Data-Efficient Multi-Modal Contrastive Learning: Prioritizing Data Quality over Quantity

Anonymous Authors

Reviewed on OpenReview: N/A

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 a subset of image-caption pairs that provably generalizes on par with the full data when trained on, has remained an open question. In this work, we propose the first theoretically rigorous data selection method for CLIP. We show that subsets that best preserve the cross-covariance of the images and captions of the full data best preserve CLIP's generalization performance. Our extensive experiments on ConceptualCaptions3M and ConceptualCaptions12M demonstrate that subsets of size 5%-10% found by ClipCov achieve over 150% and 40% the accuracy of the next best baseline on ImageNet and its shifted versions. Moreover, we show that our subsets exhibit average relative performance improvement over the next best baseline of nearly 50% across 14 downstream datasets.

1 Introduction

The success of Contrastive Language-Image Pre-training (CLIP) models like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), trained on enormous datasets of 400M and 1B image-caption pairs, respectively, underscores a pivotal moment in AI where scale and data diversity dramatically enhance model capabilities. These models have set new benchmarks in zero-shot generalization and resilience to distribution shifts, illustrating the profound impact of large-scale, diverse datasets. However, the reliance on such vast amounts of data poses significant challenges, including computational costs and the environmental impact of training, thereby raising critical questions about the efficiency and sustainability of data usage in AI. Studies such as (Gadre et al., 2023) have started to address these concerns by showing that smaller, meticulously curated datasets can sometimes outperform their voluminous counterparts. Yet, selecting optimally small yet effective subsets for CLIP remains unsolved and the complex multimodal nature of CLIP makes existing data selection techniques inapplicable (Yang et al., 2023; Joshi and Mirzasoleiman, 2023; Coleman et al., 2020).

Addressing this, our work leverages theoretical insights from (Nakada et al., 2023) i.e. the *cross-covariance of data* determined the encoders learnt with CLIP and selects subsets that preserve the *cross-covariance of data*, thus learning encoders similar to those learnt on the full data. Empirically, we find our approach, ClipCov, can select subsets of sizes 5% and 10% from ConceptualCaptions3M and ConceptualCaptions12M (Sharma et al., 2018) that can surpass the next best baseline significantly, achieving over 150% relative performance improvement on ImageNet and its variants (Deng et al., 2009; Djolonga et al., 2021; Wang et al., 2019; Recht et al., 2019; Barbu et al., 2019; Hendrycks et al., 2021), and nearly 50% on average across 14 downstream datasets.

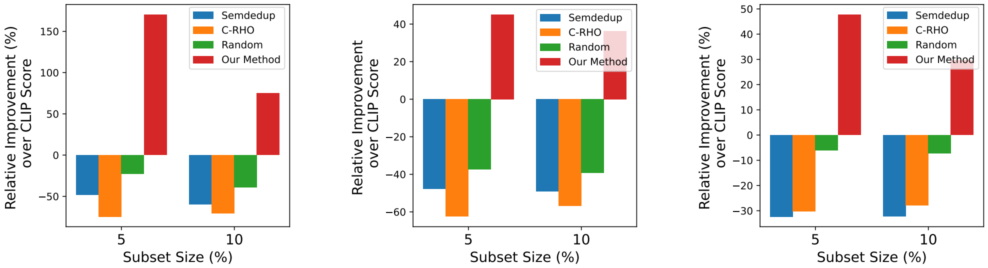


Figure 1: IN, IN Dist. Shift. and Avg. on 14 Datasets (from left to right) on CC3M

2 Method

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, where $x^i_{\mathcal{V}}$ denotes the image and $x^i_{\mathcal{L}}$ denotes the caption of i-th example. CLIP maximizes the representation similarity of paired image-captions and minimizes that of unpaired image-captions using vision encoder $f_{\mathcal{V}}$ and language encoder $f_{\mathcal{L}}$, respectively. After training, the quality of the learned representations is evaluated by zero-shot classification on different downstream image classification datasets i.e. images are classified by the similarity of their representation to the text representation of the downstream class names. Our goal is to find a subset of training data $S \subseteq V$ of size $n_s \ll n$, 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.

The training dynamics on the full data V can be determined by the *cross-covariance of the full data* (Nakada et al., 2023). Thus, the trained encoders on the full data and subset are determined by their respective data cross-covariances. Hence, we can see that if the *cross-covariance of the subset* S closely approximates the *cross-covariance of the full data*, the encoders learnt on the subset S will be similar to the encoders learnt on the full data V. Our method, ClipCov, selects subsets to do so by 1) preserving the centers of vision and language data 2) capturing the alignment & spread (covariance) of vision and language data. The pseudocode for ClipCovas well as theoretical guarantees for how similar encoders learnt on these subsets are to those learnt on the full data appear in Appendix A.

3 Results and Conclusion

Here, we compare the performance of the 5% and 10% training subsets found by ClipCov to subsets found by data-filtering baselines, including C-RHO, SemDeDup, CLIP score and Random selection. We use Conceptual Captions 3M and 12M used previously in (Goel et al., 2022). We evaluate all the methods on ImageNet (IN), IN Variants (Distribution Shift) and 14 downstream tasks as proposed by (Chen et al., 2020). Further details about the experiments appears in B. In conclusion, we show that our theoretically backed method, ClipCov, that selects subsets to preserve the *cross-covariance of the full data* can enable data-efficient multi-modal contrastive learning, thus significantly speeding up training as well as improving downstream generalization.

Table 1: Performance on CC12M

Subset	ImageNet	IN Dist. Shift	Avg.
CLIPCov 5%	13.61%	7.99%	11.68%
CLIP Score 5%	5.10%	4.42%	9.49%
CLIPCov 10%	22.71%	12.76%	16.87%
CLIP Score 10%	11.02%	8.55%	14.69%

Broader Impact Statement

Data-efficient learning democratizes AI by lowering computational costs, making advanced technologies accessible to a wider range of innovators and communities. However, before deploying subset selection algorithms broadly, it's crucial to assess their impact on different sub-populations of data to ensure fairness in AI applications.

References

A. Abbas, K. Tirumala, D. Simig, S. Ganguli, and A. S. Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication, 2023.

A. Barbu, D. Mayo, J. Alverio, W. Luo, C. Wang, D. Gutfreund, J. Tenenbaum, and B. 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. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/97af07a14cacba681feacf3012730892-Paper.pdf`.

N. Buchbinder, M. Feldman, J. Seffi, and R. Schwartz. A tight linear time (1/2)-approximation for unconstrained submodular maximization. *SIAM Journal on Computing*, 44(5):1384–1402, 2015.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In H. D. III and A. 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. URL `https://proceedings.mlr.press/v119/chen20j.html`.

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.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

J. Djolonga, J. Yung, M. Tschannen, R. Romijnders, L. Beyer, A. Kolesnikov, J. Puigcerver, M. Minderer, A. D'Amour, D. Moldovan, S. Gelly, N. Houlsby, X. Zhai, and M. Lucic. On robustness and transferability of convolutional neural networks, 2021.

S. Y. Gadre, G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang, E. Orgad, R. Entezari, G. Daras, S. Pratt, V. Ramanujan, Y. Bitton, K. Marathe, S. Mussmann, R. Vencu, M. Cherti, R. Krishna, P. W. Koh, O. Saukh, A. Ratner, S. Song, H. Hajishirzi, A. Farhadi, R. Beaumont, S. Oh, A. Dimakis, J. Jitsev, Y. Carmon, V. Shankar, and L. Schmidt. Datacomp: In search of the next generation of multimodal datasets, 2023.

S. Goel, H. Bansal, S. Bhatia, R. A. Rossi, V. Vinay, and A. Grover. Cyclip: Cyclic contrastive language-image pretraining, 2022.

D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization, 2021.

W. Ji, Z. Deng, R. Nakada, J. Zou, and L. Zhang. The power of contrast for feature learning: A theoretical analysis, 2021.

C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

S. Joshi and B. Mirzasoleiman. Data-efficient contrastive self-supervised learning: Most beneficial examples for supervised learning contribute the least. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. 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. URL https://proceedings.mlr.press/v202/joshi23b.html.

Y. Li, F. Liang, L. Zhao, Y. Cui, W. Ouyang, J. Shao, F. Yu, and J. Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=zq1iJkNk3uN.

P. Maini, S. Goyal, Z. C. Lipton, J. Z. Kolter, and A. Raghunathan. T-mars: Improving visual representations by circumventing text feature learning, 2023.

S. Mindermann, J. Brauner, M. Razzak, M. Sharma, A. Kirsch, W. Xu, B. Höltgen, A. N. Gomez, A. Morisot, S. Farquhar, and Y. Gal. Prioritized training on points that are learnable, worth learning, and not yet learnt, 2022.

M. 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.

B. Mirzasoleiman, A. Badanidiyuru, and A. Karbasi. Fast constrained submodular maximization: Personalized data summarization. In *International Conference on Machine Learning*, pages 1358–1367. PMLR, 2016.

R. Nakada, H. I. Gulluk, Z. Deng, W. Ji, J. Zou, and L. Zhang. Understanding multimodal contrastive learning and incorporating unpaired data. In F. Ruiz, J. Dy, and J.-W. 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. URL https://proceedings.mlr.press/v206/nakada23a.html.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In M. Meila and T. 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. URL `https://proceedings.mlr.press/v139/radford21a.html`.

B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do imagenet classifiers generalize to imagenet?, 2019.

P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

H. Wang, S. Ge, E. P. Xing, and Z. C. Lipton. Learning robust global representations by penalizing local predictive power, 2019.

Y. Xue, S. Joshi, E. Gan, P.-Y. Chen, and B. Mirzasoleiman. Which features are learnt by contrastive learning? On the role of simplicity bias in class collapse and feature suppression. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. 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. URL `https://proceedings.mlr.press/v202/xue23d.html`.

W. Yang and B. Mirzasoleiman. Robust contrastive language-image pretraining against adversarial attacks, 2023.

Y. Yang, H. Kang, and B. 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. URL `https://proceedings.mlr.press/v202/yang23g.html`.

Appendix A. Finding the Most Generalizable Subset

A.1 Problem Formulation

Data Distribution
Let $\mathcal{D} = \{(x_{\mathcal{V}}^i, x_{\mathcal{L}}^i)\}_{i \in V}$ be a set of $n = |V|$ image-caption pairs i.e. the full training data available to us, where $x_{\mathcal{V}}^i$ denotes the image and $x_{\mathcal{L}}^i$ denotes the caption of 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 $(x_{\mathcal{V}}^i, x_{\mathcal{L}}^i) \in \mathcal{D}$ be generated as follows:

$$x_{\mathcal{V}}^i = T_{\mathcal{V}}(u^i + \epsilon_{\mathcal{V}}) \qquad x_{\mathcal{L}}^i = T_{\mathcal{L}}(u^i + \epsilon_{\mathcal{L}}) \tag{1}$$

where: $u^i \in \mathbb{R}^d$ is the shared underlying feature vector of 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, respectively; $\epsilon_{\mathcal{V}}, \epsilon_{\mathcal{L}}$ are the noise in underlying features for vision and language, respectively. We refer to $\overline{u}_{\mathcal{V}}^i = u^i + \epsilon_{\mathcal{V}}^i$ and $\overline{u}_{\mathcal{L}}^i = u^i + \epsilon_{\mathcal{L}}^i$ as the noisy underlying feature for the image and caption, respectively. The underlying feature u^i, for each image-caption pair is sampled independently of each other and 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.
Similar to prior work Xue et al. (2023) studying uni-modal contrastive learning, we have that the k-th coordinate of the underlying feature vector corresponds to the underlying feature of latent class k. This data-distribution is identical to the data-distribution in Nakada et al. (2023) to study multi-modal contrastive learning, with the sole addition of latent-classes. 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 maximizes the representation similarity of paired image-captions and minimizes that of unpaired image-captions using vision encoder $f_{\mathcal{V}} : \mathbb{R}^{d_{\mathcal{V}}} \to \mathbb{R}^K$ and language encoder $f_{\mathcal{L}} : \mathbb{R}^{d_{\mathcal{L}}} \to \mathbb{R}^K$ that map input data in vision and language data space into a shared K-dimensional representation space, respectively (where K is the number of latent classes as defined earlier). The CLIP loss is defined as follows:

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

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 Nakada et al. (2023); Ji et al. (2021); Xue et al. (2023). Additionally, we use the linear multimodal contrastive loss used in Nakada et al. (2023):

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

where $A_{ij} \coloneqq (F_{\mathcal{V}} x_{\mathcal{V}}^i)^{\top} (F_{\mathcal{L}} x_{\mathcal{L}}^j)$. Both the CLIP loss and the linear multimodal contrastive loss are derived from the same generalized multimodal contrastive loss and achieve similar empirical performance Nakada et al. (2023).

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

Zero-Shot Classification

After training, the quality of the learned representations is evaluated by zero-shot classification on different downstream image classification datasets. A downstream task $\mathcal{D}_{\mathcal{Y}}$ is defined as a classification task on unseen data where the latent classes $\mathcal{Y} \subseteq [K]$ are a subset of latent classes of the pre-training data. For zero-shot classification on downstream task $\mathcal{D}_{\mathcal{Y}}$, we use the language encoder $f_{\mathcal{L}}$ to encode the label (e.g. the name of the class) corresponding to each latent class $y \in \mathcal{Y}$; then, the classification of an example $x_{\mathcal{V}}$ is $zs_{f_V, f_L}(x_{\mathcal{V}}) = \arg\max_{k\in[K]} \frac{f_{\mathcal{V}} x_{\mathcal{V}} \cdot z_k}{\|f_{\mathcal{V}} x_{\mathcal{V}}\| \|z_k\|}$, 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 the label corresponding to class k i.e. an example $x_{\mathcal{V}}$ is classified by the closest (average) label representation. In practice, Radford et al. (2021) proposed using a set of pre-engineered templates, e.g. '`A photo of a {label}`' to create several captions representing '`{label}`'. Thus, the average representation of the label would be the average of the representations of the templates obtained using $f_{\mathcal{L}}$. 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:

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

Finding Generalizable Multimodal Subsets

Our goal is to find a subset of training data $S \subseteq V$ of size n_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, 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.

Now, we present our method ClipCov. We first theoretically characterize how well the encoders learnt on a 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.

The training dynamics on the full data V can be determined by the cross-covariance of the full data $C_{\mathcal{D}}^{V}$ Nakada et al. (2023). First, the linear loss function (Eq. (3)) can be rewritten as the SVD objective function:

$$\mathcal{L}(F_{\mathcal{V}}, F_{\mathcal{L}}) = \mathrm{Tr}(F_{\mathcal{V}} A F_{\mathcal{L}}^{\top}) - \frac{\rho}{2}\|F_{\mathcal{V}}^{\top} F_{\mathcal{L}}\|_F^2 \tag{6}$$

and then Nakada et al. (2023) shows that A is equal to the centered cross-covariance matrix of the full data $C_{\mathcal{D}}^{V}$

$$A = C_{\mathcal{D}}^{V} := \frac{1}{|V|}\sum_{i \in V}(x_{\mathcal{V}}^{i} - \mu_{x_{\mathcal{V}}})(x_{\mathcal{L}}^{i} - \mu_{x_{\mathcal{L}}})^{\top} \tag{7}$$

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.

Thus, the trained encoders on the full data and subset are determined by their respective data cross-covariance matrices:

$$\mathcal{L}(F_{\mathcal{V}}, F_{\mathcal{L}}) = \underset{F_{\mathcal{V}}, F_{\mathcal{L}}}{\arg\max}\, \mathrm{Tr}(F_{\mathcal{V}} C_{\mathcal{D}}^{V} F_{\mathcal{L}}^{\top}) - \frac{\rho}{2}\|F_{\mathcal{V}}^{\top} F_{\mathcal{L}}\|_F^2 \tag{8}$$

$$\mathcal{L}(F_{\mathcal{V}}^{S}, F_{\mathcal{L}}^{S}) = \underset{F_{\mathcal{V}}, F_{\mathcal{L}}}{\arg\max}\, \mathrm{Tr}(F_{\mathcal{V}}^{S} C_{\mathcal{D}}^{S} {F_{\mathcal{L}}^{S}}^{\top}) - \frac{\rho}{2}\|{F_{\mathcal{V}}^{S}}^{\top} F_{\mathcal{L}}^{S}\|_F^2 \tag{9}$$

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

Hence, we can 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, the encoders learnt on the subset S will be similar to the encoders learnt on the full data V.

A.2 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 $\overline{C_{\mathcal{U}}^{V}} = \frac{1}{|V|}\sum_{i \in V}(\overline{u}_{\mathcal{V}}^{i} - \mathbb{E}_{i \in V}\,\overline{u}_{\mathcal{V}}^{i})(\overline{u}_{\mathcal{L}}^{i} - \mathbb{E}_{i \in V}\,\overline{u}_{\mathcal{L}}^{i})^{\top}$ and $\overline{C_{\mathcal{U}}^{S}} = \frac{1}{|S|}\sum_{i \in S}(\overline{u}_{\mathcal{V}}^{i} - \mathbb{E}_{i \in S}\,\overline{u}_{\mathcal{V}}^{i})(\overline{u}_{\mathcal{L}}^{i} - \mathbb{E}_{i \in S}\,\overline{u}_{\mathcal{L}}^{i})^{\top}$ be the cross-covariance of noisy underlying features for the full data V and subset S, respectively.

$$\left\|C_{\mathcal{D}}^{V} - C_{\mathcal{D}}^{S}\right\| = \left\|T_{\mathcal{V}}\overline{C_{\mathcal{U}}^{V}}T_{\mathcal{L}}^{\top} - T_{\mathcal{V}}\overline{C_{\mathcal{U}}^{S}}T_{\mathcal{L}}^{\top}\right\| \tag{10}$$

$$\leq \|T_{\mathcal{V}}\|\,\|T_{\mathcal{L}}\|\left\|\overline{C_{\mathcal{U}}^{V}} - \overline{C_{\mathcal{U}}^{S}}\right\| \tag{11}$$

$$\leq \left\|\overline{C_{\mathcal{U}}^{V}} - \overline{C_{\mathcal{U}}^{S}}\right\| \qquad \text{(Since, } \|T_{\mathcal{V}}\| = \|T_{\mathcal{L}}\| \leq 1\text{)} \tag{12}$$

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

With this, we have that if the subset S preserves the cross-covariance of noisy underlying feature of the full data V, it can preserve the data cross-covariance of the full data.

From the definition of cross-covariance, we have that the $[\overline{C^V_{\mathcal{U}}}]_{(k_1,k_2)}$ captures the cross-covariance between the k_1-th and k_2-th co-ordinate of the underlying feature vector, which correspond to the underlying features for the k_1-th and k_2-th latent classes respectively. Since underlying features in different latent classes are sampled independently of each other, the cross-covariance between the *noisy* underlying features of different latent classes is entirely due to noise. Hence, the objective is to preserve the cross-covariance within latent classes and ensure cross-covariance across latent classes is 0 (since it is caused by the noise). To do so, we can minimize the following objective:

$$\min_{S\subseteq V,|S|\leq n_s} \left\| diag(\overline{C^S_{\mathcal{U}}}) - diag(\overline{C^V_{\mathcal{U}}}) \right\| + \left\| \overline{C^S_{\mathcal{U}}} - diag(\overline{C^S_{\mathcal{U}}}) \right\| \tag{13}$$

where we preserve the cross-covariance within latent classes with the first term and destroy the cross-covariance across latent classes with the second term.

Preserving the Cross-Covariance within Latent Classes

We now discuss how to preserve the cross-covariance within a given latent class k i.e. minimizing the absolute value of $[\overline{C^S_{\mathcal{U}}}]_{(k,k)} - [\overline{C^V_{\mathcal{U}}}]_{(k,k)}$.

Let $\mu^V_{\mathcal{V}} = \mathbb{E}_{i\in V}\, \overline{u}^i_{\mathcal{V}}$ and $\mu^V_{\mathcal{L}} = \mathbb{E}_{i\in V}\, \overline{u}^i_{\mathcal{L}}$ be the mean noisy shared features of images and captions in the full data V, respectively. Similarly, let $\mu^S_{\mathcal{V}} = \mathbb{E}_{i\in S}\, \overline{u}^i_{\mathcal{V}}$ and $\mu^S_{\mathcal{L}} = \mathbb{E}_{i\in S}\, \overline{u}^i_{\mathcal{L}}$ be mean noisy shared features images and captions in the subset S.

$$\begin{aligned}
&[C^S_{\mathcal{Z}_p}]_{(k,k)} - [C^V_{\mathcal{Z}_p}]_{(k,k)} = \\
&\frac{1}{|S|}\sum_{i\in S}\left[(\overline{u}^i_{\mathcal{V}} - \mu^S_{\mathcal{V}})(\overline{u}^i_{\mathcal{L}} - \mu^S_{\mathcal{L}})^\top\right]_{(k,k)} \\
&-\frac{1}{|V|}\sum_{j\in V}\left[(\overline{u}^j_{\mathcal{V}} - \mu^V_{\mathcal{V}})(\overline{u}^j_{\mathcal{L}} - \mu^V_{\mathcal{L}})^\top\right]_{(k,k)}
\end{aligned} \tag{14}$$

Without the noise in the image and caption underlying features, the cross-covariance for underlying feature of latent class k is entirely determined by examples in latent class k. Thus, we should only preserve the terms of the cross-covariance for underlying feature of latent class k contributed by examples in latent class k in the full data V and the subset S.

$$\begin{aligned}
\text{Eq.(14)} = \frac{1}{|V||S|}\sum_{i\in S_k}\sum_{j\in V_k}\Bigg[&(\overline{u}^i_{\mathcal{V}} - \mu^S_{\mathcal{V}})(\overline{u}^i_{\mathcal{L}} - \mu^S_{\mathcal{L}})^\top \\
-(\overline{u}^j_{\mathcal{L}} - \mu^V_{\mathcal{L}})(\overline{u}^j_{\mathcal{V}} - \mu^V_{\mathcal{V}})^\top\Bigg]_{(k,k)} &
\end{aligned} \tag{15}$$

$$\begin{aligned}
&\leq \frac{1}{|V||S|}\left(\sum_{i\in S_k}\sum_{j\in V_k}\left\|\overline{u}^i_{\mathcal{V}} - \overline{u}^j_{\mathcal{L}}\right\| + \left\|\overline{u}^i_{\mathcal{L}} - \overline{u}^j_{\mathcal{V}}\right\|\right) \\
&+ \underbrace{\left\|\mu^S_{\mathcal{V}} - \mu^V_{\mathcal{L}}\right\| + \left\|\mu^S_{\mathcal{L}} - \mu^V_{\mathcal{V}}\right\|}_{\text{cross-modal distance of means}}
\end{aligned} \tag{16}$$

The complete derivation from Eq. (14) to Eq. (16) appears in Appendix D.

But minimizing the first term is sufficient to ensure that cross-modal distance of means is minimized. Here, we show this for the difference between $\mu_{\mathcal{V}}^{S}$ and $\mu_{\mathcal{L}}^{V}$:

$$\left\|\mu_{\mathcal{V}}^{S} - \mu_{\mathcal{L}}^{V}\right\| = \left\| \frac{1}{|S|} \sum_{i \in S} \overline{u}_{\mathcal{V}}^{i} - \frac{1}{|V|} \sum_{j \in V} \overline{u}_{\mathcal{L}}^{j} \right\| \tag{17}$$

$$\leq \frac{1}{|V||S|} \sum_{i \in S} \sum_{j \in V} \left\| \overline{u}_{\mathcal{V}}^{i} - \overline{u}_{\mathcal{L}}^{j} \right\| \tag{18}$$

A symmetric argument holds for the difference between $\mu_{\mathcal{L}}^{S}$ and $\mu_{\mathcal{V}}^{V}$.
Thus, preserving the cross-covariance within all latent classes $k \in [K]$ can be done by minimizing $\forall k \in K$:

$$\frac{1}{|V||S|} \sum_{i \in S_k} \sum_{j \in V_k} \left\| \overline{u}_{\mathcal{V}}^{i} - \overline{u}_{\mathcal{L}}^{j} \right\| + \left\| \overline{u}_{\mathcal{L}}^{i} - \overline{u}_{\mathcal{V}}^{j} \right\| \tag{19}$$

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 on the full data. We refer to these encoders as the proxy vision encoder $f_{\mathcal{V}}^{p}$ and the proxy language encoder $f_{\mathcal{L}}^{p}$. Nakada et al. (2023) shows that vision and language encoders trained on the full data recover the corresponding noisy underlying features [2], respectively.

$$f_{\mathcal{V}}^{p}(x_{\mathcal{V}}) = \overline{u}_{\mathcal{V}}, \forall x_{\mathcal{V}} \in \mathcal{X}_{\mathcal{V}} \tag{20}$$

$$f_{\mathcal{L}}^{p}(x_{\mathcal{L}}) = \overline{u}_{\mathcal{L}}, \forall x_{\mathcal{L}} \in \mathcal{X}_{\mathcal{L}} \tag{21}$$

Using the proxy encoders, we now introduce the notion of cross-modal similarity.
Definition. Cross Modal Similarity We define the cross-modal similarity between any two examples $i, j \in V$ as the sum of the similarities between the representations of the images of examples i and j with the other's caption. Formally, $\forall i, j \in V$, we have:

$$sim(i,j) = f_{\mathcal{V}}^{p}(x_{\mathcal{V}}^{i})^{\top} f_{\mathcal{L}}^{p}(x_{\mathcal{L}}^{j}) + f_{\mathcal{V}}^{p}(x_{\mathcal{V}}^{j})^{\top} f_{\mathcal{L}}^{p}(x_{\mathcal{L}}^{i}) \tag{22}$$

We can now maximize the cross-modal similarity $sim(i,j)$ to minimize $\left\| \overline{u}_{\mathcal{V}}^{i} - \overline{u}_{\mathcal{L}}^{j} \right\| + \left\| \overline{u}_{\mathcal{L}}^{i} - \overline{u}_{\mathcal{V}}^{j} \right\|$ for all $i \in S, j \in V$. Thus, minimizing (19) $\forall k \in K$ is equivalent to maximizing $\sum_{i \in S_k} \sum_{j \in V_k} sim(i,j), \forall k \in K$. To maximize $\forall k \in K$ simultaneously, we maximize:

$$\sum_{k \in K} \frac{1}{|V_k|} \sum_{i \in S_k} \sum_{j \in V_k} sim(i,j) \tag{23}$$

where normalizing by latent class size V_k is to optimize equally for smaller and larger latent classes.
In practice, since 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 selected examples. Thus, we optimize to

[2] up to orthogonal transformation, for clarity, we assume that this recovery is exact.

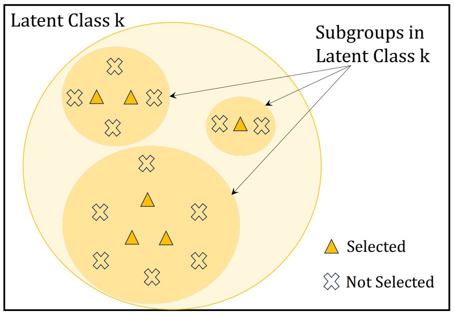
preserve the cross-covariance within latent classes by maximizing $F_{\text{intra-class}}(S) :=$

$$\sum_{k\in K}\frac{1}{|V_k|}\left(\sum_{\substack{i\in S_k\\ j\in V_k}} sim(i,j)-\frac{1}{2}\sum_{\substack{i\in S_k\\ j\in S_k}} sim(i,j)\right) \tag{24}$$

To guide the objective towards selecting diverse examples within latent class, in addition to penalizing the similarity of selected examples, we add $F_{\text{self}}(S)=\sum_{i\in S} sim(i,j)$ i.e. prioritizing examples with high cross-modal similarity to themselves. For a trained CLIP model, examples that are most centrally located in the data have the highest cross-modal similarity to themselves. Intuitively, 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 dense groups will have the highest cross-modal similarities at the end of training. Thus, picking examples with high cross-modal similarity to themselves allows us to capture the different dense subgroups within a latent classes. Hence, we preserve the cross-covariance within latent classes by maximizing $F_{\text{intra-class}}(S)+F_{\text{self}}(S) =$

$$\underbrace{\sum_{k\in K}\frac{1}{|V_k|}\left(\sum_{\substack{i\in S_k\\ j\in V_k}} sim(i,j)-\frac{1}{2}\sum_{\substack{i\in S_k\\ j\in S_k}} sim(i,j)\right)}_{F_{\text{intra-class}}(S)} + \underbrace{\sum_{i\in S} sim(i,j)}_{F_{\text{self}}(S)} \tag{25}$$

Figure 2: Visualization of examples selected by $F_{\text{intra-class}}(S)+F_{\text{self}}(S)$ in Cross-Modal Similarity Space



To illustrate what kinds of examples are selected by this objective, we provide a visualization in Fig. 2 which shows the selected examples are similar to all the examples in the latent class, even from smaller subgroups. From Fig. 2, we can see how such a subset is representative of the latent class and thus can capture the cross-covariance within the latent class.

Destroying Cross-Covariance Across Latent Classes

Algorithm 1 ClipCov: Finding S_*

```
Input: Dataset $V$, Subset size $n_s$, proxy encoders: $f^p_{\mathcal{V}}$ and $f^p_{\mathcal{L}}$,
Output: Subset $S$
$\{V_1, ..., V_K\} \leftarrow$ approximate latent classes
$S \leftarrow \{\}$
$F(S) := F_{\substack{cross\\cov}}(S) - F^{\text{reg}}_{\text{intra-class}}(S) + F_{\text{label-sim}(S)}$ (Objective 30)
$S \leftarrow \emptyset$
while $|S| \leq n_s$ do
    $e \leftarrow \arg\max_{e \in \mathcal{D} \setminus S} F(e|S)$
    $S \leftarrow S \cup \{e\}$
end while
return double-greedy($S$)
```

We now discuss how to ensure cross-covariance across latent classes is destroyed i.e. as close to 0 as possible. The cross-covariance across latent classes k_1 and k_2 is entirely due to noise in the underlying feature vector resulting in image-caption pair in latent class k_1 appearing with the underlying feature of latent class k_2 and vice-a-versa. Thus, we can minimize the cross-covariance across latent classes k_1 and k_2 by selecting image-caption pairs from latent class k_1 that are most dissimilar to examples in latent class k_2 i.e. do not have the underlying feature of latent class k_2 (and vice-a-versa). Thus, we minimize the average similarity of examples to other latent classes. The following objective, $F_{\text{inter-class}}(S)$, formalizes this:

$$F_{\text{inter-class}}(S) := \sum_{\substack{k_1, k_2 \in [K] \\ k_1 \neq k_2}} \sum_{i \in S_k} \sum_{j \in V_{k_2}} \frac{sim(i,j)}{|V_{k_2}|} \tag{26}$$

where $sim(i,j)$ is the cross-modal similarity between image-caption pair i and j as defined in Def. A.2 and $\sum_{i \in S_k} \sum_{j \in V_{k_2}} \frac{sim(i,j)}{|V_{k_2}|}$ is exactly 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}.

Preserving Cross-Covariance of Data

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

$$F_{\substack{cross\\cov}}(S) := F_{\text{intra-class}}(S) + F_{\text{self}}(S) - F_{\text{inter-class}}(S) \tag{27}$$

A.3 Deriving the Final Objective for Finding the Most Generalizable Subset

We now discuss two practical considerations that often arise when learning from large vision-language datasets and also account for them in the final objective to find subset S.

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 labels for 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 $F_{\text{label-sim}(S)}$

$$\begin{aligned} F_{\text{label-sim}(S)} = & \sum_{k\in[K]} \sum_{i\in S_k} \alpha f^p_{\mathcal{L}}(x^i_{\mathcal{L}})^\top f^p_{\mathcal{L}}(y_k) \\ & - \sum_{i\in S_k} \alpha \frac{f^p_{\mathcal{L}}(x^i_{\mathcal{L}})^\top f^p_{\mathcal{L}}(y_k)}{|V_k|} \end{aligned} \tag{28}$$

where α is the ratio of average cross-modal similarity to the average similarity in text [3] Here, the second term prevents domination of classes with very good similarity to the label. This improves the zero-shot performance on various datasets.

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

$$F^{\text{reg}}_{\text{intra-class}}(S) = \sum_{k\in K} \frac{1}{|V_k|} \sum_{\substack{i\in S_k \\ j\in V_k}} \frac{sim(i,j)}{|V_k|} \tag{29}$$

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

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

$$S_* \approx \underset{S\subseteq V, |S|\leq n_s}{\arg\max}\ F_{\substack{cross \\ cov}}(S) - F^{\text{reg}}_{\text{intra-class}}(S) + F_{\text{label-sim}(S)} \tag{30}$$

A.4 ClipCov: Efficiently Finding the Most Generalizable Subset

Here, we discuss how the proxy representations and latent classes required to optimize Objective (30) 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 Radford et al. (2021), which are trained on massive amounts of data and obtain impressive zero-shot generalization, thus likely recover the underlying features of the full data V well.

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

Approximating Latent Classes In practice, we do not have access to latent classes required to optimize Objective (30). Instead, we approximately recover latent classes using zero-shot classification using the proxy encoders. Moreover, in practice, models trained using CLIP are evaluated on a variety of downstream tasks that do not always have common latent classes; thus, in finding the subset S_*, we use fine-grained latent classes (e.g. ImageNet-1k latent classes) to capture nearly all downstream latent classes.

Scaling to Massive Datasets Since $F_{\text{inter-class}}(S)$ 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 (30) is NP-hard as it requires evaluating an exponential number of subsets. To efficiently find a near-optimal subset, we note that $F_{\substack{cross \\ cov}}(S)$ is approximately non-monotone submodular, and $F_{\text{label-sim}(S)}$, $F^{\text{reg}}_{\text{intra-class}}(S)$ are modular. So Objective (30) is approximately submodular. Thus, we can find a good subset using an algorithm for non-monotone submodular function maximization under a cardinality constraint. In particular, we first use the greedy algorithm to find a subset, and then filter the subset by applying unconstrained submodular maximization Mirzasoleiman et al. (2016). 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 Buchbinder et al. (2015), which calculates $a_e = F(e|\emptyset)$ and $b_e = F(V \setminus \{e\})$ for all $e \in S$, and then keeps examples for which $a_e \geq b_e$. 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 Minoux (2005). 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.

Appendix B. Experiments

Subset Size	Method	ImageNet	ImageNet Dist. Shift	Avg. over 14 Datasets
5%	Random	1.27%	1.10%	6.99%
	C-RHO	0.42%	0.59%	6.05%
	SemDeDup	0.85%	0.82%	5.89%
	CLIP Score	1.65%	1.76%	7.42%
	ClipCov	**4.46%**	**2.55%**	**9.91%**
10%	Random	3.95%	2.43%	11.50%
	C-RHO	1.89%	1.56%	8.90%
	SemDeDup	2.60%	1.87%	9.40%
	CLIP Score	6.48%	4.44%	12.84%
	ClipCov	**11.33%**	**5.97%**	**16.14%**

Table 2: Comparing Performance of 5% and 10% Subsets Selected from ConceptualCaptions3M

In this section, we compare the zero-shot performance of the 5% and 10% training subsets found by ClipCov subsets found by data-filtering 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 Sharma et al. (2018) which includes 3 million image-captions pairs, and has been widely employed for benchmark evaluations in various studies focusing on contrastive language-image pre-training Yang and Mirzasoleiman (2023); Goel et al. (2022); Li et al. (2022). We evaluate all the methods on downstream tasks proposed by Chen et al. (2020) and used in prior work for evaluating CLIP Yang and Mirzasoleiman (2023); Goel et al. (2022); Li et al. (2022). The exact list of datasets and corresponding accuracies appears in Appendix C.1.

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 Yang and Mirzasoleiman (2023).

Baselines The data-filtering baselines we consider are: (1) CLIP Score Gadre et al. (2023), (2) C-RHO Maini et al. (2023), (3) SemDeDup Abbas et al. (2023), 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 Mindermann et al. (2022) 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.

Zero-Shot Performance Table 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 from the these methods performing worse even than Random subsets. In contrast, ClipCov successfully

Method	ImageNet	ImageNet Dist. Shift	Avg. over 14 Datasets
$F_{\underset{cov}{cross}}(S)$	9.00%	5.30%	14.40%
$F_{\underset{cov}{cross}}(S) - F^{\text{reg}}_{\text{intra-class}}(S)$	8.94%	5.10%	14.26%
$F_{\underset{cov}{cross}}(S) + F_{\text{label-sim}(S)}$	10.87%	5.73%	13.91%
ClipCov	**11.33%**	**5.97%**	**16.14%**

Table 3: Ablation over Objective

Method	ImageNet	ImgNet Shift	Avg.
CLIP Score	5.01%	3.16%	10.53%
ClipCov	**6.70%**	**3.48%**	**13.68%**

Table 4: Ablation over Proxy Encoders

extract subsets that can preserve the downstream generalization performance on various datasets and outperforms CLIP Score. Fig. 1 shows that ClipCov achieves over 150% and 40% relative performance improvement over CLIP Score (the next best baseline) on ImageNet and its shifted versions. Moreover, it also shows that ClipCov nearly 50% relative performance improvement on average across the 14 downstream tasks.

Ablation Study Table 3 ablates over the objective and shows that both $F_{\text{label-sim}(S)}$ and $F^{\text{reg}}_{\text{intra-class}}(S)$ are essential additions to $F_{\underset{cov}{cross}}(S)$. Table 4 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 Radford et al. (2021). These results show that ClipCov is not sensitive to choice of proxy model. The drop in performance for both CLIP Score and ClipCovwhen compared to the subsets in Table 2, shows that using cross-modal similarities from encoders trained on more diverse and balanced data (e.g. CLIP from Radford et al. (2021))is beneficial to both CLIP Score and ClipCov.

Appendix C. Experimental Details

C.1 Downstream Datasets and Accuracies

Datasets

The 14 downstream datasets we evaluate on are the following (similar to the downstream datasets used by Yang and Mirzasoleiman (2023); Goel et al. (2022)):

Table 5: Downstream Datasets

Datasets
Caltech101
CIFAR10
CIFAR100
DTD
Food101
ImageNet
STL10
SVHN
SUN397
ImageNet-Sketch
ImageNet-V2
ImageNet-A
ImageNet-R
ObjectNet

Accuracies

Datasets	**Random**	**C-RHO**	**SemDeDup**	**CLIP Score**	**ClipCov**
Caltech101	5.62%	2.93%	4.37%	11.29%	14.15%
CIFAR10	15.43%	16.57%	12.51%	14.79%	18.64%
CIFAR100	3.98%	1.56%	2.12%	3.82%	2.96%
DTD	1.91%	2.61%	1.17%	1.97%	3.67%
Food101	2.38%	1.19%	1.49%	2.88%	3.04%
ImageNet	1.27%	0.42%	0.85%	1.65%	4.46%
STL10	18.01%	16.91%	19.41%	16.79%	23.46%
SVHN	8.74%	11.03%	8.55%	9.25%	10.97%
SUN397	5.56%	1.27%	2.58%	4.33%	7.81%
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: 5% Subset Sizes Per Dataset Accuracies

Datasets	**Random**	**C-RHO**	**SemDeDup**	**CLIP Score**	**ClipCov**
Caltech101	19.90%	10.19%	14.77%	28.76%	28.13%
CIFAR10	20.27%	22.06%	18.55%	18.55%	26.81%
CIFAR100	4.60%	4.28%	3.99%	6.60%	6.53%
DTD	3.19%	2.61%	1.01%	2.07%	2.93%
Food101	3.54%	1.85%	2.77%	5.79%	6.03%
ImageNet	3.95%	1.89%	2.60%	6.48%	11.33%
STL10	26.32%	22.95%	24.84%	26.59%	34.75%
SVHN	8.18%	8.78%	8.35%	7.61%	10.14%
SUN397	13.57%	5.45%	7.75%	13.15%	18.62%
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: 10% Subset Sizes Per Dataset Accuracies

C.2 Additional Experiments Comparing CLIPScore (next best baseline) and ClipCov(Subset Sizes 6%, 8%)

Subset Size	Method	ImageNet	ImageNet Dist. Shift	Avg. over 14 Datasets
6%	CLIPScore	2.75%	2.37%	8.77%
	ClipCov	**6.30%**	**3.53%**	**11.33%**
8%	CLIPScore	4.69%	3.66%	11.24%
	ClipCov	**9.70%**	**4.83%**	**13.72%**

Table 8: Comparing Performance of 6% and 8% Subsets Selected from ConceptualCaptions3M

C.3 Additional Training Details

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

Appendix D. Full Derivation from Eq. (14) to Eq. (16)

Let $\mu^V_{\mathcal{V}} = \mathbb{E}_{i\in V}\,\overline{u}^i_{\mathcal{V}}$ and $\mu^V_{\mathcal{L}} = \mathbb{E}_{i\in V}\,\overline{u}^i_{\mathcal{L}}$ be the mean noisy shared features of images and captions in the full data V, respectively. Similarly, let $\mu^S_{\mathcal{V}} = \mathbb{E}_{i\in S}\,\overline{u}^i_{\mathcal{V}}$ and $\mu^S_{\mathcal{L}} = \mathbb{E}_{i\in S}\,\overline{u}^i_{\mathcal{L}}$ be mean noisy shared features images and captions in the subset S.

$$\begin{aligned}
&[C^S_{\mathcal{Z}_p}]_{(k,k)} - [C^V_{\mathcal{Z}_p}]_{(k,k)} = \\
&\frac{1}{|S|}\sum_{i\in S}\Big[(\overline{u}^i_{\mathcal{V}} - \mu^S_{\mathcal{V}})(\overline{u}^i_{\mathcal{L}} - \mu^S_{\mathcal{L}})^\top\Big]_{(k,k)} \\
&-\frac{1}{|V|}\sum_{j\in V}\Big[(\overline{u}^j_{\mathcal{V}} - \mu^V_{\mathcal{V}})(\overline{u}^j_{\mathcal{L}} - \mu^V_{\mathcal{L}})^\top\Big]_{(k,k)}
\end{aligned} \tag{31}$$

Since the (k,k)-th element of $(\overline{u}^j_{\mathcal{V}} - \mu^V_{\mathcal{V}})(\overline{u}^j_{\mathcal{L}} - \mu^V_{\mathcal{L}})^\top$ is the same as the (k,k)-th element of $(\overline{u}^j_{\mathcal{L}} - \mu^V_{\mathcal{L}})(\overline{u}^j_{\mathcal{V}} - \mu^V_{\mathcal{V}})^\top$

$$\begin{aligned}
&= \frac{1}{|S|}\sum_{i\in S}\Big[(\overline{u}^i_{\mathcal{V}} - \mu^S_{\mathcal{V}})(\overline{u}^i_{\mathcal{L}} - \mu^S_{\mathcal{L}})^\top\Big]_{(k,k)} \\
&-\frac{1}{|V|}\sum_{j\in V}\Big[(\overline{u}^j_{\mathcal{L}} - \mu^V_{\mathcal{L}})(\overline{u}^j_{\mathcal{V}} - \mu^V_{\mathcal{V}})^\top\Big]_{(k,k)}
\end{aligned} \tag{32}$$

$$\begin{aligned}
&\frac{1}{|V||S|}\sum_{i\in S}\sum_{j\in V}\Bigg[(\overline{u}^i_{\mathcal{V}} - \mu^S_{\mathcal{V}})(\overline{u}^i_{\mathcal{L}} - \mu^S_{\mathcal{L}})^\top \\
&-(\overline{u}^j_{\mathcal{L}} - \mu^V_{\mathcal{L}})(\overline{u}^j_{\mathcal{V}} - \mu^V_{\mathcal{V}})^\top\Bigg]_{(k,k)}
\end{aligned} \tag{33}$$

For the population data, the cross-covariance for underlying feature of latent class k is entirely determined by examples in latent class k. Thus, we should only preserve the cross-covariance for underlying feature of latent class k, for examples in latent class k in the full data and the subset S.

$$= \frac{1}{|V||S|} \sum_{i \in S_k} \sum_{j \in V_k} \Bigg[(\overline{u}^i_{\mathcal{V}} - \mu^S_{\mathcal{V}})(\overline{u}^i_{\mathcal{L}} - \mu^S_{\mathcal{L}})^\top$$

$$- (\overline{u}^j_{\mathcal{L}} - \mu^V_{\mathcal{L}})(\overline{u}^j_{\mathcal{V}} - \mu^V_{\mathcal{V}})^\top \Bigg]_{(k,k)} \tag{34}$$

$$= \frac{1}{|V||S|} \sum_{i \in S_k} \sum_{j \in V_k} \left[\overline{u}^i_{\mathcal{V}} {\overline{u}^i_{\mathcal{L}}}^\top - \overline{u}^j_{\mathcal{L}} {\overline{u}^j_{\mathcal{V}}}^\top \right]_{(k,k)}$$

$$- \left[\overline{u}^i_{\mathcal{V}} {\mu^S_{\mathcal{L}}}^\top - \overline{u}^j_{\mathcal{L}} {\mu^V_{\mathcal{V}}}^\top \right]_{(k,k)} - \left[\mu^S_{\mathcal{V}} {\overline{u}^i_{\mathcal{L}}}^\top - \mu^V_{\mathcal{L}} {\overline{u}^j_{\mathcal{V}}}^\top \right]_{(k,k)}$$

$$+ \left[\mu^S_{\mathcal{V}} {\mu^S_{\mathcal{L}}}^\top - \mu^V_{\mathcal{L}} {\mu^V_{\mathcal{V}}}^\top \right]_{(k,k)}$$

Since the norm of all the vectors above is bounded by 1

$$\leq \frac{1}{|V||S|} \sum_{i \in S_k} \sum_{j \in V_k} \left\| \overline{u}^i_{\mathcal{V}} - \overline{u}^j_{\mathcal{L}} \right\| \left\| \overline{u}^i_{\mathcal{L}} - \overline{u}^j_{\mathcal{V}} \right\|$$

$$+ \left\| \overline{u}^i_{\mathcal{V}} - \overline{u}^j_{\mathcal{L}} \right\| \|\mu^S_{\mathcal{L}} - \mu^V_{\mathcal{V}}\| + \left\| \overline{u}^i_{\mathcal{L}} - \overline{u}^j_{\mathcal{V}} \right\| \|\mu^S_{\mathcal{L}} - \mu^V_{\mathcal{V}}\| \tag{35}$$

$$\leq \frac{1}{|V||S|} \sum_{i \in S_k} \sum_{j \in V_k} \left\| \overline{u}^i_{\mathcal{V}} - \overline{u}^j_{\mathcal{L}} \right\| + \left\| \overline{u}^i_{\mathcal{L}} - \overline{u}^j_{\mathcal{V}} \right\|$$

$$+ \|\mu^S_{\mathcal{V}} - \mu^V_{\mathcal{L}}\| + \|\mu^S_{\mathcal{L}} - \mu^V_{\mathcal{V}}\| \tag{36}$$

$$\leq \frac{1}{|V||S|} \left(\sum_{i \in S_k} \sum_{j \in V_k} \left\| \overline{u}^i_{\mathcal{V}} - \overline{u}^j_{\mathcal{L}} \right\| + \left\| \overline{u}^i_{\mathcal{L}} - \overline{u}^j_{\mathcal{V}} \right\| \right)$$

$$+ \underbrace{\|\mu^S_{\mathcal{V}} - \mu^V_{\mathcal{L}}\| + \|\mu^S_{\mathcal{L}} - \mu^V_{\mathcal{V}}\|}_{\text{cross-modal distance of means}} \tag{37}$$