Data-Efficient Contrastive Language-Image Pretraining: Prioritizing Data Quality over Quantity

Siddharth Joshi UCLA CS
Arnav Jain UCLA CS
Ali Payani Cisco Systems Inc.
Baharan Mirzasoleiman UCLA CS

Abstract

Contrastive Language-Image Pre-training (CLIP) on large-scale image-caption datasets learns representations that can achieve remarkable zero-shot generalization. However, such models require a massive amount of pre-training data. Improving the quality of the pre-training data has been shown to be much more effective in improving CLIP's performance than increasing its volume. Nevertheless, finding small subsets of training data that provably generalize the best has remained an open question. In this work, we propose the first theoretically rigorous data selection method for CLIP. We show that subsets that closely preserve the cross-covariance of the images and captions of the full data provably achieve a superior generalization performance. Our extensive experiments on ConceptualCaptions3M and ConceptualCaptions12M demonstrate that subsets found by ClipCov achieve over 2.7x and 1.4x the accuracy of the next best baseline on ImageNet and its shifted versions. Moreover, we show that our subsets obtain 1.5x the average accuracy across 11 downstream datasets, of the next best baseline. The code is available at: `https://github.com/BigML-CS-UCLA/clipcov-data-efficient-clip`.

1 INTRODUCTION

Contrastive Language-Image Pretraining (CLIP) has recently showed an impressive success, by enabling zero-shot recognition ability, transferability to downstream tasks, and learning robust representations to distribution shift [27]. CLIP is trained on large image-caption datasets, by maximizing the alignment between paired image-captions representations and minimizing the agreement between image-caption representations of different pairs. Achieving this success, however, requires 1000 times larger datasets than traditional vision datasets like ImageNet [7]. For example, CLIP [27] and ALIGN [15] are trained on 400M and 1B image-captions pairs crawled from the internet. This raises a key question of whether such a massive data is necessary to achieve superior performance and robustness.

There have been recent efforts in answering this question. [10] showed that smaller, more stringently filtered datasets can lead to models that generalize better than larger datasets coming from the same pool. For example, 1.4B images-caption pairs with highest similarity to ImageNet images, and with a high image-caption similarity outperform the full 12.8B data for zero-shot classification on ImageNet. While the effectiveness of such simple filtering strategies confirms the importance of the data quality for training CLIP, such strategies cannot further improve the data efficiency of language-image pretraining. In fact, it is not clear how one can select *small subsets* of the training data that *provably* generalize best, when trained on.

Finding small image-caption subsets with superior generalizability is indeed very challenging and demands fundamental understanding of the representation learning mechanism of CLIP. Indeed, the complex multimodal nature of CLIP makes existing data selection techniques inapplicable. Supervised data selection techniques that select examples based on per-example gradient [33, 26], loss [25], or entropy of the predictions [6] are not applicable to CLIP. This is because the contrastive CLIP loss and its gradient depend on the entire dataset and excluding any example affects the gradient of all the examples. The recent data selection technique of [16] for unimodal contrastive learning, which finds images with central representations, also does not generalize to the multimodal scenario. Data selection for CLIP is

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

inherently more complicated, due to the interaction between the image and text modalities.

In this work, we address the above challenge for the first time. We rely on recent theoretical results of [24] that showed that the CLIP representations are determined by the cross-covariance matrix of the image-caption data. We show that the subsets that closely capture the cross-covariance of the image-caption pairs in the data can guarantee similar zero-shot generalization performance for CLIP.

We confirm the effectiveness and scalability of our proposed technique through extensive experiments on the Conceptual Caption (CC) 3M [29] and CC 12M [4] datasets. We show that our subsets (of sizes 5% - 50%) outperform equal size subsets found using several CLIP data filtering baselines, including CLIP score [10], C-RHO [19], SemDeDup [1] and random selection. The subsets selected by ClipCov achieve over 2.7x and 1.4x the accuracy of the next best baseline on ImageNet and its shifted versions [2, 8, 9, 13, 28, 30], respectively. Additionally, we demonstrate that our selected subsets obtain 1.5x the accuracy of the next best baseline, across 11 different downstream datasets.

2 RELATED WORK

Multimodal Contrastive Learning Recently, Contrastive Language-Image Pre-training (CLIP) on large datasets comprised of paired images and captions has shown remarkable zero-shot generalization performance and transferability to a variety of downstream tasks. In particular, CLIP [27] and ALIGN [15] trained on 400M/1B image-caption pairs achieve comparable accuracy to SOTA supervised learning across several tasks, without the need for any further training. Several recent studies aimed to improve the data-efficiency and performance of CLIP via data augmentation on image and text modalities [17, 23], and imposing geometrically consistency in the image and text space [11].

Multimodal Contrastive Learning Theory A few recent works have studied dynamics of multimodal contrastive learning. [35] extends the results of [12], which showed the equivalence between the matrix factorization objective and the spectral contrastive loss, to the spectral multimodal contrastive loss. [24] showed that for linear models, each step of loss minimization by gradient descent can be seen as performing SVD on a contrastive cross-covariance matrix. We utilize the theory of [24] to characterize the subsets that contribute the most to MMCL and guarantee superior generalization performance for CLIP.

Data Filtering for Multi-Modal Contrastive Learning Large image-caption datasets crawled from the internet often contain image-caption pairs that are uninformative, or contain unreliable or wrong captions. Hence, such datasets are often filtered before being used for training. Several data filtering methods have been proposed recently. Such methods often use a pre-trained CLIP model to filter examples based on their similarity of image-caption representations [10, 19]. Some other methods [19, 34] address dataset-specific problems, e.g. the presence of text in large number of images, and do not yield useful subsets on other datasets. While data filtering methods are essential to filter potentially wrong or irrelevant examples, they cannot find generalizable subsets from filtered datasets. Another recently proposed method aims to reduce the redundancy in the dataset by eliminating examples with similar image representations [1]. However, this method too drastically fails to find the most generalizable subsets from large image-caption datasets, as we will confirm experimentally.

Data Selection for Supervised and Self-supervised Learning Data efficiency in supervised learning has been the subject of extensive research, as evidenced by a long line of work [6, 20, 26, 33]. However, applying these techniques directly to MMCL is not possible due to the absence of labels. Additionally, loss-based methods may not be suitable for MMCL because the loss of examples in MMCL depends on all examples in a batch. While data-efficiency for unimodal contrastive learning has been studied before in [16], it cannot be transferred to multimodal learning due to the fundamental differences in data i.e. no augmentations and paired data from different modalities.

3 PROBLEM FORMULATION

Data Distribution Let $\mathcal{D} = \{(x^i_{\mathcal{V}}, x^i_{\mathcal{L}})\}_{i \in V}$ be a set of $n = |V|$ image-caption pairs i.e. the full training data available to us, drawn from K latent classes i.e. $V = \cup_{k \in [K]} V_k$, where $x^i_{\mathcal{V}}$ denotes the image and $x^i_{\mathcal{L}}$ denotes the caption of the i-th example. Moreover, let $\mathcal{X}_{\mathcal{V}}$ be the set of images and $\mathcal{X}_{\mathcal{L}}$ be the set of captions in $\mathcal{D}$. To model the notion that paired image-captions describe the same underlying object, let image-caption pair $(x^i_{\mathcal{V}}, x^i_{\mathcal{L}}) \in \mathcal{D}$ be generated as follows:

$$x^i_{\mathcal{V}} = T_{\mathcal{V}}(u^i + \epsilon_{\mathcal{V}}) \qquad x^i_{\mathcal{L}} = T_{\mathcal{L}}(u^i + \epsilon_{\mathcal{L}}), \qquad (1)$$

where $u^i \in \mathbb{R}^d$ is the shared underlying feature vector for example i; $T_{\mathcal{V}} : \mathbb{R}^d \to \mathbb{R}^{d_{\mathcal{V}}}$ and $T_{\mathcal{L}} : \mathbb{R}^d \to \mathbb{R}^{d_{\mathcal{L}}}$ are the mappings from underlying feature space to the vision and language data spaces; and $\epsilon_{\mathcal{V}}, \epsilon_{\mathcal{L}}$ are the noise in underlying features for vision and language, respectively. We refer to $\overline{u}^i_{\mathcal{V}} = u^i + \epsilon^i_{\mathcal{V}}$ and $\overline{u}^i_{\mathcal{L}} = u^i + \epsilon^i_{\mathcal{L}}$ as the noisy underlying feature for the image and caption of example i. The underlying feature u^i, for each image-caption pair is sampled independently

of other pairs and of the noise $\epsilon_{\mathcal{V}}, \epsilon_{\mathcal{L}}$. Additionally, we assume $\forall i, \|\overline{u}_{\mathcal{V}}\|^i, \|\overline{u}_{\mathcal{L}}\|^i$ and $\|T_{\mathcal{V}}\|, \|T_{\mathcal{L}}\|$ is ≤ 1.

The shared underlying feature helps us capture the notion that paired image-captions represent the same underlying object (feature) e.g. '`a dog`'. The noise in the data distribution allows us to model both the occurrence of mismatched pairs e.g. an image of '`a dog`' matched with caption '`a cat`' as well as noise in data space for both images and texts e.g. an image of '`a dog with a cat in the background`' paired the caption '`a dog`' or the caption '`a dog with a cat`' paired with an image of '`a dog`'.

Contrastive Language-Image Pre-training (CLIP) CLIP is composed of a vision encoder $f_{\mathcal{V}} : \mathbb{R}^{d_{\mathcal{V}}} \to \mathbb{R}^r$ and a language encoder $f_{\mathcal{L}} : \mathbb{R}^{d_{\mathcal{L}}} \to \mathbb{R}^r$ that map input data in vision and language data space into a shared r-dimensional representation space, respectively. The vision and language encoders are trained by maximizing the representation similarity of paired image-captions and minimizing that of unpaired image-captions in every mini-batch, using the following multimodal contrastive loss:

$$\begin{aligned}&\mathcal{L}_{\text{CLIP}}(f_{\mathcal{V}}, f_{\mathcal{L}}) = \\ &- \mathop{\mathbb{E}}_{x_{\mathcal{V}}, x_{\mathcal{L}} \sim \mathcal{D}} \log \frac{\exp(f_{\mathcal{V}}(x_{\mathcal{V}})^\top f_{\mathcal{L}}(x_{\mathcal{L}}))}{\mathbb{E}_{x_{\mathcal{L}}^- \sim \mathcal{X}_{\mathcal{L}}} \exp(f_{\mathcal{V}}(x_{\mathcal{V}})^\top f_{\mathcal{L}}(x_{\mathcal{L}}^-))} \\ &- \mathop{\mathbb{E}}_{x_{\mathcal{V}}, x_{\mathcal{L}} \sim \mathcal{D}} \log \frac{\exp(f_{\mathcal{V}}(x_{\mathcal{V}})^\top f_{\mathcal{L}}(x_{\mathcal{L}})}{\mathbb{E}_{x_{\mathcal{V}}^- \sim \mathcal{X}_{\mathcal{V}}} \exp(f_{\mathcal{V}}(x_{\mathcal{V}}^-)^\top f_{\mathcal{L}}(x_{\mathcal{L}}))}. \end{aligned} \tag{2}$$

For simplicity of theoretical analysis, we consider linear encoders where $f_{\mathcal{V}}(x_{\mathcal{V}}) = F_{\mathcal{V}} \cdot x_{\mathcal{V}}$ and $f_{\mathcal{L}}(x_{\mathcal{L}}) = F_{\mathcal{L}} \cdot x_{\mathcal{L}}$ where $F_{\mathcal{V}} \in \mathbb{R}^{r \times d_{\mathcal{V}}}$ and $F_{\mathcal{L}} \in \mathbb{R}^{r \times d_{\mathcal{L}}}$, used widely across machine learning literature [24, 14, 31]. Additionally, we use the linear multimodal contrastive loss used in [24]:

$$\begin{aligned}\mathcal{L}(F_{\mathcal{V}}, F_{\mathcal{L}}) = &-\frac{1}{2n(n-1)} \sum_{i \in V} \sum_{\substack{j \in V \\ j \neq i}} (A_{ij} - A_{ii}) \\ &-\frac{1}{2n(n-1)} \sum_{i \in V} \sum_{\substack{j \in V \\ j \neq i}} (A_{ji} - A_{ii}) + \frac{\rho}{2} \|F_{\mathcal{V}}^\top F_{\mathcal{L}}\|_F^2,\end{aligned} \tag{3}$$

where $A_{ij} \coloneqq (F_{\mathcal{V}} x_{\mathcal{V}}^i)^\top (F_{\mathcal{L}} x_{\mathcal{L}}^j)$. [24] shows that both the CLIP loss and the linear multimodal contrastive loss can be derived from a generalized form of the multimodal contrastive loss i.e. aligning representations of paired image-captions and separating representations of unpaired image-captions.

Note that we only use linear encoders and the linear multi-modal contrastive loss function in our theoretical analysis; the experiments in Section 5 are conducted with non-linear encoders and the CLIP loss in Eq. (2).

Zero-Shot Classification After training, the model is evaluated via zero-shot classification on different downstream image classification tasks. A downstream task $\mathcal{D}_{\mathcal{Y}}$ is defined as a classification task on unseen data from a set of $\mathcal{Y}$ classes. For zero-shot classification on downstream task $\mathcal{D}_{\mathcal{Y}}$, we use the language encoder $f_{\mathcal{L}}$ to encode the label of each class $y \in \mathcal{Y}$; using a set of pre-engineered templates, e.g. '`A photo of a {label}`' to create several captions representing '`{label}`' [27]. Then, the classification of an example $x_{\mathcal{V}}$ is $zs_{f_V, f_L}(x_{\mathcal{V}}) = \arg\max_{k \in \mathcal{D}_{\mathcal{Y}}} \frac{f_{\mathcal{V}} x_{\mathcal{V}} \cdot z_k}{\|f_{\mathcal{V}} x_{\mathcal{V}}\| \|z_k\|}$, where $z_k = \mathbb{E}_{x_{\mathcal{L}} s.t. y(x_{\mathcal{L}}) = k}[f_{\mathcal{L}}(x_{\mathcal{L}})]$ is the average representation of templates obtained using $f_{\mathcal{L}}$ for class label k. That is an example $x_{\mathcal{V}}$ is classified by the closest (average) template representation. The zero-shot error of $f_{\mathcal{V}}, f_{\mathcal{L}}$ is defined as the fraction of misclassified examples using the trained vision and language encoders f_V, f_L:

$$\mathcal{E}_{zs}(f_{\mathcal{V}}, f_{\mathcal{L}}) := \mathbb{P}_{x_{\mathcal{V}} \sim \mathcal{D}_{\mathcal{Y}}} \left[y(x_{\mathcal{V}}) \neq zs_{f_V, f_L}(x_{\mathcal{V}}) \right]. \tag{4}$$

Finding Generalizable Multimodal Subsets Our goal is to find a subset of training image-caption data $S \subseteq V$ of size at most $n_s \geq |S|$, such that encoders trained on the subset achieve similar generalization, across downstream tasks using zero-shot evaluation, to encoders trained on the full training data V. To do so, we formulate the problem as finding a subset S such that the encoders learnt on the subset closely approximate the encoders learnt on the full training data V:

$$S^* = \mathop{\arg\min}_{S \subseteq V, |S| \leq n_s} \left\| F_{\mathcal{V}}^S - F_{\mathcal{V}} \right\| + \left\| F_{\mathcal{L}}^S - F_{\mathcal{L}} \right\| \tag{5}$$

where $F_{\mathcal{V}}^S, F_{\mathcal{L}}^S$ are the vision and language encoders learnt on the subset S and $F_{\mathcal{V}}, F_{\mathcal{L}}$ are the encoders learnt on the the full training data V.

4 FINDING THE MOST GENERALIZABLE SUBSETS

In this section, we first theoretically characterize how well the encoders learnt on an arbitrary subset S approximate the encoders learnt on the full (training) data V. Then, we present CLIPCOV, our algorithm for efficiently finding S^*, the most generalizable subset, from a massive corpus of image-caption pairs.

To do so, we rely on the recent theoretical results showing that the training dynamics on the full data V are determined by the cross-covariance matrix of all the image-caption pairs in the dataset [24]. The centered cross-covariance matrix of the full data $C_{\mathcal{D}}^V$ is defined as follows:

$$C_{\mathcal{D}}^V := \frac{1}{|V|} \sum_{i \in V} (x_{\mathcal{V}}^i - \mu_{x_{\mathcal{V}}})(x_{\mathcal{L}}^i - \mu_{x_{\mathcal{L}}})^\top, \tag{6}$$

where $\mu_{x_\mathcal{V}} = \mathbb{E}_{x_\mathcal{V} \in \mathcal{X}_\mathcal{V}} x_\mathcal{V}$ is the center of vision data and $\mu_{x_\mathcal{L}} = \mathbb{E}_{x_\mathcal{L} \in \mathcal{X}_\mathcal{L}} x_\mathcal{L}$ is the center of language data. [1] The cross-covariance matrix for image-caption data captures the covariance between paired image-captions.

The linear loss function in Eq. (3) can be rewritten as the SVD objective function:

$$\mathcal{L}(F_\mathcal{V}, F_\mathcal{L}) = -\operatorname{Tr}(F_\mathcal{V} C_\mathcal{D}^V F_\mathcal{L}^\top) + \frac{\rho}{2}\|F_\mathcal{V}^\top F_\mathcal{L}\|_F^2. \quad (7)$$

Likewise, dynamics of training on the subset is determined by the cross-covariance of the subset:

$$\mathcal{L}(F_\mathcal{V}^S, F_\mathcal{L}^S) = -\operatorname{Tr}(F_\mathcal{V}^S C_\mathcal{D}^S {F_\mathcal{L}^S}^\top) + \frac{\rho}{2}\|{F_\mathcal{V}^S}^\top F_\mathcal{L}^S\|_F^2, \quad (8)$$

where $C_\mathcal{D}^S$ is the data cross-covariance matrix of the subset S.

Hence, we see that if $C_\mathcal{D}^S$, the cross-covariance of the subset S, closely approximates $C_\mathcal{D}^V$, the cross-covariance of the full data, by minimizing the contrastive multimodal loss, the encoders learnt on the subset S will be similar to the encoders learnt on the full data V.

4.1 Preserving the Cross-Covariance of Data

To preserve the cross-covariance of the full data, we can preserve the cross-covariance of noisy image and caption underlying features. Let

$$\overline{C_\mathcal{U}^V} = \frac{1}{|V|}\sum_{i\in V}(\overline{u}_\mathcal{V}^i - \mathop{\mathbb{E}}_{i\in V}\overline{u}_\mathcal{V}^i)(\overline{u}_\mathcal{L}^i - \mathop{\mathbb{E}}_{i\in V}\overline{u}_\mathcal{L}^i)^\top,$$
$$\overline{C_\mathcal{U}^S} = \frac{1}{|S|}\sum_{i\in S}(\overline{u}_\mathcal{V}^i - \mathop{\mathbb{E}}_{i\in S}\overline{u}_\mathcal{V}^i)(\overline{u}_\mathcal{L}^i - \mathop{\mathbb{E}}_{i\in S}\overline{u}_\mathcal{L}^i)^\top,$$

be the cross-covariance of noisy underlying features for the full data V and subset S, respectively. Then we have that,

$$\|C_\mathcal{D}^V - C_\mathcal{D}^S\| = \left\|T_\mathcal{V}\overline{C_\mathcal{U}^V}T_\mathcal{L}^\top - T_\mathcal{V}\overline{C_\mathcal{U}^S}T_\mathcal{L}^\top\right\| \quad (9)$$
$$\leq \|T_\mathcal{V}\|\,\|T_\mathcal{L}\|\left\|\overline{C_\mathcal{U}^V} - \overline{C_\mathcal{U}^S}\right\| \quad (10)$$
$$\leq \left\|\overline{C_\mathcal{U}^V} - \overline{C_\mathcal{U}^S}\right\|, \quad (11)$$

where the last inequality holds since $\|T_\mathcal{V}\| = \|T_\mathcal{L}\| \leq 1$.

We see that if the subset S preserves the cross-covariance of noisy underlying feature of V, then it also preserves the data cross-covariance of the full data.

To preserve the cross-covariance matrix of underlying features in Eq. (6), we need to (1) preserve the centers of images $\mu_{x_\mathcal{V}}$ and captions $\mu_{x_\mathcal{L}}$ in every latent class; and (2) preserve the cross-covariance of examples in every latent class by selecting image-caption pairs that are centrally located in different subpopulations of the latent class and represent its different subgroups.

Next, we show how we can find a subset with the above two properties to closely preserve the cross-covariance of the full data.

Preserving Centers of Latent Classes

Let $\mu_\mathcal{V}^{S_k} = \mathbb{E}_{i\in S_k}\overline{u}_\mathcal{V}^i$ and $\mu_\mathcal{L}^{S_k} = \mathbb{E}_{i\in S_k}\overline{u}_\mathcal{L}^i$ be the centers of the noisy underlying features for images and captions in subset S_k selected from latent class k, respectively. Likewise, let $\mu_\mathcal{V}^{V_k} = \mathbb{E}_{i\in V_k}\overline{u}_\mathcal{V}^i$ and $\mu_\mathcal{L}^{V_k} = \mathbb{E}_{i\in V_k}\overline{u}_\mathcal{L}^i$ be the centers of the noisy underlying features for images and captions of latent class k in full data, respectively.

We now bound the error in preserving the centers of images and captions:

$$\begin{aligned}
&\left\|\mu_\mathcal{V}^{S_k} - \mu_\mathcal{V}^{V_k}\right\| + \left\|\mu_\mathcal{L}^{S_k} - \mu_\mathcal{L}^{V_k}\right\| \\
&\quad= \left\|\frac{1}{|S_k|}\sum_{i\in S_k}\overline{u}_\mathcal{V}^i - \frac{1}{|V_k|}\sum_{j\in V_k}\overline{u}_\mathcal{V}^j\right\| \\
&\qquad+ \left\|\frac{1}{|S_k|}\sum_{i\in S_k}\overline{u}_\mathcal{L}^i - \frac{1}{|V_k|}\sum_{j\in V_k}\overline{u}_\mathcal{L}^j\right\| \\
&\quad\leq \left\|\frac{1}{|S_k|}\sum_{i\in S_k}\overline{u}_\mathcal{V}^i - \frac{1}{|V_k|}\sum_{j\in V_k}\overline{u}_\mathcal{L}^j\right\| \\
&\qquad+ \left\|\frac{1}{|S_k|}\sum_{i\in S_k}\overline{u}_\mathcal{L}^i - \frac{1}{|V_k|}\sum_{j\in V_k}\overline{u}_\mathcal{V}^j\right\| \\
&\qquad+ \underbrace{2\left\|\mu_\mathcal{V}^{V_k} - \mu_\mathcal{L}^{V_k}\right\|}_{\text{alignment of full data centers}}. \quad (12)
\end{aligned}$$

The alignment of centers of latent class V_k in full data refers to the similarity of the image and caption centers of the noisy underlying features and is independent of the subset S_k. Moreover, with sufficiently small noise in underlying feature, full data class centers are nearly identical as the true underlying feature is shared across images and captions within a latent class. Hence we get:

$$\begin{aligned}
&\left\|\mu_\mathcal{V}^{S_k} - \mu_\mathcal{V}^{V_k}\right\| + \left\|\mu_\mathcal{L}^{S_k} - \mu_\mathcal{L}^{V_k}\right\| \\
&\lessapprox \frac{1}{|S_k|}\frac{1}{|V_k|}\left\|\sum_{\substack{i\in S_k\\ j\in V_k}}\overline{u}_\mathcal{V}^i - \overline{u}_\mathcal{L}^j\right\| + \frac{1}{|S_k|}\frac{1}{|V_k|}\left\|\sum_{\substack{i\in S_k\\ j\in V_k}}\overline{u}_\mathcal{L}^i - \overline{u}_\mathcal{V}^j\right\| \quad (13)
\end{aligned}$$

That is, we can minimize Eq. (13) to find a subset that

[1] Since $|V|$ is large, we replace $|V|-1$ with $|V|$ for simplicity.

preserves image and caption centers, for each latent class k.

However, in practice, we do not have access to these noisy underlying features. Instead, we approximate them using the representations of a vision and language encoder trained with the multimodal contrastive loss on the full data. We refer to these encoders as the proxy vision encoder $f^p_{\mathcal{V}}$ and the proxy language encoder $f^p_{\mathcal{L}}$. Vision and language encoders trained on the full data recover the corresponding noisy underlying features, up to orthogonal transformation [24]. Using the proxy encoders, we now introduce the notion of cross-modal similarity to help us minimize Eq. (13).

Definition 4.1 (**Cross-Modal Similarity**). We define the cross-modal similarity between any two image-caption pairs $i, j \in V$ as the sum of the cosine similarities between the representations of the image of i with caption of j and caption of i with image of j. Formally, $\forall i, j \in V$, we have:

$$\text{sim}(i,j) = \left\langle f^p_{\mathcal{V}}(x^i_{\mathcal{V}})^\top, f^p_{\mathcal{L}}(x^j_{\mathcal{L}}) \right\rangle + \left\langle f^p_{\mathcal{V}}(x^j_{\mathcal{V}})^\top, f^p_{\mathcal{L}}(x^i_{\mathcal{L}}) \right\rangle.$$

Since, the norm of underlying feature vectors ≤ 1 and orthogonal transformations preserve inner products, we can minimize Eq. (13) by maximizing the cross-modal similarity of the corresponding examples: $\sum_{\substack{i\in S_k \\ j\in V_k}} \text{sim}(i,j)$.

Thus, we can preserve centers for the full data by maximizing the following objective:

$$\sum_{k\in[K]} \frac{1}{|V_k|} \sum_{\substack{i\in S_k \\ j\in V_k}} \text{sim}(i,j), \tag{14}$$

where we normalize the objective for latent class k by the size of the latent class $|V_k|$ to prevent larger classes from dominating the objective.

In practice, the data within latent classes is often imbalanced. To prevent large subgroups within latent classes from dominating the objective, we penalize the similarity between selected examples to encourage diversity in the subset. Thus, we preserve the centers and alignment of centers, while ensuring diversity of the selected examples, using the following objective:

$$F_{\text{class}}(S) := \sum_{k\in K} \frac{1}{|V_k|} \Big(\sum_{\substack{i\in S_k \\ j\in V_k}} \text{sim}(i,j) - \frac{1}{2} \sum_{\substack{i\in S_k \\ j\in S_k}} \text{sim}(i,j) \Big).$$

Preserving Cross-covariance with CLIP Score

Next, we aim to select a subset of examples that capture the cross-covariance between image-caption pairs within every latent class. To do so, we need to find examples that are centrally located in different subpopulations within every latent class and represent its different subgroups. By minimizing the CLIP loss on such subsets and aligning their image-caption pairs, we also align other image-caption pairs in the corresponding subgroups within the latent class.

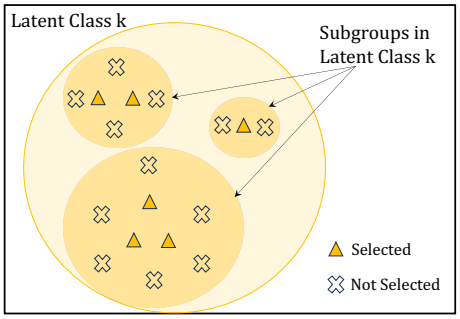


Figure 1: Visualization of examples selected by $F_{\text{class}}(S) + F_{\text{self}}(S)$ in cross-modal similarity space. CLIPCOV selects central examples that are representative of different subgroups in every latent class.

To effectively find subsets that capture the covariance within every latent class, we use a pre-trained CLIP model to find examples that represent different subgroups of the data. Having a pre-trained model with the CLIP loss, the above examples can be efficiently identified as pairs with largest cross-modal similarity between their own image and captions, using the following CLIP score objective:

$$F_{\text{self}}(S) = \sum_{i\in S} \text{sim}(i,i).$$

Examples found using the above CLIP score objective are similar to many other pairs in their latent class, and are most centrally located in different subgroups.

CLIP score is effective to find larger subsets. Effectively, training with the CLIP loss aligns different image-caption pairs. In doing so, subgroups of image-caption pairs that have large cross-modal similarity to each other get close in the representation space. This is because groups of similar examples together introduce a large gradient during the training to pull their images and captions together. Hence, the most central example in each subgroup will have the largest cross modal similarity between its image and caption. Therefore, CLIP score can efficiently find examples that are centrally located in different subgroups of the training data. If the pre-training data is large and diverse, examples found by CLIP score obtain a superior performance on various downstream tasks, as they contain the most central examples in various subgroups of the data. Nevertheless, when the subset size is small, such examples cannot capture the center of latent classes

Algorithm 1 CLIPCOV

Input: Dataset V, Subset size n_s, proxy encoders: $f^p_{\mathcal{V}}$ and $f^p_{\mathcal{L}}$ to calculate $F_{\text{ClipCov}}(S)$
Output: Subset S
$\{V_1, ..., V_K\} \leftarrow$ approximate latent classes
$S \leftarrow \{\}$
$F(S) := F_{\text{ClipCov}}(S)$ ▷ Eq. (16)
$S \leftarrow \emptyset$ ▷ **Greedy**
while $|S| \leq n_s$ **do**
  $e \leftarrow \arg\max_{e \in V \setminus S} F(e|S)$
  $S \leftarrow S \cup \{e\}$
end while
$S_1 \leftarrow \emptyset, S_2 \leftarrow S$ ▷ **Double Greedy**
for $e \in S$ **do**
  $a \leftarrow F(e|S_1)$
  $b \leftarrow F(S_2 \setminus \{e\}) - F(S_2)$
  if $a \geq b$ **then**
    $S_1 \leftarrow S_1 \cup \{e\}$
  else
    $S_2 \leftarrow S_2 \setminus \{e\}$
  end if
end for
return S_1 or equivalently S_2

accurately and $F_{\text{class}}(S)$ is crucial to achieve superior generalization performance.

Hence, we can preserve the cross-covariance of the data by maximizing the following objective $F_{\text{cov}}(S)$:

$$F_{\text{cov}}(S) := F_{\text{class}}(S) + F_{\text{self}}(S).$$

To illustrate what kinds of examples are selected by this objective, we provide a visualization in Fig. 1 which shows the selected examples are similar to all the examples in the latent class, even from smaller subgroups. From Fig. 1, we can see how such a subset is representative of the latent class and thus can capture the cross-covariance within the latent class.

4.2 Deriving the Final Objective for Finding the Most Generalizable Subset

We now discuss three practical considerations that often arise when learning from large vision-language datasets, and account for them in the final objective to find the most generalizable subsets.

Label Centrality for Zero-shot Classification While preserving the cross-covariance within latent classes allows us to ensure that images in a given latent class can correctly be paired with their corresponding captions, zero-shot classification measures similarity of images representations to the text representations of the *labels* of the latent classes. This is highly sensitive to the name of the label being similar to the captions of the corresponding latent class. To explicitly ensure that the selected captions are similar to the labels used, we introduce $F_{\text{label}}(S)$:

$$F_{\text{label}}(S) = \sum_{k \in [K]} \sum_{i \in S_k} \alpha f^p_{\mathcal{L}}(x^i_{\mathcal{L}})^\top f^p_{\mathcal{L}}(y_k) - \sum_{i \in S_k} \alpha \frac{f^p_{\mathcal{L}}(x^i_{\mathcal{L}})^\top f^p_{\mathcal{L}}(y_k)}{|V_k|},$$

where α is the ratio of average cross-modal similarity to the average similarity in text [2] Here, the second term prevents domination of classes with very good similarity to the label. This improves the zero-shot performance on various downstream datasets.

Dealing with Imbalanced Data In practice, when the sizes of latent classes are extremely imbalanced i.e. some latent classes in the training data are much larger than others, this leads to $F_{\text{class}}(S)$ for large latent classes dominating the objective. Hence, we further regularize $F_{\text{class}}(S)$ to avoid only selecting examples from large latent classes by deducting the following regularization term from the objective.

$$F^{\text{reg}}_{\text{class}}(S) = \sum_{k \in K} \frac{1}{|V_k|} \sum_{\substack{i \in S_k \\ j \in V_k}} \frac{\text{sim}(i,j)}{|V_k|},$$

which is approximately the average sum of intra-class cross-modal similarity of the selected subset S.

Penalizing Inter-class Similarity Empirically, we find that ensuring that the examples selected for different latent classes are dissimilar yields more distinguishable representations for latent classes, improving performance across various downstream tasks. Thus, we minimize the average similarity of examples to other latent classes. The following objective, $F_{\text{inter}}(S)$, formalizes this:

$$F_{\text{inter}}(S) := -\sum_{\substack{k_1, k_2 \in [K] \\ k_1 \neq k_2}} \sum_{i \in S_k} \sum_{j \in V_{k_2}} \frac{\text{sim}(i,j)}{|V_{k_2}|},$$

where $\sum_{i \in S_k} \sum_{j \in V_{k_2}} \frac{\text{sim}(i,j)}{|V_{k_2}|}$ is the average cross-modal similarity of image-caption pair i to image-caption pairs in V_{k_2}. In practice, we can compute this average cross-modal similarity efficiently, by first averaging the image-caption representations of latent class k_2 and then computing the cross-modal similarity between examples $i \in V_{k_1}$ and the average image-caption representations of V_{k_2}.

[2] Empirically, we find $\alpha \approx \frac{1}{2}$.

Final Objective Hence, the final objective for finding the most generalizable subset S^* is:

$$S^* \in \underset{S \subseteq V, |S| \leq n_s}{\arg\max} F_{\text{ClipCov}}(S), \quad \text{where} \tag{15}$$

$$F_{\text{ClipCov}}(S) := \tag{16}$$
$$F_{\text{cov}}(S) + F_{\text{label}}(S) - F^{\text{reg}}_{\text{class}}(S) + F_{\text{inter}}(S).$$

4.3 ClipCov: Efficiently Finding the Most Generalizable Subset

Here, we discuss how the proxy representations and latent classes required to solve Problem (16) are obtained. We then present ClipCov and show how it can efficiently find this subset from massive datasets.

Obtaining Proxy Representations We can use any pretrained CLIP as the proxy encoders to determine the proxy representations cross-covariance matrix. The effectiveness of ClipCov is dependent on how closely the proxy representations recover the underlying features of the full data. Hence, we use the open-source pretrained CLIP encoders provided by [27], which are trained on massive amounts of data and obtain impressive zero-shot generalization, thus are likely effectively recover the underlying features of the full data V.

Approximating Latent Classes In practice, we do not have access to latent classes required to solve Problem (16). Instead, we approximately recover latent classes via zero-shot classification using proxy CLIP encoders. For zero-shot classification, we use 1000 labels of ImageNet-1k. This allows finding fine-grained latent classes.

Scaling to Massive Datasets Since $F_{\text{inter}}(S)$ can be computed using the average representations of latent classes, in practice, ClipCov only needs to compute pairwise cross-modal similarities within latent classes. Here, the fine-grained latent classes used also ensure that computing pairwise cross-modal similarities within latent classes is inexpensive.

Maximizing Objective (16) is NP-hard as it requires evaluating an exponential number of subsets. To efficiently find a near-optimal subset, we note that $F_{\text{cov}}(S)$ is non-monotone submodular, and $F_{\text{inter}}(S)$, $F_{\text{label}}(S)$, $F^{\text{reg}}_{\text{class}}(S)$ are modular. Hence, Objective (16) is non-monotone submodular. Thus, we can find a near-optimal subset using algorithms for non-monotone submodular function maximization under a cardinality constraint. To do so, we first use the greedy algorithm to find a subset, and then filter the subset by applying unconstrained submodular maximization [22]. The greedy algorithm starts with the empty set $S_0 = \emptyset$, and at each iteration t, it chooses an element $e \in V$ that maximizes the marginal utility $F(e|S_t) = F(S_t \cup \{e\}) - F(S_t)$. Formally, $S_t = S_{t-1} \cup \{\arg\max_{e \in V} F(e|S_{t-1})\}$. For unconstrained maximization, we use the double-greedy algorithm [3], which initializes $S_1 = \emptyset$ and $S_2 = S_T$, where S_T is the subset found in the final iteration of the greedy algorithm, and calculates $a_e = F(e|S_1)$ and $b_e = F(S_2 \setminus \{e\})$ for all $e \in V$, and then adds examples for which $a_e \geq b_e$ to S_1 and removes examples for which $a_e < b_e$ from S_2 and eventually returns $S_1 = S_2$. The complexity of the greedy algorithm is $\mathcal{O}(nk)$ to find k out of n examples, and can be further speed up using lazy evaluation [21]. The double-greedy applied to the subset has a complexity of $\mathcal{O}(k)$. Hence, the subset can be found efficiently. Algorithm 1 illustrates our pseudocode.

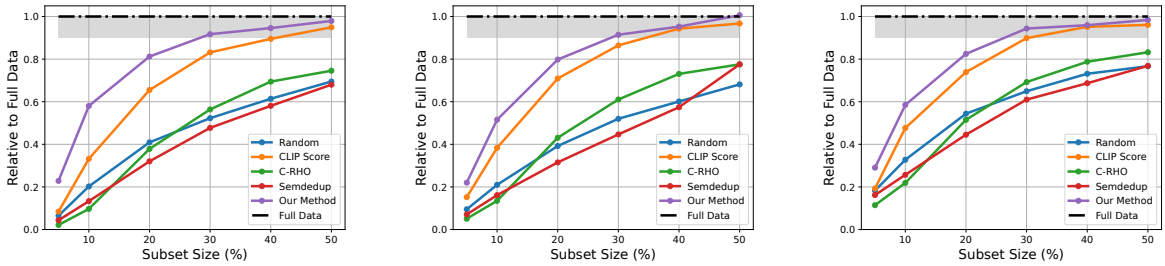
5 EXPERIMENTS

In this section, we compare the zero-shot performance of training on subsets of sizes 5%-50% found by ClipCov and those found by baselines, including C-RHO, SemDeDup, CLIP score and Random selection. Moreover, we conduct an extensive ablation on the various components of ClipCov.

Dataset & Evaluation We use Conceptual Captions 3M and 12M [29] which include 3 and 12 million image-captions pairs, respectively, and have been widely employed for benchmark evaluations in various studies focusing on contrastive language-image pre-training [32, 11, 18]. We evaluate all the methods on downstream tasks proposed by [5] and used in prior work for evaluating CLIP [32, 11, 18]. The exact list of datasets and corresponding accuracies appears in Appendix A.

Training Setup For pre-training, we use an open-source implementation of CLIP, with default ResNet-50 as the image encoder and a Transformer as the text encoder. Each experiment is run with a batch size of 512 for 30 epochs, consistent with [32].

Baselines The data-filtering baselines we consider are: (1) CLIP Score [10], (2) C-RHO [19], (3) SemDeDup [1], and (4) random subsets. CLIP score discard image-caption pairs with the smallest similarity between their image and caption representations, obtained using a pretrained CLIP. C-RHO is an extension to RHO [20] for CLIP. It computes the similarity of paired image-caption representations using a pre-trained CLIP and compares it to the similarity obtained using a model partially trained (for 5 epochs) on the full data. Then, image-captions pairs with the smallest difference between these similarities are discarded. SemDeDup clusters the image representations of examples and then discards examples from each cluster that are most similar to each other. Due to computational constraints, we



(a) ImageNet (b) ImageNet Dist. Shift (c) Avg. over 11 Datasets

Figure 2: Performance across subset of different sizes selected from ConceptualCaptions3M. Gray region indicates accuracy within 90% of that of full data.

Table 1: Performance of 5% and 10% subsets selected from ConceptualCaptions12M

Subset Size	Method	ImageNet	ImageNet Dist. Shift	Avg. over 11 Datasets
5%	CLIP Score	5.10%	4.42%	9.49%
	ClipCov	**13.61%**	**7.99%**	**11.68%**
10%	CLIP Score	11.02%	8.55%	14.69%
	ClipCov	**22.71%**	**12.76%**	**16.87%**

Table 2: Ablation over proxy encoders

Method	ImageNet	ImgNet Shift	Avg.
CLIP Score	5.01%	3.16%	7.35%
ClipCov	**6.70%**	**3.48%**	**9.10%**

only evaluate CLIPScore and ClipCov, the two best performing methods on CC12M.

Zero-Shot Performance Fig. 2 shows that, both specifically on ImageNet and across datasets, ClipCov is able to outperform previous baselines. Moreover, our results demonstrate that all common data-filtering baselines, except CLIP Score, fail to extract generalizable subsets from datasets that are already filtered. This is evidenced by these methods performing worse even than random subsets. In contrast, ClipCov successfully extracts subsets that can preserve the downstream generalization performance on various datasets and outperforms CLIP Score. Moreover, Fig. 2 shows that ClipCov can discard 50% of the data without losing any accuracy, outperforming all baselines. In fact, only 30% of the data is needed for performance within 90% of training on the full data. Table 4 shows that ClipCov achieves over 2.7x and 1.4x the accuracy of CLIP Score (the next best baseline) on ImageNet and its shifted versions. Moreover, it also shows that ClipCov obtains 1.5x the accuracy of CLIP Score, across 11 downstream tasks. Table 1 verifies that ClipCov scales to larger datasets as well, with ClipCov's subsets, achieving over 2.5x and 1.9x the accuracy of CLIP Score subsets, on ImageNet and its shifted versions, as well as nearly 1.25x the average accuracy over 11 downstream tasks.

Ablation Study Table 3 ablates over the objective and shows that $F_{\text{inter}}(S)$, $F_{\text{label}}(S)$ and $F^{\text{reg}}_{\text{class}}(S)$ are all useful practical additions to $F_{\text{cov}}(S)$. Table 2 compares the performance of ClipCov and CLIP Score where the similarities are computed using a model trained on ConceptualCaptions3M rather than the open-source CLIP provided in [27]. These results show that ClipCov can outperform prior art, regardless of choice of proxy model. The drop in performance for both CLIP Score and ClipCov when compared to the subsets in Table 4, shows that using cross-modal similarities from encoders trained on more diverse and balanced data (e.g. CLIP from [27]) is beneficial to both CLIP Score and ClipCov.

6 CONCLUSION

We identified subsets of examples that contribute the most to contrastive language image pre-training (CLIP). Theoretically, we characterized the most beneficial subsets with rigorous generalization guarantees for downstream zero-shot performance, as those that preserve the cross-covariance matrix of the full training data. Empirically, we compare the performance of our method to baselines and show that it achieves over 2.7x and 1.4x the accuracy of the next best baseline, on ImageNet and distribution shifted versions of ImageNet. Moreover, we also show ClipCov achieve 1.5x the average accuracy across 11 downstream datasets. To conclude, ClipCov enables data-efficient CLIP pre-training on massive web-scale datasets.

Table 3: Ablation over objective for 10% subset selected from ConceptualCaption3M

Method	ImageNet	ImageNet Dist. Shift	Avg. over 11 Datasets
$F_{\text{cov}}(S)$	8.74%	5.17%	10.82%
$F_{\text{cov}}(S) + F_{\text{inter}}(S)$	9.00%	5.30%	10.96%
$F_{\text{cov}}(S) + F_{\text{inter}}(S) - F^{\text{reg}}_{\text{class}}(S)$	8.94%	5.10%	11.29%
$F_{\text{cov}}(S) + F_{\text{inter}}(S) + F_{\text{label}}(S)$	10.87%	5.73%	11.27%
ClipCov	**11.33%**	**5.97%**	**12.64%**

Table 4: Performance across subset of different sizes selected from ConceptualCaptions3M

Subset Size	Method	ImageNet	ImageNet Dist. Shift	Avg. over 11 Datasets
5%	Random	1.27%	1.10%	3.95%
	C-RHO	0.42%	0.59%	2.47%
	SemDeDup	0.85%	0.82%	3.50%
	CLIP Score	1.65%	1.76%	4.16%
	ClipCov	**4.46%**	**2.55%**	**6.28%**
10%	Random	3.95%	2.43%	7.08%
	C-RHO	1.89%	1.56%	4.73%
	SemDeDup	2.60%	1.87%	5.55%
	CLIP Score	6.48%	4.44%	10.30%
	ClipCov	**11.33%**	**5.97%**	**12.64%**
20%	Random	7.99%	4.54%	11.75%
	C-RHO	7.39%	4.99%	11.13%
	SemDeDup	6.25%	3.65%	9.62%
	CLIP Score	12.79%	8.21%	15.87%
	ClipCov	**15.86%**	**9.24%**	**17.82%**
30%	Random	10.21%	6.02%	14.03%
	C-RHO	11.01%	7.07%	14.96%
	SemDeDup	9.32%	5.17%	13.18%
	CLIP Score	16.24%	10.01%	19.42%
	ClipCov	**17.91%**	**10.59%**	**20.39%**
40%	Random	11.99%	6.96%	15.80%
	C-RHO	13.56%	8.46%	17.02%
	SemDeDup	11.34%	6.65%	14.84%
	CLIP Score	17.48%	10.92%	20.56%
	ClipCov	**18.47%**	**11.03%**	**20.73%**
50%	Random	13.54%	7.89%	16.57%
	C-RHO	14.56%	8.98%	17.98%
	SemDeDup	13.28%	8.98%	16.60%
	CLIP Score	18.54%	11.20%	20.76%
	ClipCov	**19.12%**	**11.65%**	**21.26%**

Acknowledgements

This research was partially supported by the National Science Foundation CAREER Award 2146492 and Cisco Systems.

References

[1] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S. Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication, 2023.

[2] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[3] Niv Buchbinder, Moran Feldman, Joseph Seffi, and Roy Schwartz. A tight linear time (1/2)-approximation for unconstrained submodular maximization. *SIAM Journal on Computing*, 44(5):1384–1402, 2015.

[4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts, 2021.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020.

[6] C Coleman, C Yeh, S Mussmann, B Mirzasoleiman, P Bailis, P Liang, J Leskovec, and M Zaharia. Selection via proxy: Efficient data selection for deep learning. In *International Conference on Learning Representations (ICLR)*, 2020.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] Josip Djolonga, Jessica Yung, Michael Tschannen, Rob Romijnders, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Matthias Minderer, Alexander D'Amour, Dan Moldovan, Sylvain Gelly, Neil Houlsby, Xiaohua Zhai, and Mario Lucic. On robustness and transferability of convolutional neural networks, 2021.

[10] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets, 2023.

[11] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A. Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining, 2022.

[12] Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss, 2022.

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

[14] Wenlong Ji, Zhun Deng, Ryumei Nakada, James Zou, and Linjun Zhang. The power of contrast for feature learning: A theoretical analysis, 2021.

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

[16] Siddharth Joshi and Baharan Mirzasoleiman. Data-efficient contrastive self-supervised learning: Most beneficial examples for supervised learning contribute the least. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15356–15370. PMLR, 23–29 Jul 2023.

[17] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pretraining paradigm. In *International Conference on Learning Representations*, 2021.

[18] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pretraining paradigm. In *International Conference on Learning Representations*, 2022.

[19] Pratyush Maini, Sachin Goyal, Zachary C. Lipton, J. Zico Kolter, and Aditi Raghunathan. T-mars: Improving visual representations by circumventing text feature learning, 2023.

[20] Sören Mindermann, Jan Brauner, Muhammed Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N. Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. Prioritized training on points that are learnable, worth learning, and not yet learnt, 2022.

[21] Michel Minoux. Accelerated greedy algorithms for maximizing submodular set functions. In

Optimization Techniques: Proceedings of the 8th IFIP Conference on Optimization Techniques Würzburg, September 5–9, 1977, pages 234–243. Springer, 2005.

[22] Baharan Mirzasoleiman, Ashwinkumar Badanidiyuru, and Amin Karbasi. Fast constrained submodular maximization: Personalized data summarization. In *International Conference on Machine Learning*, pages 1358–1367. PMLR, 2016.

[23] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pages 529–544. Springer, 2022.

[24] Ryumei Nakada, Halil Ibrahim Gulluk, Zhun Deng, Wenlong Ji, James Zou, and Linjun Zhang. Understanding multimodal contrastive learning and incorporating unpaired data. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 4348–4380. PMLR, 25–27 Apr 2023.

[25] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training, 2023.

[26] Omead Pooladzandi, David Davini, and Baharan Mirzasoleiman. Adaptive second order coresets for data-efficient machine learning, 2022.

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[28] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet?, 2019.

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

[30] Haohan Wang, Songwei Ge, Eric P. Xing, and Zachary C. Lipton. Learning robust global representations by penalizing local predictive power, 2019.

[31] Yihao Xue, Siddharth Joshi, Eric Gan, Pin-Yu Chen, and Baharan Mirzasoleiman. Which features are learnt by contrastive learning? On the role of simplicity bias in class collapse and feature suppression. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 38938–38970. PMLR, 23–29 Jul 2023.

[32] Wenhan Yang and Baharan Mirzasoleiman. Robust contrastive language-image pretraining against adversarial attacks, 2023.

[33] Yu Yang, Hao Kang, and Baharan Mirzasoleiman. Towards sustainable learning: Coresets for data-efficient deep learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 39314–39330. PMLR, 23–29 Jul 2023.

[34] Haichao Yu, Yu Tian, Sateesh Kumar, Linjie Yang, and Heng Wang. The devil is in the details: A deep dive into the rabbit hole of data filtering, 2023.

[35] Qi Zhang, Yifei Wang, and Yisen Wang. On the generalization of multi-modal contrastive learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 41677–41693. PMLR, 23–29 Jul 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes**
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Yes**
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **Yes**
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Yes**
   (b) Complete proofs of all theoretical results. **Yes**
   (c) Clear explanations of any assumptions. **Yes**
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **Yes**
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes**
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **N/A**
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Yes**
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. **N/A**
   (b) The license information of the assets, if applicable. **N/A**
   (c) New assets either in the supplemental material or as a URL, if applicable. **N/A**
   (d) Information about consent from data providers/curators. **N/A**
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **N/A**
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. **N/A**
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **N/A**
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **N/A**

A Experimental Details

A.1 Accuracy on downstream tasks

Table 5: 5% ClipCov subset selected from CC3M

Datasets	Random	C-RHO	SemDeDup	CLIP Score	ClipCov
Caltech101	9.71%	3.24%	5.52%	11.22%	16.93%
DTD	3.72%	2.23%	2.61%	3.78%	4.10%
Food101	1.90%	1.05%	1.15%	2.58%	3.24%
ImageNet	1.27%	0.42%	0.85%	1.65%	4.46%
STL10	17.57%	16.45%	22.44%	15.10%	22.31%
SUN397	3.82%	0.79%	1.89%	2.70%	5.33%
ImageNet-Sketch	0.23%	0.24%	0.28%	0.68%	0.84%
ImageNet-V2	1.18%	0.42%	0.75%	1.53%	3.76%
ImageNet-A	1.17%	0.76%	1.15%	1.44%	1.55%
ImageNet-R	2.32%	1.13%	1.37%	4.22%	5.52%
ObjectNet	0.60%	0.41%	0.53%	0.91%	1.07%

Table 6: 10% ClipCov subset selected from CC3M

Datasets	Random	C-RHO	SemDeDup	CLIP Score	ClipCov
Caltech101	18.67%	13.82%	16.60%	36.45%	32.44%
DTD	3.72%	4.31%	2.02%	8.94%	9.31%
Food101	3.40%	1.68%	0.99%	5.03%	5.36%
ImageNet	3.95%	1.89%	2.60%	6.48%	11.33%
STL10	26.10%	19.56%	23.93%	22.96%	35.01%
SUN397	9.92%	3.01%	5.49%	11.18%	15.78%
ImageNet-Sketch	1.13%	0.77%	0.64%	2.76%	3.89%
ImageNet-V2	3.66%	1.66%	2.55%	5.00%	9.04%
ImageNet-A	1.37%	1.28%	1.45%	1.69%	2.07%
ImageNet-R	4.87%	3.21%	3.65%	11.02%	12.71%
ObjectNet	1.11%	0.86%	1.07%	1.74%	2.12%

Table 7: 20% ClipCov subset Sizes Per Dataset Accuracies

Datasets	Random	C-RHO	SemDeDup	CLIP Score	ClipCov
Caltech101	34.30%	33.36%	30.70%	40.28%	45.83%
DTD	7.93%	7.71%	4.41%	12.39%	11.49%
Food101	5.83%	4.05%	4.22%	9.51%	9.49%
ImageNet	7.99%	7.39%	6.25%	12.79%	15.86%
STL10	32.72%	32.31%	28.07%	36.30%	42.24%
SUN397	17.81%	12.71%	13.93%	22.25%	24.89%
ImageNet-Sketch	2.48%	3.69%	1.82%	6.94%	7.77%
ImageNet-V2	7.12%	6.34%	5.23%	10.72%	13.38%
ImageNet-A	1.91%	2.17%	1.83%	2.43%	2.17%
ImageNet-R	9.14%	10.26%	7.32%	17.55%	19.32%
ObjectNet	2.06%	2.45%	2.05%	3.40%	3.54%

Table 8: 30% ClipCov subset selected from CC3M

Datasets	Random	C-RHO	SemDeDup	CLIP Score	ClipCov
Caltech101	38.18%	42.47%	36.12%	47.25%	48.17%
DTD	8.62%	10.90%	9.15%	14.63%	12.82%
Food101	6.85	5.86%	5.76%	12.65%	12.54%
ImageNet	10.21%	11.01%	9.32%	18.54%	19.12%
STL10	39.24%	39.85%	39.56%	47.83%	45.95%
SUN397	21.09%	19.14%	19.19%	31.41%	37.04%
ImageNet-Sketch	4.28%	5.94%	3.12%	10.19%	10.35%
ImageNet-V2	8.94%	9.20%	7.09%	16.06%	16.09%
ImageNet-A	2.12%	2.67%	2.51%	2.47%	3.35%
ImageNet-R	12.21%	14.32%	10.56%	22.39%	23.50%
ObjectNet	2.57%	3.20%	2.56%	4.91%	4.98%

Table 9: 40% ClipCov subset selected from CC3M

Datasets	Random	C-RHO	SemDeDup	CLIP Score	ClipCov
Caltech101	39.31%	45.47%	40.67%	51.73%	52.08%
DTD	12.61%	11.12%	9.57%	12.93%	12.39%
Food101	8.51%	7.08%	8.01%	12.68%	12.61%
ImageNet	11.99%	13.56%	11.34%	17.48%	18.47%
STL10	42.14%	43.01%	40.27%	46.11%	45.85%
SUN397	24.41%	23.39%	24.08%	30.63%	31.49%
ImageNet-Sketch	5.05%	6.56%	4.60%	9.84%	10.17%
ImageNet-V2	9.98%	11.86%	9.34%	14.57%	15.25%
ImageNet-A	2.51%	2.91%	2.64%	2.97%	2.80%
ImageNet-R	14.09%	16.90%	13.36%	22.43%	22.10%
ObjectNet	3.16%	4.05%	3.32%	4.79%	4.84%

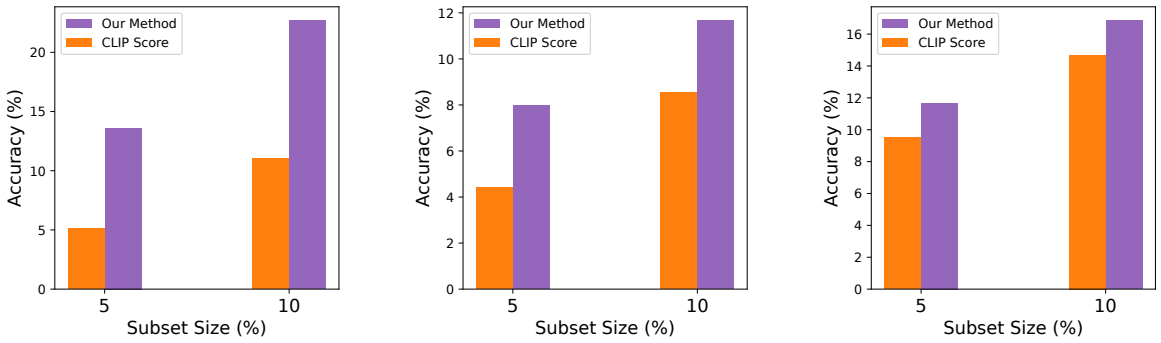
Table 10: 50% ClipCov subset selected from CC3M

Datasets	Random	C-RHO	SemDeDup	CLIP Score	ClipCov
Caltech101	41.96%	46.80%	44.06%	47.25%	48.17%
DTD	9.20%	10.96%	8.51%	14.63%	12.82%
Food101	8.44%	8.42%	8.39%	12.65%	12.54%
ImageNet	13.45%	14.56%	13.28%	18.54%	19.12%
STL10	45.29%	46.40%	44.46%	47.83%	45.95%
SUN397	24.41%	25.78%	25.88%	31.41%	37.04%
ImageNet-Sketch	6.45%	7.96%	5.65%	10.19%	10.35%
ImageNet-V2	11.55%	12.11%	11.08%	16.06%	16.09%
ImageNet-A	2.37%	3.05%	3.03%	2.47%	3.35%
ImageNet-R	15.63%	17.40%	14.85%	22.39%	23.50%
ObjectNet	3.47%	4.36%	3.45%	4.91%	4.98%

A.2 Additional Training Details

The experiments were conducted using NVIDIA A100s and NVIDIA RTX A6000 GPUs.

A.3 Visualization of CC12M Results



(a) ImageNet (b) ImageNet Dist. Shift (c) Avg. over 11 Datasets

Figure 3: Performance across subset of sizes 5% and 10% from CC12M