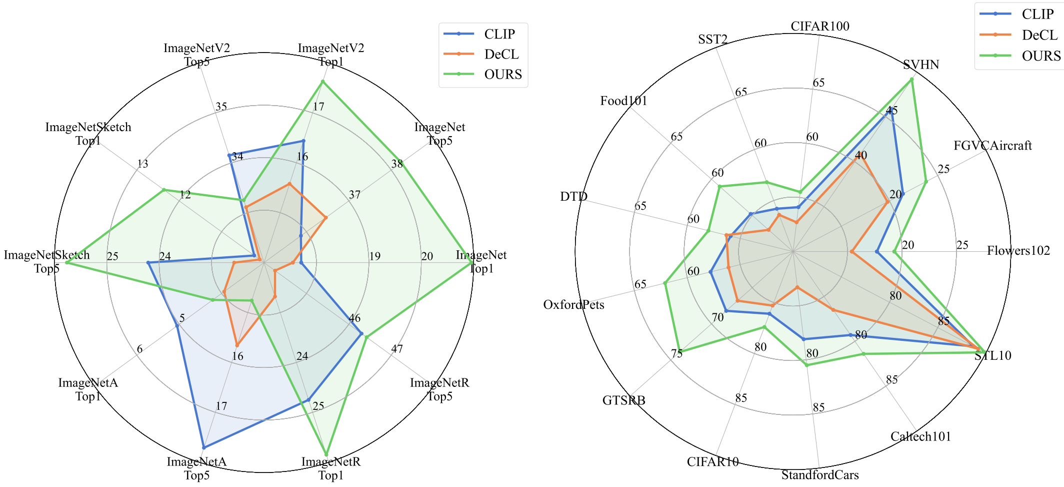
We perform zero-shot transfer learning on standard image classification tasks using the ImageNet1K dataset (Russakovsky et al., 2015). We adopt the common strategy of prompt engineering, constructing text prompts for each dataset using class names and templates such as "a photo of the [class name]" and "a sketch of the [class name]". The normalized class text embeddings are derived using multiple standard prompts, while image embeddings are obtained from the pre-trained encoder. During evaluation, the class whose text embedding exhibits the highest similarity score to the image embedding is utilized as the predicted label. We report Top-K classification accuracy with $K = 1$ and 5.

Table 1 presents the zero-shot transfer learning performance, including baselines for reference, with a focus on comparing with CLIP and DeCL. DeCL improves upon CLIP performance by 1% in Top-5 accuracy by addressing gradient bias issues, while our approach surpasses CLIP by 3% in both Top-1 and Top-5 accuracy through stochastic training pairs re-weighting. Notably, both DeCL and our method entail no additional computing overhead beyond the original CLIP baseline, except for the sampling processes, which are negligible relative to the total training cost.

3.3 Natural Distribution Shift Evaluation

We evaluate variations of the ImageNet1K dataset featuring shifted distributions (Recht et al., 2019; Wang et al., 2019; Hendrycks et al., 2021b,a), which incorporate sketches, cartoons, and adversarially generated images. These datasets serve to assess model generalizability and robustness. Employing the same processes outlined in the previous section, we conduct zero-shot evaluation and report classification accuracy on Top-1 and Top-5.

Table 2 showcases the zero-shot transfer learn-



(a) Zero-shot Performance.

(b) Linear Probing Performance.

Figure 2: Visualization of model performance. Every axis denotes the performance on a particular dataset measured using wither Top1 or Top5 accuracy metric. Distinct colors signify different methods or approaches. An approach that spans a larger area demonstrates superior overall performance.

ing performance on the Natural Distribution Shift benchmark. DeCL exhibits the poorest performance across all four benchmarks, while the CLIP baseline demonstrates the best performance on ImageNet-A. Additionally, CLIP exhibits decent performance on Top-5 accuracy for ImageNetV2. Our method enhances the CLIP baseline performance by 1-2% in Top-1 accuracy for three out of four benchmarks (ImageNet-V2, ImageNetSketch, and ImageNet-R), and by around 1% for two out of four benchmarks (ImageNetSketch and ImageNet-R). This indicates that by leveraging our approach to weight training pairs with stochastic approximation, we improve the robustness and generalizability of learned embeddings. However, our method underperforms CLIP on ImageNet-A, a dataset with adversarial noise, possibly due to the ineffectiveness of correcting noisy pairs in training to combat adversarial noise in data.

3.4 Linear Probing Evaluation

We further perform evaluations on linear probing classification tasks, wherein we fit a linear classifier with a downstream training dataset by leveraging the fixed learned visual encoder. The finetuned model is then evaluated on the testing dataset. This setting is used to evaluate how well the learned embeddings can generalize to new tasks with further supervision that requires only minimum finetuning effort. Following standard setup, we test on 14 standard benchmarks (Krizhevsky, 2009; Russakovsky et al., 2015; Fei-Fei et al., 2006; Netzer et al., 2011; Coates et al., 2011; Cimpoi et al., 2014; Maji et al., 2013; Parkhi et al., 2012; Socher et al., 2013; Bossard et al., 2014; Houben et al., 2013; Krause et al., 2013; Nilsback and Zisserman, 2008).

As shown in Table 3, our method outperforms both CLIP and DeCL on all the datasets, leading to an average gain of 3-4%. This further validates that our approach enables more flexible training with a higher tolerance for noisy data pairs, which can improve the model performance for better representations.

We visualize the model performance in Figure 2 where each color represents a different approach and the larger the area one approach covers indicates the better performance. We can see that our method outperforms baselines on both tasks with more advantage on linear probing tasks.

4 Related Works and Limitation

Vision-Language Representation Learning: Recent advances in vision-language representation learning can be broadly classified based on the manner in which information from two modalities is utilized for joint learning. The first category leverages unified models (Wang et al., 2021a, 2022b,c) to process both images and texts. Typically, these inputs are tokenized into sequences (Peng et al., 2022; Bao et al., 2022). The latter methods deploy separate encoders (Radford et al., 2021; Mokady et al., 2021; Shen et al., 2021; Li et al., 2021; Yang et al., 2022; Kwon et al., 2022; Jia et al., 2021a) for images and texts. To align the different modalities, they utilize the contrastive loss (Oord et al., 2018;

He et al., 2020; Chen et al., 2020). It's noteworthy that these techniques have been demonstrated to achieve state-of-the-art (SOTA) results on multiple downstream tasks. How to obtain robust and representational embeddings from CL is vital to benefit downstream tasks. Specifically, we focus on how to cope with noisy positive-negative pairs for CL.

Noisy Pairs in Contrastive Learning: While most works directly utilize large scale dataset for contrastive learning, some argue the noisy dataset issue. Noisy contrastive learning is an advanced technique that addresses the challenges of standard contrastive learning when faced with inconsistencies or "noise" within paired data. Traditional contrastive methods often struggle with mislabeled or ambiguous pairs, leading to decreased accuracy and efficiency. Noisy contrastive learning, on the other hand, incorporates mechanisms, often probabilistic in nature, to accommodate these uncertainties. By assigning confidence or probability weights to each pair, this approach allows for more adaptive and flexible learning. Rather than being limited by the binary classification of pairs, it embraces the inherent complexities and variations in real-world data, enhancing the model's robustness and performance. NLIP (Huang et al., 2023) enforces the pairs with larger noise probability to have fewer similarities. (Han et al., 2022) apply noise estimation component to adjust the consistency between different modalities for the action recognition task. RINCE (Hoffmann et al., 2022) uses a ranked ordering of positive samples to improve InfoNCE loss. Another recent work (Chen et al., 2022) studies the gradient bias issue in contrastive learning and proposes a stochastic approach to mitigate it with an bayesian augmentation. This method transforms the contrastive loss into a decomposable form. Consequently, conventional stochastic optimization can be applied without inducing gradient bias. Our approach uses a stochastic approach from a different perspective to address the noisy data issue instead of the gradient bias issue. To combat this challenge, we are introducing a probability extension. This innovative approach assigns a probability weight to each pair, whether positive or negative. By doing so, the model is no longer rigidly committed to a binary classification of the pairs but can now take into consideration the uncertainties or noise present in the data. This not only provides more nuanced information to the model but also enhances its robustness.

Stochastic Expectation Maximization Stochastic EM (Nielsen, 2000) stands as a pivotal algorithm in machine learning and probabilistic modeling. Building upon the foundations of the classical Expectation-Maximization (EM) algorithm (Lin, 2011), Stochastic EM offers an efficient solution for parameter estimation in situations involving vast datasets or latent variables, *e.g.*, to maximize the log-likelihood of $p(\mathbf{z}, \mathcal{D}|\boldsymbol{\theta})$, where $\mathcal{D}$ is the dataset, $\mathbf{z}$ is the local random variable and $\boldsymbol{\theta}$ is the global model parameter. By leveraging the power of mini-batch sampling, Stochastic EM strikes a balance between computational scalability and estimation accuracy. It has found widespread utility in various domains, including clustering (Allassonnière and Chevallier, 2021), topic modeling (Zaheer et al., 2016), and latent variable modeling (Zhang and Chen, 2020), making it an indispensable tool to cope with complex probabilistic models and extensive data and a natural fit to our problem.

Limitations: Our work is based on the contrastive learning framework where the training data are present in pairs. Expansion to broader contrastive learning setting, as well as other modalities would be our next steps.

5 Conclusion

We address a significant yet often overlooked limitation of standard CL, arising in data containing noisy positive-negative pairs. We overcome the limitation by presenting a principled solution, which reformulates CL within a probability framework and introduces random weights for data pairs. Leveraging Bayesian data augmentation techniques, we efficiently infer these random weights through sampling, and optimize the model parameters effectively using stochastic expectation maximization. Our innovative approach demonstrates its effectiveness through thorough evaluations on standard benchmarks, including applications in multi-modal contrastive learning using the CLIP framework. The results underscore the broad applicability and enhanced robustness of our proposed method. We believe our approach constitutes a valuable addition to the contrastive representation learning literature, capable of significantly improving the performance of state-of-the-art representation learning foundation models when applied to larger datasets.

References

Stéphanie Allassonnière and Juliette Chevallier. 2021. A new class of stochastic em algorithms. escaping local maxima and handling intractable sampling. *Comput. Stat. Data Anal.*, 159:107159.

Alex Andonian, Shixing Chen, and Raffay Hamid. 2022. Robust cross-modal representation learning with progressive self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16430–16441.

Hangbo Bao, Li Dong, and Furu Wei. 2022. Beit: Bert pre-training of image transformers. *ArXiv*, abs/2106.08254.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101 – mining discriminative components with random forests. In *Proc. ECCV*.

Changyou Chen, Jianyi Zhang, Yi Xu, Liqun Chen, Jiali Duan, Yiran Chen, Son Tran, Belinda Zeng, and Trishul Chilimbi. 2022. Why do we need large batchsizes in contrastive learning? a gradient-bias perspective. *Proc. NeurIPS*, 35:33860–33875.

Jianfei Chen, Jun Zhu, Yee Whye Teh, and Tong Zhang. 2018. Stochastic expectation maximization with variance reduction. In *NeurIPS*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proc. ICML*.

Ching-Yao Chuang, R Devon Hjelm, Xin Wang, Vibhav Vineet, Neel Joshi, Antonio Torralba, Stefanie Jegelka, and Yale Song. 2022. Robust contrastive learning against noisy views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16670–16681.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. 2014. Describing textures in the wild. In *Proc. CVPR*.

Adam Coates, A. Ng, and Honglak Lee. 2011. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*.

Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. 2022. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *Proc. ECCV*, pages 88–105, Cham. Springer Nature Switzerland.

Bernard Delyon, Marc Lavielle, and Éric Moulines. 1999. Convergence of a stochastic approximation version of the em algorithm. *Annals of Statistics*, 27:94–128.

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

Li Fei-Fei, R. Fergus, and P. Perona. 2006. One-shot learning of object categories. *IEEE TPAMI*.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2021. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2023. Clip-adapter: Better vision-language models with feature adapters. *IJCV*, pages 1–15.

Shashank Goel, Hritik Bansal, Sumit Kaur Bhatia, Ryan A. Rossi, Vishwa Vinay, and Aditya Grover. 2022. Cyclip: Cyclic contrastive language-image pretraining. *ArXiv*, abs/2205.14459.

Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. 2023. Calip: Zero-shot enhancement of clip with parameter-free attention. In *Proc. AAAI*, volume 37, pages 746–754.

Haochen Han, Qinghua Zheng, Minnan Luo, Kaiyao Miao, Feng Tian, and Yan Chen. 2022. Noise-tolerant learning for audio-visual action recognition. *arXiv preprint arXiv:2205.07611*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proc. CVPR*.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *Proc. CVPR*.

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

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Xiaodong Song. 2021b. Natural adversarial examples. In *Proc. CVPR*.

David T Hoffmann, Nadine Behrmann, Juergen Gall, Thomas Brox, and Mehdi Noroozi. 2022. Ranking info noise contrastive estimation: Boosting contrastive learning via ranked positives. In *Proc. AAAI*, volume 36, pages 897–905.

Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. 2013. Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark. In *International Joint Conference on Neural Networks*, 1288.

Peng Hu, Zhenyu Huang, Dezhong Peng, Xu Wang, and Xi Peng. 2023. Cross-modal retrieval with partially mismatched pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Runhui Huang, Yanxin Long, Jianhua Han, Hang Xu, Xiwen Liang, Chunjing Xu, and Xiaodan Liang. 2023. Nlip: Noise-robust language-image pre-training. In *Proc. AAAI*, volume 37, pages 926–934.

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. 2021. Openclip.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021a. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proc. ICML*.

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

Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, and Trishul Chilimbi. 2023. Understanding and constructing latent modality structures in multi-modal representation learning. In *Proc. CVPR*, pages 7661–7671.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *IEEE Workshop on 3D Representation and Recognition (3dRR-13)*.

Alex Krizhevsky. 2009. Learning multiple layers of features from tiny images.

Gukyeong Kwon, Zhaowei Cai, Avinash Ravichandran, Erhan Bas, Rahul Bhotika, and Stefan 0 Soatto. 2022. Masked vision and language modeling for multi-modal representation learning. *ArXiv*, abs/2208.02131.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Proc. NeurIPS*.

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proc. CVPR*, pages 7061–7070.

Dahua Lin. 2011. An introduction to expectation-maximization.

Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2022. Frozen clip models are efficient video learners. In *Proc. ECCV*, pages 388–404. Springer.

Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. 2023. More control for free! image synthesis with semantic diffusion guidance. pages 289–299.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proc. ICLR*.

S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. 2013. Fine-grained visual classification of aircraft. Technical report.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

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

Yuval Netzer, Tao Wang, Adam Coates, A. Bissacco, Bo Wu, and A. Ng. 2011. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*.

Søren Nielsen. 2000. The stochastic em algorithm: estimation and asymptotic results. *Bernoulli*, 6:457–489.

Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. *Indian Conference on Computer Vision, Graphics & Image Processing*.

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

Kwanyong Park, Sanghyun Woo, Seoung Wug Oh, In So Kweon, and Joon-Young Lee. 2022. Per-clip video object segmentation. In *Proc. CVPR*, pages 1352–1361.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. 2012. Cats and dogs. In *Proc. CVPR*.

Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. 2022. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *ArXiv*, abs/2208.06366.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proc. ICML*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Fine-tuned clip models are efficient video learners. In *Proc. CVPR*, pages 6545–6554.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet? In *Proc. ICML*.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proc. CVPR*, pages 22500–22510.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. Imagenet large scale visual recognition challenge. *IJCV*, 115.

Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. 2022. Clip-forge: Towards zero-shot text-to-shape generation. In *Proc. CVPR*, pages 18603–18613.

Aditya Sanghi, Rao Fu, Vivian Liu, Karl DD Willis, Hooman Shayani, Amir H Khasahmadi, Srinath Sridhar, and Daniel Ritchie. 2023. Clip-sculptor: Zero-shot generation of high-fidelity and diverse shapes from natural language. In *Proc. CVPR*, pages 18339–18348.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Proc. NeurIPS*, volume 35, pages 25278–25294. Curran Associates, Inc.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. 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)*. Association for Computational Linguistics.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2021. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*.

Mustafa Shukor, Guillaume Couairon, and Matthieu Cord. 2022. Efficient vision-language pretraining with visual concepts and hierarchical alignment. *ArXiv*, abs/2208.13628.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. EMNLP*. Association for Computational Linguistics.

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

Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. 2022a. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proc. CVPR*, pages 3835–3844.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. 2019. Learning robust global representations by penalizing local predictive power. In *Proc. NeurIPS*.

Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2021a. Ufo: A unified transformer for vision-language representation learning. *ArXiv*, abs/2111.10023.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022b. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *Proc. ICML*.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 2022c. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *ArXiv*, abs/2208.10442.

Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. 2021b. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *ArXiv*, abs/2111.02358.

Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. 2023. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *Proc. CVPR*, pages 20908–20918.

Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. 2022. Vision-language pre-training with triple contrastive learning. In *Proc. CVPR*.

Manzil Zaheer, Michael Wick, Jean-Baptiste Tristan, Alex Smola, and Guy Steele. 2016. Exponential stochastic cellular automata for massively parallel inference. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 966–975, Cadiz, Spain. PMLR.

Siliang Zhang and Yunxiao Chen. 2020. Computation for latent variable model estimation: A unified stochastic proximal framework. *Psychometrika*, 87:1473 – 1502.

Chong Zhou, Chen Change Loy, and Bo Dai. 2022. Extract free dense labels from clip. In *Proc. ECCV*, pages 696–712. Springer.

Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. 2023. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proc. CVPR*, pages 11175–11185.

Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyao Zeng, Shanghang Zhang, and Peng Gao. 2022. Pointclip v2: Adapting clip for powerful 3d open-world learning. *arXiv preprint arXiv:2211.11682*.

A Analysis

A.1 Sensitivity to Sampling Parameters

We conduct an analysis to investigate the sensitivity of our method to different sampling parameters. As detailed in Section 2.2 and Algorithm 1, there are several hyperparameters associated with the Gamma distributions that require determination. Following the same setting as in DeCL, we introduce a Gamma prior for u_i's with shape and rate parameters $a_u = 1$ and $b_u = 0$. Subsequently, we determine the parameters for the prior Gamma distribution for $\mathbf{w}$, where we need to establish a_- and b_- for negative pairs, as well as a_+ and b_+ for positive pairs.

To simplify and without loss of generality, we set b_- and b_+ to be 0. To reduce the search space, we fix a_+ and perform a grid search for the optimal value of a_-. Specifically, we set $a_+ = 5$ and explore $1, 5, 10, 20$ for a_-, where a higher value indicates a stronger preference for higher weight in the prior on negative pairs.

The results are presented in Table 4. Optimal a_- is observed to be twice that of a_+, with the trend indicating that neither higher nor lower values yield greater gains. This suggests that slightly higher weights on negative pairs are preferable in noisy dataset training scenarios, while excessive attention to negative pairs is undesirable as it may diminish the learning signal from positive pairs.

Furthermore, we visualize the learned distribution sample results in Figure 3. It is observed that with proper hyperparameter settings, the majority of sampled weights cluster around 1, with some pairs associated with significantly higher or lower weights. This observation aligns with expectations, as our objective is to enable the model to adaptively determine lower weights for noisy training pairs.

A.2 Visualization of Embeddings.

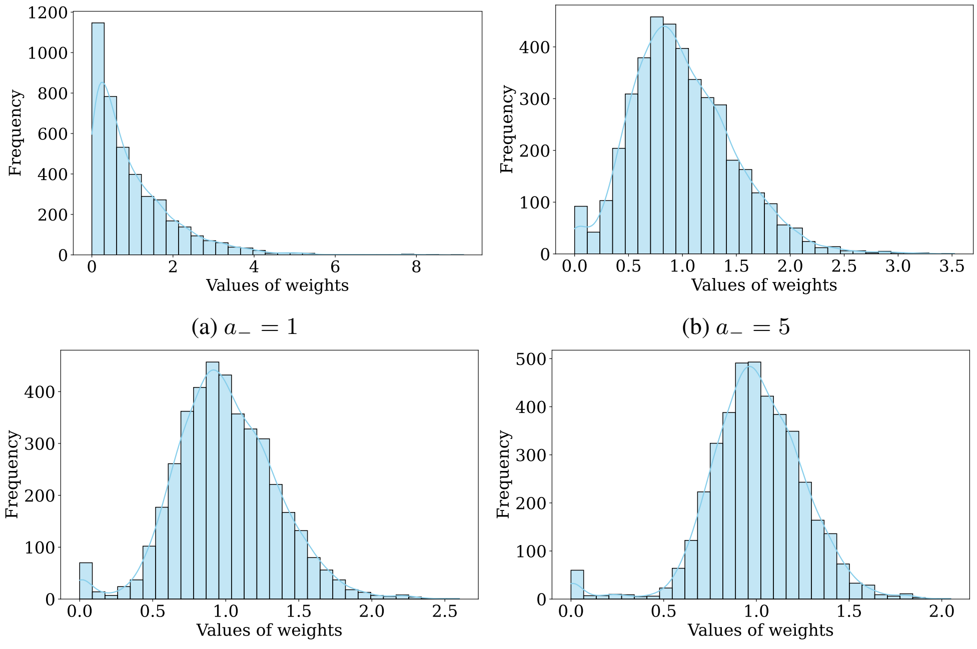
In our study, we examine the visualization of image and text class embeddings of CIFAR10 dataset from trained models employing distinct approaches. Our analysis reveals that both RINCE and our method facilitate a more uniform distribution of text embeddings.

B Implementation Details

We maintain consistency with the OpenCLIP codebase and hyper-parameter settings, except for the number of GPUs. Training is executed from scratch across 8 NVIDIA V100 GPUs for 32 epochs, with a batch size of 128 per GPU and a feature dimension of 1024. An initial learning rate of 5×10^{-4} is employed, with a warm-up period of 10000 iterations and subsequent cosine decay scheduling. AdamW optimizer (Loshchilov and Hutter, 2019) is utilized, along with a weight decay of 0.2. To underscore the effectiveness of our approach with noisy datasets, we introduce random noise of 10% into the training data by randomly selecting 10% of data pairs within each batch and re-sampling the positive labels, resulting in 10% of the training data featuring incorrect positive pairs. All baselines are trained from scratch using the same codebase, fixed random seed, and consistent hyper-parameters for equitable comparisons. After pre-training, we evaluate the model trained on the final epoch for all baselines and our approach.

Table 4: Effect of Changing Sampling Parameters on ImageNet zero-Shot Classification (%).

	$a_- = 1$		$a_- = 5$		$a_- = 10$		$a_- = 20$	
	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
$a_+ = 5$	18.00	34.57	18.02	34.55	**20.96**	**38.24**	18.39	35.38



(a) $a_- = 1$

(b) $a_- = 5$

(c) $a_- = 10$

(d) $a_- = 20$

Figure 3: Posterior sample distribution of pair weights $\mathbf{w}$ with different prior choices, where $a_+ = 5$. $a_- = 10$ features the best performance.

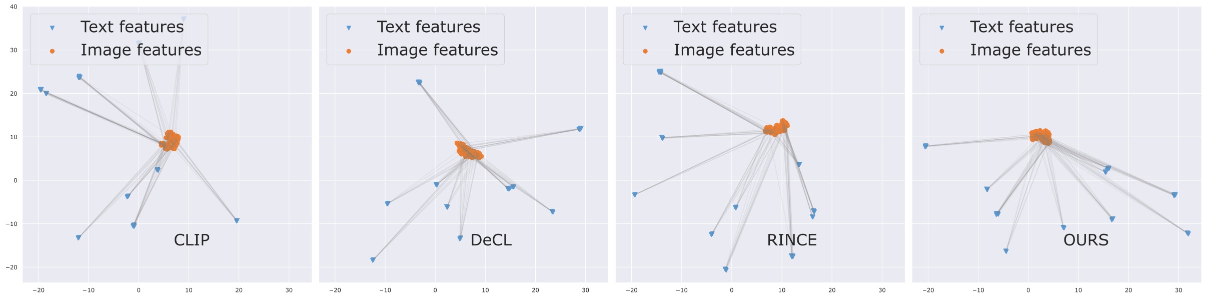


Figure 4: Visualization of Embeddings