LEARNING MULTI-MODAL REPRESENTATION ALIGNMENTS FROM NOISY DATA-PAIRS

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive learning (CL) represents one of the most successful paradigms for self-supervised representation learning, which has been applied to SOTA multi-modal learning applications. One overlooked limitation of standard contrastive learning, however, is that it is not designed for robust learning in the presence of noisy data pairs. For example, not all negative samples are truly negative, *e.g.*, within a mini-batch there can be negative samples that are semantically as positive as the positive sample. This is common in most web-sourced multi-modal datasets such as CC3M and YFCC that are frequently used for CL, due to the noisy nature when crawling the datasets. Consequently, the noise in the datasets could significantly impair the power of CL. To remedy this issue, we propose a novel solution by reformulating the standard CL into a probability framework, and introducing learnable random weights to associate with data pairs, so as to allow automatic inference of the degree of noisiness for each data pair. Within our probability framework, posterior inference of the random weights can be done efficiently with Bayesian data augmentation. Consequently, the model can be effectively optimized by a novel learning algorithm based on stochastic expectation maximization. We demonstrate the effectiveness of our approach on several standard multi-modal contrastive learning benchmarks, which significantly outperforms standard contrastive learning.

1 INTRODUCTION

Contrastive learning has become increasingly popular in multi-modal representation learning due to its effectiveness in aligning representations from different modalities. In the context of vision-language representation learning, the model aims to learn generic representations from images and texts that could benefit multi-modal downstream applications such as zero-shot image classification and image-text retrieval. Recent advances (Radford et al., 2021; Jia 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 applying contrastive loss to pre-train the model with a substantial volume of web-sourced paired image-text data such as Conceptual Caption (Sharma et al., 2018), YFCC (Thomee et al., 2016), Laion (Schuhmann et al., 2022). While some studies combine the representations of two modalities into a single encoder (Wang et al., 2021a; 2022b;c; 2021b), it is more prevalent to represent the image and text modalities separately using modality-specific encoders similar to the CLIP framework (Mokady et al., 2021; Shen et al., 2021; Jia et al., 2021; Li et al., 2021; Duan et al., 2022; Yang et al., 2022; Shukor et al., 2022). After pre-training, the model can produce general representations of both image and text inputs, demonstrating exceptional performance in subsequent tasks. Recent advances show that these high-quality representations can be adapted to 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 shape (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), etc.

In multi-modal representation learning, standard contrastive loss seeks to maximize the similarity between corresponding image-text pairs (termed "positive pairs") while distinguishing them from all the non-matching image-text pairs (termed "negative pairs"). Such an objective aligns the true image-text pairs together to build meaningful representations. Although contrastive loss has proven effective in empirical applications for multi-modal representation learning, there remain two open

Figure 1: Examples from CC3M (Sharma et al., 2018) dataset that contain noisy pairs.

questions that have been largely ignored in previous works. First, are the ground truth labels of "positive" and "negative" from the web-sourced dataset truly reliable? Most common web-sourced datasets consider images and their corresponding descriptions as the **only** true positive pairs. Yet in those datasets there can be multiple image-text pairs containing similar contents while being labeled as negative pairs. In other words, web-sourced datasets, due to their large volume and automated collection processes without human labeling, naturally contain substantial noisy pairs. For example, in Figure 1, the first image is considered as a true positive to the text "*man and woman hold hands, walk to the beach*". Both the other texts in the same batch would be considered as negative samples that should be pushed away from the representation of the image. However, the second text "*loving couple on a beach*" can also be considered semantically positive to reflect the content of the first image, while being labeled as "negative" during training. In addition, there also can be other positive pairs in the dataset that contain dissimilar or vague descriptions such as the right example in Figure 1. Such noisy data pairs could potentially lead to mixed training signals and loss of accuracy in performance.

The second open question is whether contrastive learning can handle such noisy pairs. The design of conventional contrastive learning amplifies the importance of true positive pairs within every mini-batch and pushes away all the negative pairs equally. Thus it could be susceptible to inconsistent training signals. For instance, in Figure 1, although the second text contains more similar content to the image, it is treated equally "negative" as other texts in the same batch. Without the flexibility to adjust the importance of each data pair, contrastive learning could overfit into the noisy data pairs within the web-source dataset, leading to sub-optimal solutions.

To address these limitations, we propose a fundamental approach to incorporate stochastic weighting into contrastive learning. Specifically, we augment the contrastive loss by assigning a probability weight to each data pair to allow automatic inference on the degree of nosiness level of the pair. By doing so, we can imbue the system with a degree of flexibility, allowing it to better discern and adapt to the inherent uncertainties in the data. This ensures that data pairs are treated more accurately based on their likelihood of being genuine positive or negative pairs, rather than relying on potentially erratic batch-specific determinations. For efficient learning and inference, we first reformulate the problem into a probability framework with Bayesian data augmentation. The formulation allows us to efficiently infer the weight of each data pair in contrastive learning, such that the learned representation is robust towards noisy training data. Finally, we develop a stochastic expectation maximization algorithm to incorporate the inferred random weights for efficient learning of model parameters. To summarize, our paper has the following major contributions:

- For the first time, we identify the inherent noise problem for some most commonly-used datasets for contrastive learning, and formulate the problem as contrastive learning with noisy data pairs.
- We propose a principled method to solve the problem by reformulating it into a probability framework with Bayesian data augmentation techniques. Based on the reformulation, a novel stochastic expectation maximization algorithm is developed to effectively learn the robust model while simultaneously inferring the stochastic data-pair weights.
- With extensive and large-scale experiments, we demonstrate improved performance on several public benchmarks for multi-modal contrastive learning.

2 METHOD

We start by describing the basic setup and notation in contrastive learning, where a backbone network, parameterized by $\boldsymbol{\theta}$, is used to generate generalized representations, written as $\mathbf{z} = \mathsf{enc}(\mathbf{x}; \boldsymbol{\theta})$ for input data $\mathbf{x}$. The multi-modal data is represented in terms of positive and negative data pairs. Specifically, given a multi-modal dataset $\mathcal{D} \triangleq \{(\mathbf{x}_i^1, \mathbf{x}_i^2)\}$ where the superscript indexes different modalities and subscript indexes data samples, each $(\mathbf{x}_i^1, \mathbf{x}_i^2)$ represents a positive pair and each $(\mathbf{x}_i^1, \mathbf{x}_j^2)$ with $i \neq j$ represents a negative pair. Denote $s_{i^+} \triangleq \mathsf{sim}(\mathsf{enc}(\mathbf{x}_i^1; \boldsymbol{\theta}), \mathsf{enc}(\mathbf{x}_i^2; \boldsymbol{\theta}))$ as the similarity score between the positive pair $(\mathbf{x}_i^1, \mathbf{x}_i^2)$ after the encoder; and $s_{ik^-} \triangleq \mathsf{sim}(\mathsf{enc}(\mathbf{x}_i^{m_1}; \boldsymbol{\theta}), \mathsf{enc}(\mathbf{x}_k^{m_2}; \boldsymbol{\theta}))$ as 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 $\mathsf{sim}(\cdot, \cdot)$ denotes a similarity metric (positive value). We adopt the exponential cosine similarity used in most contrastive learning methods in this paper, *i.e.*, $\mathsf{sim}(\mathbf{x}_1, \mathbf{x}_2) \triangleq e^{\mathbf{x}_1^T \mathbf{x}_2}$. Note the similarity scores depend on the model parameter $\boldsymbol{\theta}$, but we omit it in our development for notation simplicity.

2.1 PROBABILITY WEIGHTED CONTRASTIVE LEARNING

As discussed in the Introduction, contrastive learning is designed specifically for the ideal case of clean pair data. Specifically, consider the standard setup with one positive pair and K negative pairs for each data sample. The contrastive loss is defined as:

$$\mathcal{L}_{\text{con}}(\mathcal{D}; \boldsymbol{\theta}) = -\frac{1}{|\mathcal{D}|} \sum_{\mathbf{x}_i \in \mathcal{D}} \log(\mathcal{L}_{\mathbf{x}_i}), \text{ with } \mathcal{L}_{\mathbf{x}_i} \triangleq \frac{s_{i^+}}{s_{i^+} + \sum_{k=1}^K s_{ik^-}} .$$

However, real data usually come with noisy pairs, rendering directly applying contrastive learning problematic. In the following, we describe our fundamental method to deal with such a noisy pair data setting for contrastive representation learning. Our basic idea is intuitive, which is to generalize the standard contrastive loss by adding learnable stochastic weights for all the data pairs. Specifically, we introduce local learnable weights $\{w_i^+, w_{ik}^-\}$ associated with the data pairs, and define the following noise-robust weighted contrastive loss:

$$\mathcal{L}_{\text{con}}^r(\mathcal{D}; \boldsymbol{\theta}) = -\frac{1}{|\mathcal{D}|} \sum_{\mathbf{x}_i \in \mathcal{D}} \log(\mathcal{L}_{\mathbf{x}_i}^r), \text{ with } \mathcal{L}_{\mathbf{x}_i}^r \triangleq \frac{w_i^+ s_{i^+}}{w_i^+ s_{ij^+} + \sum_{k=1}^K w_{ik}^- s_{ik^-}} , \tag{1}$$

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

One challenge with such a loss, however, is that these auxiliary random weights are local random variables that grow quadratically w.r.t. the training data size (including augmented data), which is essentially infinite and thus infeasible to be stored in the setting of continuous data augmentation. To overcome the challenge, inspired by 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, which are then integrated into the contrastive loss to optimize the global model parameter.

Specifically, we reformulate the problem from a Bayesian inference perspective, where we assign appropriate priors for the weights. We can consider either Bernoulli priors to model weights as binary random variables, or Gamma priors to model them as positive values. For modeling convenience, we consider Gamma priors, *i.e.*,

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

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}^-, as

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

This probability weighting mechanism can be seen as a measure of confidence in the pairing, offering a more flexible and adaptive learning process. It can accommodate the variations and possible inconsistencies in the data, allowing the model to better adapt to real-world complexities.

Another challenge, however, is that directly performing Bayesian inference on such a posterior distribution is infeasible, due to the non-conjugacy between the priors and likelihood. Fortunately, we can borrow ideas from Chen et al. (2022) to introduce an augmented random variable u_i to associate to each data point, giving us an augmented joint posterior distribution equivalent to $p(\{w_i^+\}, \{w_{ik}^-\}, \boldsymbol{\theta}|\mathcal{D})^*$, as

$$p(\boldsymbol{\theta}, \mathbf{u}, \mathbf{w}\,|\mathcal{D}) \propto \prod_{i:\mathbf{x}_i\in\mathcal{D}} w_i^+ s_{i^+} e^{-\mathbf{u}_i w_i^+ s_{i^+}} \prod_k e^{-u_i w_{ik}^- s_{ik^-}} p(\{w_i^+\})p(\{w_{ik}^-\})p(\boldsymbol{\theta}) \,, \tag{2}$$

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

2.2 Efficient Inference and Learning with Stochastic Expectation Maximization (Stochastic EM)

Based on the idea in Chen et al. (2022), we propose a stochastic EM algorithm for efficient inference and learning of our model. Stochastic EM is a stochastic variant of the popular EM algorithm, which alternatively infers local random variables and optimizes global model parameters for a latent variable model (Allassonnière & Chevallier, 2021; Chen et al., 2018; Delyon et al., 1999). It consists of three steps: *simulation, stochastic approximation, and maximization*. In our setting, *simulation* corresponds to sampling local random variables for a batch of data, *e.g.*, $\mathbf{u}$ and $\mathbf{w}$; *stochastic approximation* then uses the sampled auxiliary random variables to update a stochastic objective $Q(\boldsymbol{\theta})$ at each iteration t as: $Q_{t+1}(\boldsymbol{\theta}) = Q_t(\boldsymbol{\theta}) + \lambda_t(\log p(\boldsymbol{\theta}, \mathbf{u}, \mathbf{w}\,|\mathcal{D}) - Q_t(\boldsymbol{\theta}))$, where $\{\lambda_t\}$ is a sequence of decreasing weights; Finally, in *maximization*, we optimize the model parameter $\boldsymbol{\theta}$ by maximizing the stochastic objective $Q_{t+1}(\boldsymbol{\theta})$. We describe more details below:

Simulation Given the joint posterior distribution in equation 2 and the current batch of data, one can easily sample the local random variables $\mathbf{u}$ and $\mathbf{w}$, which simply follow Gamma distributions of the following forms:

$$u_i|\{w_i^+, w_{ik}^-, \boldsymbol{\theta}\} \sim \mathsf{Gamma}(a_u, b_u + w_i^+ s_{i^+} + \sum_k w_{ik}^- s_{ik^-}), \;\; \forall i, \text{ and} \tag{3}$$

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

These sampled random variables for the current batch of data will then be used in the stochastic approximation step described below. Optionally, to make the algorithm more stable, we propose to update u_i's with moving averages after sampling, *e.g.*, we maintain $\{u_i\}$ in the memory and update them as: $u_i \leftarrow \alpha u_i + (1-\alpha)\tilde{u}_i$, 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 hyper-parameter 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 We then proceed to calculate the stochastic approximation based on the simulated local random variables above. For notation simplicity, we define $Q_0(\boldsymbol{\theta}) = 0$. Then we can reformulate $Q_{t+1}(\boldsymbol{\theta})$ by decomposing the recursion, resulting in

$$Q_{t+1}(\boldsymbol{\theta}) = \sum_{\tau=0}^{t} \tilde{\lambda}_\tau \log p(\boldsymbol{\theta}, \mathbf{u}_\tau, \mathbf{w}_\tau\,|\mathcal{D}_\tau), \text{ where } \tilde{\lambda}_\tau \triangleq \lambda_\tau \prod_{t'=\tau+1}^{t} (1 - \lambda_{t'}) \,, \tag{4}$$

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

*In the sense that marginalizing over the augmented random variables $\{w_i^+\}$ and $\{w_{ik}^-\}$ in $p(\theta, \mathbf{U}, \{w_i^+\}, \{w_{ik}^-\}|\mathcal{D})$ gives back to 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 3
    Sample $\mathbf{w}$ according to equation 3
  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

Maximization The stochastic approximation objective in equation 4 provides a convenient form for stochastic optimization over time, similar to online optimization (Bent & Van Hentenryck, 2005). Specifically, at each time t, we can initialize the parameter $\boldsymbol{\theta}$ from the last step, and update it by stochastic gradient descent calculated from the current batch of data. To reduce variance, we propose to optimize a marginal version of $p(\boldsymbol{\theta}, \mathbf{u}_\tau, \mathbf{w}_\tau \,|\mathcal{D}_\tau)$ by integrating out $\mathbf{u}_\tau$, which essentially reduces to our original weighted contrastive loss in equation 1.

With the above steps, it is ready to optimize the model by stochastic EM. The details are provided in Algorithm 1.

3 Related Works

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; 2021b) 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; Duan et al., 2022; Yang et al., 2022; Shukor et al., 2022; Kwon et al., 2022; Jia et al., 2021) 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 in embedding space to improve the model training. 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 levitate it. To combat the gradient bias, the authors introduce a Bayesian data augmentation approach. This new method transforms the contrastive loss into a decomposable form. Consequently, conventional stochastic optimization can be applied with-

out inducing gradient bias. Our approach uses a stochastic approach from a different perspective to address the noisy data 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 & Chevallier, 2021), topic modeling (Zaheer et al., 2016), and latent variable modeling (Zhang & Chen, 2020), making it an indispensable tool to cope with complex probabilistic models and extensive data and a natural fit to our problem.

4 EXPERIMENTS

We conduct experiments focused on image-text contrastive learning using CLIP-based models, wherein two distinct encoders are trained to align features between image and text modalities. We then evaluate on standard benchmarks including zero-shot, distribution shift, and linear probing tasks. We also provide ablation study and analysis on the sampling hyper-parameters and sampled weights.

4.1 EXPERIMENTS SETUP

For encoders, our CLIP model adopts ResNet-50 (He et al., 2016) as the image encoder and BERT (Devlin et al., 2018) as the text encoder. We adopt the official code from OpenCLIP (Ilharco et al., 2021) and DeCL (Chen et al., 2022) to reproduce the baselines and our methods. Our reproduced CLIP results are consistent with the recent works (Mu et al., 2021; Gao et al., 2021; Duan et al., 2022; Jiang et al., 2023), although their results are slightly lower than reported in the original CLIP paper. One possible reason is that we use fewer GPUs, thus leading to a smaller effective batch size. It is important to highlight that all the methods adopt the same OpenCLIP codebase and identical hyper-parameter configurations, thus ensuring a fair comparison.

Pre-training: We follow the standard practice and pre-train the model with the CC3M (Sharma et al., 2018) dataset with 3M unique images and 4M image-text pairs.

Evaluation: For zero-shot image classification evaluation, we take the pre-trained image encoder to obtain image representation, as well as the pre-trained text encoder and prompts to construct class descriptions to obtain class representations. We evaluate on ImageNet for embedding quality and its distribution shifted benchmarks to evaluate the robustness of our methods. We further evaluate linear probing performance, where the encoders are fixed and one linear layer is trained with additional supervision to evaluate the quality of the learned representations

Implementation Details: We follow the same code base and hyper-parameters setting as OpenCLIP except for the number of GPUs. We train the model from scratch on 8 NVIDIA V100 GPUs for 32 epochs. Our batch size is set to 128 per GPU and the feature dimension is 1024. We use an initial learning rate of $5e^{-4}$. We warm up the learning rate for 10000 iterations and follow the cosine decay scheduling. AdamW (Loshchilov & Hutter, 2019) optimizer is used along with a weight decay of 0.2. To further demonstrate the effectiveness of our approach for noisy datasets, we add a random noise of 10% into the training data by randomly selecting 10% of data pairs within every batch and re-sample the positive labels such that 10% of the training data has incorrect positive pairs. For all the baselines we use the same codebase to train from scratch with fixed random seed and the same hyper-parameters for fair comparisons. After pre-training, we evaluate the model trained on the last epoch for all baselines and our approach.

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

Method	Top1 Accuracy ↑	Top5 Accuracy ↑
CLIP	17.71	35.87
DeCL	17.55	36.46
OURS	**20.96**	**38.24**

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

Method	ImageNetV2		ImageNetSketch		ImageNet-A		ImageNet-R	
	Top1 ↑	Top5 ↑	Top1 ↑	Top5 ↑	Top1 ↑	Top5 ↑	Top1 ↑	Top5 ↑
CLIP	16.44	**34.15**	10.23	24.21	**5.05**	**17.71**	24.75	46.30
DeCL	15.58	33.11	10.1	22.57	3.94	15.66	22.68	44.26
OURS	**17.63**	33.25	**12.36**	**25.76**	4.21	14.76	**25.85**	**46.42**

4.2 Zero-Shot Transfer Learning Evaluation

We conduct zero-shot transfer on standard image classification tasks using the ImageNet1K dataset (Russakovsky et al., 2015). We employ the standard evaluation strategy of prompt engineering. For each dataset, we construct text prompts using the name of the class with some templates, for example, "a photo of the `[class name]`" and "a sketch of the `[class name]`". We obtain the normalized class text embedding for each class with multiple standard prompts. We obtain the image embeddings from the pre-trained encoder. During evaluation, the class whose text embedding has the highest similarity score to the image embedding is used as the prediction of the label. Consistent with previous works, we report Top-K classification accuracy with $K = 1, 5$.

We show in Table 1 the zero-shot transfer learning performance, we include other baselines for reference while we mainly focus on comparing with CLIP and DeCL. DeCL improves the clip baseline performance by 1% on Top5 accuracy by solving the gradient bias issue, while our approach can improve over CLIP by 3% on both Top1 and Top5 accuracy with stochastic training pairs reweighting. Note that both DeCL and our method do not require additional computing except for the sampling processes compared to the original CLIP baseline, which is negligible relative to the total training cost.

4.3 Natural Distribution Shift Evaluation

We also assess variations of the ImageNet1K datasets with featuring shifted distributions (Recht et al., 2019; Wang et al., 2019; Hendrycks et al., 2021b;a). These datasets incorporate sketches, cartoons, and adversarially generated images. They are usually considered as domain-shifted versions of ImageNet and are frequently utilized to evaluate the generalizability and robustness of models, as they usually contain harder or less common data samples. We perform the zero-shot evaluation using the same processes mentioned in the previous section and report classification accuracy on Top-1 and Top-5.

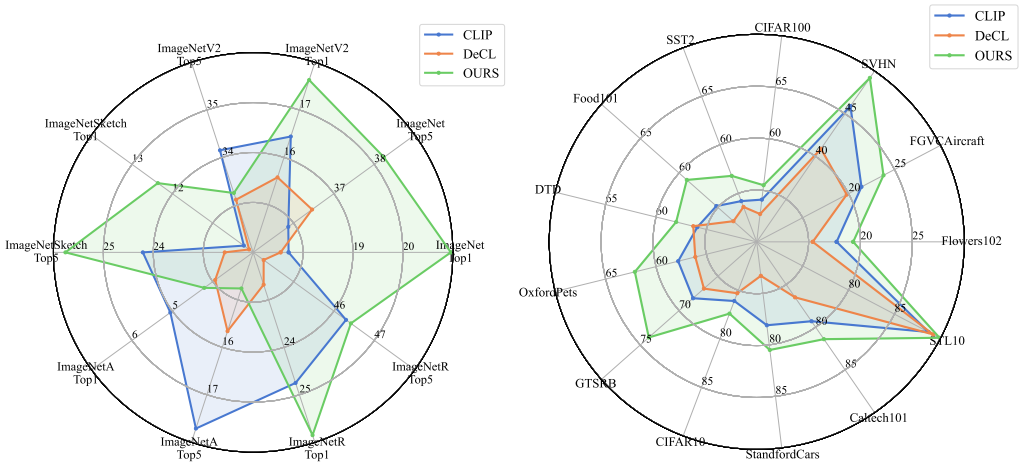
We show in Table 2 the zero-shot transfer learning performance on the Natural Distribution Shift benchmark. We can see that DeCL performs the worst on all four benchmarks, while CLIP baseline demonstrates the best performance on ImageNet-A. CLIP also features decent performance on Top5 accuracy for ImageNetV2. Our method improves the clip baseline performance by 1-2% on Top1 accuracy for three out of four benchmarks (ImageNet-V2, ImageNetSketch, and ImageNet-R), and by around 1% on two out of four benchmarks (ImageNetSketch and ImageNet-R). This indicates that by using our approach to weight training pairs with stochastic approximation we are able to improve the robustness and generalizability of the learned embeddings. Interestingly, our method under-performs CLIP on ImageNet-A, a dataset with adversarial noise. We hypothesize the reason is that correcting noisy pairs in training does not help to combat adversarial noise in data.

4.4 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

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

	Caltech101	**SVHN**	**STL10**	**CIFAR10**	**CIFAR100**	**DTD**	**FGVCAircraft**	**OxfordPets**	**SST2**	**Food101**	**GTSRB**	**StanfordCars**	**Flowers102**	**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	57.9
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	56.1
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**	**60.5**



(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.

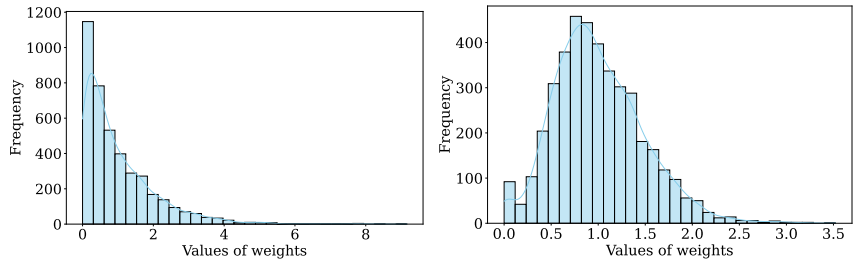
model is then evaluated on the testing dataset. This setting is used to evaluate how well the learned embeddings can generalize to new tasks with further supervision that requires only minimum fine-tuning effort. Following standard setup, we test on 14 standard benchmarks (Krizhevsky, 2009; Russakovsky 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 & 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.

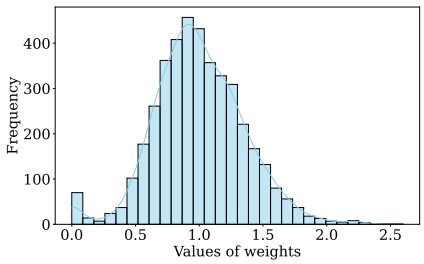
4.5 Analysis

We perform analysis to further investigate our approach. We first test the sensitivity of our method on different sampling parameters. As shown in Section 2.2 and Algorithm 1, there are several hyper-parameters of the two Gamma distributions that need to be determined. Following the same setting as in DeCL we introduce a Gamma prior for u_i's with the shape and rate parameters being $a_u = 1$ and $b_u = 0$. We then choose the parameters for the prior Gamma distribution for $\mathbf{w}$, where we need to determine a_- and b_- for the negative pairs as well as a_+ and b_+ for positive pairs. For simplicity and without loss of generality we set b_- and b_+ to be 0. To reduce the search space, we simply fix

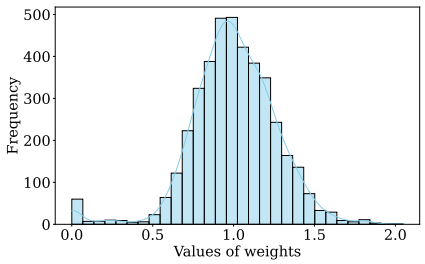


(a) $a_- = 1$

(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.

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_+ and grid search for the best value of a_-. We set $a_+ = 5$ and search over $\{1, 5, 10, 20\}$ for a_-, where a higher value prefers higher weight in prior on negative pairs.

The corresponding results are shown in Table 4. As we can see, the optimal value for a_- is twice of a_+ with the trend that neither higher or lower value brings greater gain. This indicates that slightly higher weights on negative pairs are preferable in the noisy dataset training scenarios while paying too much attention to negative pairs is not desirable as it might mitigate the learning signal from positive pairs. We also visualize the learned distribution sample results in Figure 3. We can observe that by properly setting the hyper-parameters, most of the sampled weights lie around 1 and there are pairs that are associated with much higher weights or lower weights. This observation is expected as our goal is to enable the model to have extra adaptation to automatically determine to lower weights for noisy training pairs.

5 Conclusion

In this paper, we investigate an important yet unnoticeable limitation of standard contrastive learning, where data come with noisy positive-negative pairs. Standard CL cannot handle this problem as it treats each pair equally. As a remedy, we propose a principled solution to CL by reformulating it into a probability framework and introducing random weights for data pairs. With a Bayesian data augmentation technique, the random weights can be efficiently inferred via sampling, and the model parameter can be effectively optimized via stochastic expectation maximization. The effectiveness of our innovative approach has been proven through rigorous evaluations on standard benchmarks, including applications in multi-modal contrastive learning based on the CLIP framework. The results also showcase the wide-ranging applicability and improved robustness of our proposed method. We believe our method is a valuable addition to the literature on contrastive representation learning, which can further boost the performance of state-of-the-art representation learning foundation models with larger datasets.

REFERENCES

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

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

Russell Bent and Pascal Van Hentenryck. Online stochastic and robust optimization. In Michael J. Maher (ed.), *Advances in Computer Science - ASIAN 2004. Higher-Level Decision Making*, pp. 286–300, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-30502-6.

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

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

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

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

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

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

Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Proc. ECCV*, pp. 88–105, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19836-6.

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

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

Jiali Duan, Liqun Chen, Son Tran, Jinyu Yang, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Multi-modal alignment using representation codebook. In *Proc. CVPR*, 2022.

Li Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE TPAMI*, 2006. doi: 10.1109/TPAMI.2006.79.

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

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

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

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

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

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

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. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proc. ICCV*, 2021a.

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

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

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

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

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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. Imagenet large scale visual recognition challenge. *IJCV*, 115, 2015.

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

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

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. Laion-5b: An open large-scale dataset for training next generation image-text models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Proc. NeurIPS*, volume 35, pp. 25278–25294. Curran Associates, Inc., 2022.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1238.

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

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

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

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

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

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

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

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

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

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

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

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

Manzil Zaheer, Michael Wick, Jean-Baptiste Tristan, Alex Smola, and Guy Steele. Exponential stochastic cellular automata for massively parallel inference. In Arthur Gretton and Christian C. Robert (eds.), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 966–975, Cadiz, Spain, 09–11 May 2016. PMLR.

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

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

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

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