Dynamic Cluster Data Sampling for Efficient and Long-Tail-Aware Vision-Language Pre-training

Mingliang Liang[1], Zhuoran Liu[†,2], Arjen P. de Vries[1], and Martha Larson[1]

[1] Radboud University, Nijmegen, The Netherlands
[2] University of Amsterdam, Amsterdam, The Netherlands
{mliang, z.liu, a.devries, mlarson}@cs.ru.nl

Abstract. The computational cost of training a vision-language model (VLM) can be reduced by sampling the training data. Previous work on efficient VLM pre-training has pointed to the importance of semantic data balance, adjusting the distribution of topics in the data to improve VLM accuracy. However, existing efficient pre-training approaches may disproportionately remove rare concepts from the training corpus. As a result, *long-tail concepts* remain insufficiently represented in the training data and are not effectively captured during training. In this work, we introduce a *dynamic cluster-based sampling approach (DynamiCS)* that downsamples large clusters of data and upsamples small ones. We first demonstrate the advantage of our cluster-scaling approach, which maintains the relative order of semantic clusters in the data and emphasizes the long-tail. This approach contrasts with current work, which focuses only on flattening the semantic distribution of the data. Then, we show the importance of dynamic sampling, which applies sampling at each epoch to improve cross-epoch data diversity and make upsampling practical. Our experiments show that DynamiCS reduces the computational cost of VLM training and provides a performance advantage for long-tail concepts. Code available at `https://github.com/MingliangLiang3/DynamiCS`.

Keywords: Efficient VLM Pre-training · Long-Tail Concepts · Cluster Scaling · Dynamic Sampling

1 Introduction

Vision-Language Models (VLMs) demonstrate strong transferability [4, 6, 12, 13, 19, 32, 48, 49]. Pre-trained VLMs can be applied to classification and image-text retrieval tasks, and their image encoders are widely used in Multimodal Large Language Models (MLLM) and generative models [14, 23, 28, 33]. CLIP [32], one of the popular VLMs, is trained on extremely large-scale datasets, requiring substantial GPU resources for pre-training. The training costs of CLIP have given rise to *cost-saving* training approaches, notably, RECLIP [15], FLIP [19],

[†]Corresponding author. Work was done while at Radboud University.

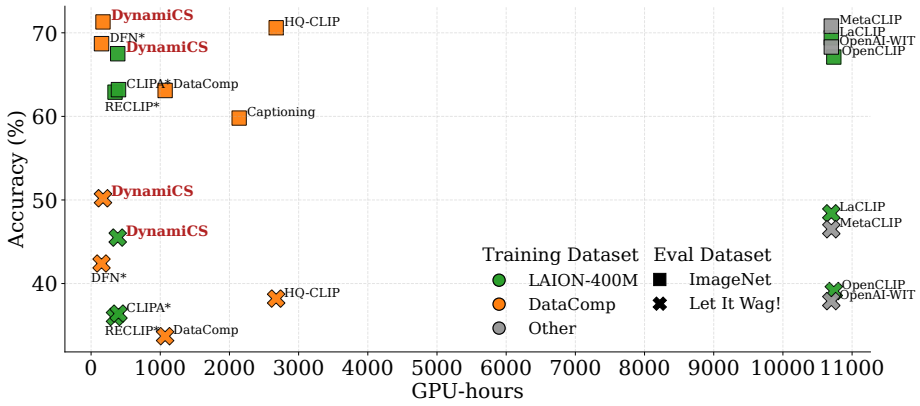

Fig. 1: Zero-shot top-1 accuracy on ImageNet-1K [5] and *Let It Wag!* [39] (long-tail test set). DynamiCS outperforms cost-saving baselines (RECLIP [15], FLIP [19], CLIPA [18]) and dual-purpose approaches (DataComp [10], DFN [9], Captioning [25]) while using less computational resources and achieves accuracy competitive with full-scale pre-training, e.g., OpenCLIP [4]. Experiments are conducted with ViT-B/16 pre-training of 6 epochs, applying different strategies to LAION-400M [35] or DataComp [10].

and CLIPA [18], which maintain VLM performance but reduce training costs by reducing the amount of text or image information used from each sample for training.

CLIP's large-scale pre-training data is collected from the web, but the data is curated to improve *semantic data balance*, i.e., to adjust the distribution of topics in the data, which in the wild typically has a very fat head and a very long tail. In [45], the advantages of CLIP data curation are described as avoiding biases and balancing the data over the metadata, which corresponds to semantic categories. These advantages motivate the data curation approach in MetaCLIP [45], which limits the number of data samples associated with each metadata category to 20k, effectively chopping off the fat head of the distribution.

Dual-purpose approaches apply data sampling for the simultaneous purposes of training cost reduction and achieving semantic balance. This gap is left open by MetaCLIP [45], whose goal is a semantically balanced training set. Metadata is not the only means of identifying semantic structure, rather clustering can be used instead, as with density-based pruning (DBP) [2]. Dual-purpose approaches can also combine data sampling for training cost reduction with sampling to improve data quality. For example, approaches have removed text-image pairs using another already-trained CLIP as a filter [9,10] and integrating information from automatically generated captions [25]. However, these approaches complicate the goal of training cost reduction since they all make critical use of another already-trained VLM.

In this paper, we propose a dual-purpose approach for VLM training reduction, DynamiCS, which reduces costs with dynamic sampling and addresses semantic data balance with cluster scaling. In Fig. 1, we observe that DynamiCS achieves a substantial speedup over OpenCLIP [4], MetaCLIP [45], and HQ-CLIP [43], while also yielding a slight improvement on ImageNet-1K zero-shot classification. It improves (top of graph) over other cost-saving training approaches (RECLIP, CLIPA, FLIP) and other dual-purpose approaches (DFN, DataComp, Captioning) in terms of training cost and also performance. The improvements are particularly remarkable on *long-tail test data* (bottom of the graph), where the pre-trained DynamiCS ViT-B/16 models substantially outperform the baselines on long-tail benchmark *Let It Wag!* [39] classification task.

The inspiration for DynamiCS is a change of perspective from previous semantic data balancing approaches, which "aim for even" distributions of pre-training data over semantic categories. The "aim for even" philosophy is represented by MetaCLIP [45], which chops off the fat head, and DBP [2], which homogenizes densities with a comparable effect. Cluster scaling instantiates an "aim for utility" philosophy, which does not necessarily result in a flattened distribution. Specifically, it downsamples the semantic fat head, without aiming to completely eliminate it. At the same time, it upsamples the semantic long tail, which improves the VLM performance on long-tail concepts.

In sum, this paper makes the following contributions:

- We point out that data sampling should involve not only down- but also upsampling, in order to improve VLM performance on the long tail concepts.
- We show that dynamic sampling complements cluster scaling: by drawing a different subset of data at each epoch, it improves cross-epoch diversity and makes upsampling practical without increasing training cost.
- On the basis of these insights, we propose a dynamic cluster-based sampling approach (DynamiCS), which outperforms other cost-reducing VLM training approaches and is competitive with full-scale OpenCLIP [4] while requiring only about 3% of its training costs.

The paper is structured as follows: in Sec. 2, we cover related work. Then, we introduce DynamiCS (Sec. 3) and our experimental setup (Sec. 4). We provide further motivation with experimental analysis on a smaller-scale dataset (Sec. 5). In Sec. 6, we present experimental results on a large-scale dataset that demonstrate the advances of DynamiCS. Sec. 7 provides discussion and outlook.

2 Related Work

2.1 Vision-Language Models

Large-scale Vision-Language Models (VLMs) have demonstrated remarkable transferability [4,6,12,13,19,32,48,49], which forms the basis for their widespread applicability and effectiveness in downstream tasks. CLIP (Contrastive Language-Image Pre-Training) exemplifies this paradigm by learning visual–semantic embeddings from language through contrastive learning [27,32]. CLIP pre-training relies on large-scale datasets and can use millions or billions of image–text pairs.

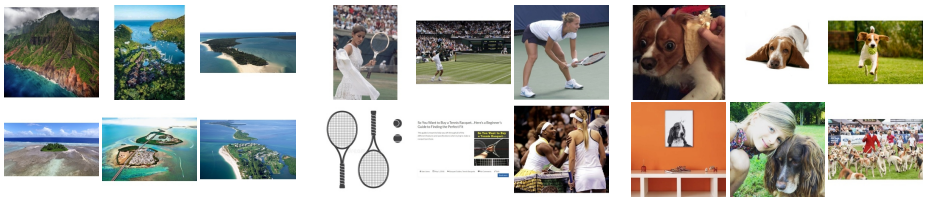

Fig. 2: Image examples from three semantic clusters (e.g., "sea," "tennis" "dog"), where visually similar concepts are grouped together.

Following CLIP [32], subsequent work, such as OpenCLIP [4] and MetaCLIP [45], has advanced the paradigm through open-source training and improved data curation strategies. Reproducible training with publicly accessible and diverse image–text pairs has become a key factor in advancing large-scale VLMs, which facilitate downstream applications. However training, as mentioned in Sec. 1, requires substantial resources. In this paper, we focus on reducing the costs of VLM training while preserving model effectiveness.

2.2 Cost-reducing VLM Training

Approaches that reduce the training costs of VLMs fall into two categories: cost-saving approaches that reduce the amount of information in each sample and data filtering approaches, which use data sampling to make the training set smaller. We discuss each in turn.

Reducing Image/Text Tokens. RECLIP [15] introduces the use of low-resolution images for CLIP pre-training followed by high-resolution fine-tuning, achieving a substantial reduction in training cost. Taking the same idea to the patch level, FLIP [19] uses random image masking, where only a subset of patches is used for training. Similar token number reduction methods based on image-token distribution have been explored with A-CLIP [46], GLIP [20], and CLIP-PGS [31], and based on word frequency with CLIPF [21]. CLIPA [18] investigates and combines different token-level strategies, including low-resolution images with truncation, random, block, and syntax-based text masking. All token-reducing pre-training methods achieve comparable zero-shot performance at substantially lower cost compared to full-scale training [18]. Reducing the number of tokens has become a common practice in cost-reducing VLM training; however, it does not account for semantic data balance. In this paper, we propose a dual-purpose approach for reducing VLM training costs while addressing semantic data balance through cluster scaling.

Training Data Filtering. The pre-training datasets used for CLIP are large-scale image–text collections crawled from the Internet, such as CC3M [37], CC12M [3], LAION-400M [35], LAION-2B [34], and DataComp [10]. These datasets are inherently noisy, contain many duplicates or near-duplicates, and are semantically highly imbalanced [30, 39, 44, 45]. Several studies have observed that CLIP's class-wise performance exhibits a log-linear relationship with class

frequency [30, 39, 44]. Recent studies focus on filtering semantically redundant images out of pre-training datasets [1, 42]. As mentioned in Sec. 1, DBP [2] is a dual-purpose approach that simultaneously seeks to reduce training cost and achieve semantic data balance. It first removes duplicates, then clusters and prunes the data based on a measure of cluster complexity, resulting in consistently better performance than random pruning. However, while pursuing the aim of uniform data density, it does not explicitly take into account long-tail concepts. Instead, long-tail clusters can lose a larger proportion of samples.

Other dual-purpose approaches combine training-cost reduction with the selection of high-quality data. DataComp [10] and DFN [9] use a pre-trained CLIP model to filter misaligned image–text pairs with low similarity scores. Synthetic captions generated with LLMs or VLLMs can also be used to improve alignment [8, 16, 17, 25, 43], however, this may reduce data diversity and ultimately limit model generalization [25]. As noted in Sec. 1, these approaches rely on the existence of another already-trained VLM. Overall, long-tail concepts remain underrepresented during VLM pre-training [39].

2.3 Dynamic Data Sampling

Previous data pruning methods utilize a fixed subset of the dataset during the training [1, 2, 45], after which the model only sees part of the training data. An alternative method chooses a different pruning criterion in different stages of pre-training. One example is ClusterClip [36], which balances the clusters within each batch during large language model training, but it does not aim to reduce the overall training cost. In our work, we balance the dataset of VLM training by subsampling high-frequency concepts and upsampling low-frequency concepts. During pre-training, data samples are dynamically drawn in each training epoch.

3 DynamiCS

In this section, we first discuss common issues with current cost-reducing VLM training, and then introduce our dynamic cluster data sampling method.

3.1 The Challenge of Semantic Data Balance

Fig. 2 provides an impression of the semantic structure of CLIP training data that is revealed via clustering. We generated image embeddings using the pre-trained DINOv2-ViT-B/16 [29] model and clustered them using K-means [1, 2]. Following the definition of "concept" [39] that includes object classes, i.e., lemmatized nouns in captions/generation prompts, we see in Fig. 2, that clusters can be associated with dominant concepts. For example, one cluster is associated with beach, view, sunset, and sea, which relate to the concept of "sea". Another cluster includes tennis, court, player, and ball, which are associated with "tennis".

However, CLIP training faces the challenge that training data contains a few very large clusters and a large number of small clusters. As mentioned in Sec. 1,

in the wild the distribution of topics in the data has a very fat head and a very long tail. Specifically, if we cluster DataComp [10] using image embeddings into 50k clusters, each cluster contains approximately 12k samples on average. However, the distribution is too imbalanced. We observe that some clusters contain more than 1 million points, while others have fewer than 1k. Other datasets, such as CC3M [37], YFCC-15M [38], and the LAION-400M dataset also exhibit an imbalance in concept frequency [2, 39]. Also, a similar long-tail concept distribution of the pre-training dataset of VLMs has been observed in previous works [30, 39, 44].

These observations concerning semantic balance motivate our introduction of a dual-purpose approach that simultaneously reduces training costs and addresses semantic data balance. However, we observe that existing methods designed to achieve semantic data balance, represented by MetaCLIP [45] and DBP [2], pursue an "aim for even" philosophy, mentioned in Sec. 1, which seeks to flatten the semantic distribution without devoting any specific attention to semantic categories that are not well represented, i.e., comprise the long tail. Because DynamiCS reduces training costs with dynamic training, we have the freedom to explore an "aim for utility" philosophy that downsamples to reduce, but not completely flatten, the fat head, and upsamples to ensure the long-tail is well represented. The "aim for utility" approach is supported by recent work that establishes that CLIP training is robust to long-tail distributions [44], meaning that it should not be necessary to aim for a completely flat distribution, and also by work that points to the continuing challenge of the long semantic tail [30, 39, 44].

3.2 Cluster Scaling

DynamiCS controls semantic balance with Eq. 1:

$$S_i = \frac{c_i^\alpha}{\sum_{j=1}^{N} c_j^\alpha} \cdot T \tag{1}$$

where N is the total number of clusters, c_i = original size of cluster i, $\alpha \geq 0$ is a scaling factor that controls how aggressively large clusters are down-sampled and small clusters are up-sampled, T is the target total number of samples across all clusters, S_i is the resampled size of cluster i. When $S_i > c_i$ (up-sampled clusters), samples are drawn with replacement.

The motivation for Eq. 1 is twofold: First, it provides the basic scaling we need, reducing the number of samples in large clusters and increasing the number in small clusters, while preserving the relative order of cluster sizes. Second, it allows us to control the shape of the semantic distribution using α, which can range from 0 to a large number. The case of $\alpha = 0$ represents the extreme of the "aim for even" philosophy in which all clusters contain the same number of samples. The case of $\alpha = 1$ reduces to random sampling, an important baseline for data sampling. When α is large, the formula mimics cases in which long-tail semantic clusters are neglected to the point of being eliminated, as happens in [1] under de-duplication thresholds corresponding to aggressive pruning.

We create the clusters using k-means with cosine similarity to group the image embeddings. After clustering the embeddings, some clusters remain semantically redundant because their centroids are close to each other in the embedding space. To reduce this redundancy, we merge centroids whose cosine similarity exceeds a threshold. The clustering approach is based on the assumption that semantic concepts are well represented in the image embedding space of pre-trained vision models. Our inspection of the clustering results, cf. Fig. 2 supports this assumption.

We note that Eq. 1 is reminiscent of the equation applied for balancing the training input for text models, specifically the sampling strategy applied in word2vec [24]. This resemblance reinforces the importance of our idea that the distribution should not be completely flattened, but rather that differences should be preserved so that the relative ordering of the original data is maintained. However, in word2vec the aim is a balance of text tokens using token reduction, and here we are balancing the distribution of samples, not tokens, with a data filtering approach.

3.3 Dynamic Sampling

Dynamic sampling randomly selects a different subset of S_i samples from the cluster in each epoch. In each model pre-training epoch, for each cluster i, S_i samples are drawn based on the scaling factor P_i defined as:

$$P_i = S_i/c_i \tag{2}$$

This sampling strategy reduces the number of samples seen from large clusters while ensuring that different samples within the cluster can be seen in each epoch. It increases the number of samples seen from small clusters by upsampling. Note that all samples within a cluster share the same probability of being selected.

4 Experimental Settings

We follow OpenCLIP [11], FLIP [19] and CLIPA [18] to pre-train and evaluate our model.

Dataset. We pre-train our models on datasets of different sizes. Specifically, we use DataComp (Large, 1.28B image-text pairs) [10], which is filtered with DFN [9] for the DynamiCS analysis (Sec.5) and all next experiments. We refer to this filtered dataset as DataComp-DFN. We also conduct DynamiCS on the LAION-400M [35] dataset. Both datasets are widely used for pre-training VLMs [4,9,10,15,18,19,43]. Due to expired URLs, we could download about 298 million samples from LAION-400M and 130 million samples from DataComp (out of 192 million candidates).

Architecture. We use ViT-B/16 and ViT-L/16 [7] as image encoders and a Transformer model [40] as the text encoder. The input image resolutions are 112×112 and 224×224, and the maximum text length is 32.

Table 1: Zero-shot classification on ImageNet-1K for different α values in Eq. 1. All models are pre-trained on **DataComp** for 106 million samples seen with ViT-B/16 image encoder and 112×112 image resolution.

	0.0	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**	**2.0**
ImageNet-1K	38.5	39.2	38.2	36.9	36.4	33.8	19.4
Let It Wag!	19.5	20.2	19.6	17.5	15.6	13.4	5.1

Training and Fine-tuning Setting. We pre-train the model on the DataComp and LAION-400M dataset for 1.28 billion and 2.56 billion samples seen, following DFN [9] and CLIPA [18]. To accelerate training and conserve computational resources as much as possible, we first pre-train at a small image resolution (112×112) with a batch size of 28k which has been shown to be an effective way to speed up CLIP [15, 18], and use this as our baseline. We then fine-tune the model for one additional step with a full-resolution image (224×224) to bridge the distribution gap between training and evaluation. We set the number of clusters to 50k for both DataComp and LAION-400M (Sec. 6). We use $\alpha = 0.2$ and set the target number of samples T to 50% of the dataset size. For post-clustering refinement, we use a cosine-similarity threshold of 0.7.

The model was trained on 2 nodes, each with 4 H100 GPUs, with identical settings to ensure that all models were run under consistent conditions. More details of the training and fine-tuning settings can be found in the Appendix.
Evaluation We adopt the evaluation settings and downstream datasets as CLIP and OpenCLIP [4, 32]. These datasets cover a broad range of modalities and domains, including natural images, fine-grained classification tasks, and cross-modal retrieval benchmarks to ensure the evaluation is both wide in scope and diverse. We also evaluate the models on the long-tail concepts dataset *Let It Wag!* [39], which consists of 290 long-tail categories curated from multiple VLMs' training datasets and 130K test samples collected from the web (448 images per category). For classification, we adopt the 80 prompts introduced in CLIP [32], which are also used for ImageNet-1K evaluation.

5 Experimental Analysis of DynamiCS

In this section, we carry out experiments on DataComp to demonstrate the contributions of dynamic sampling and cluster scaling and set key hyperparameters.
Scaling Factor α. As shown in Table 1, the model achieves its best performance when $\alpha = 0.2$. The result supports the conclusion that, in practice, we need to select a moderate α that balances the short-head and long-tail data.

Diving into more detail, we see that the "aim for even" case, where $\alpha = 0$ and each cluster selects the same number of samples, is outperformed by $\alpha = 0.2$. This result supports our "aim for utility", where the fat head is not completely eliminated and the long-tail is emphasized. Note that the $\alpha = 0$ still outperforms $\alpha = 1$, which corresponds to the dynamic random sampling approach where all samples have equal probability of being sampled.

Table 2: Comparison of zero-shot top-1 classification accuracy on ImageNet-1K. All models are pre-trained on **DataComp** for 0.64 billion samples seen with ViT-B/16 image encoder.

Models	Samples Seen@Resolution	ImageNet-1K	*Let It Wag!*
Random Pruning	0.64B@112 + 128M@224	64.5	35.5
Random-Dynamic		66.2	36.2
Cluster-Scaling (ours)		68.0	43.7
DynamiCS (ours)		69.2	46.5

When $\alpha = 2$, sampling favors large clusters and suppresses small ones, leading to a highly imbalanced distribution. Performance drops substantially as α increases from 0.2 to 2, falling from 39.2% to 19.4% on ImageNet-1K and from 20.2% to 5.1% on *Let It Wag!*. This drop can be attributed to the loss of representation of long-tailed concepts.

Overall, models trained with α values between 0.0 and 0.8 consistently outperform the random method ($\alpha = 1.0$), indicating that DynamiCS is relatively robust to the choice of α. Based on this robustness, we use $\alpha = 0.2$ for all large-scale experiments in Sec. 6, without concern that the experimental results are sensitive to this choice.

Cluster Scaling Helps Improve Long-tail Concepts Performance. As shown in Table 2, cluster scaling substantially improves pre-training under both random and dynamic sampling. It substantially increases performance by 3.5% and 3.0% on ImageNet-1K, and by 8.2% and 10.3% on *Let It Wag!*, respectively. The improvement can be attributed to DynamiCS enhancing the performance of long-tail concepts while maintaining that of head concepts.

Dynamic Sampling Increases Training Data Diversity. Then, we show the influence of dynamic sampling on CLIP pre-training. The results in Table 2 show the improvement that dynamic sampling delivers compared to standard, static sampling. DynamiCS outperforms the Random pruning method by 4.7% and 11.0% on the ImageNet-1K and *Let It Wag!* datasets. Here, random pruning means that it randomly selects a fixed 50% subset of the dataset. Random-Dynamic outperforms random pruning by 1.7% and 0.7% on the ImageNet-1K and *Let It Wag!* datasets. Here, Random-Dynamic means that each sample has a 50% probability of being selected during training. And DynamiCS outperforms Cluster-Scaling by 1.2% and 2.8% on the ImageNet-1K and *Let It Wag!* datasets. The gains of dynamic sampling can be attributed to an improvement in the diversity of the data used for training, compared with static sampling.

6 Comparative Experimental Results

In this section, we present an evaluation of DynamiCS trained on LAION-400M [35] and DataComp [9, 10] on the classic zero-shot ImageNet-1K task, on a long-tail task, and on a range of other datasets conventionally used in the literature to evaluate CLIP performance.

Table 3: Random sampling vs. DynamiCS. Results on ImageNet-1K and *Let It Wag!* under same training settings on LAION-400M and DataComp-DFN.

Models	Dataset (Data Size)	Samples Seen@Resolution	ImageNet-1K	*Let It Wag!*	GPU hours
Random	LAION-400M	1.28B@112 + 128M@224	59.8	31.9	151
DynamiCS (Ours)	(298M)	1.28B@112 + 128M@224	65.0	42.1	163
Random	DataComp-DFN	0.64B@112 + 128M@224	64.5	35.5	90
DynamiCS (Ours)	(130M)	0.64B@112 + 128M@224	69.2	46.5	95

Table 4: Comparison of zero-shot top-1 classification on ImageNet-1K. Our model was pre-trained on the LAION-400M dataset and a subset of **DataComp-Large** that was filtered by DFN-2B. FLIP is pre-trained with 75% image masking, resulting in the same number of image tokens as a 112×112 image size. CLIPA is pre-trained by syntax masking with 16 text tokens. The symbol $*$ indicates results we reproduced under the same low-resolution and training setting to enable fairer comparison wherever possible. The symbol $\approx$ indicates estimated values, because GPU hours are not reported in their paper. All models use the ViT-B/16 image encoder.

Models	Dataset (Data Size)	Samples Seen@Resolution	ImageNet-1K	*Let It Wag!*	GPU hours
OpenCLIP [4]	LAION-400M	2.56B@224	64.2	—	2140
FLIP [19]	(400M)	2.56B@224 + 128M@224	60.9	—	—
CLIPA* [18]	LAION-400M (298M)	2.56B@112 + 128M@224	63.2	36.4	269
RECLIP* [15]		2.56B@112 + 128M@224	62.9	36.0	280
DynamiCS (Ours)		1.28B@112 + 128M@224	**65.0**	42.1	163
DynamiCS (Ours)		2.56B@112 + 128M@224	**67.5**	45.5	299
DataComp [10]	DataComp	1.28B@224	63.1	33.7	$\approx$ 1070
Captioning [25]	(1.28B)	2.56B@224	59.8	—	$\approx$ 2140
WhatIf [17]	Recap-DataComp-1B (1.4B)	2.56B@112 + 128M@224	69.2	—	—
DFN [9]	DataComp-DFN (130M)	1.28B@224	67.8	—	$\approx$ 1070
HQ-CLIP [43]		3.20B@224	70.6	38.2	$\approx$ 2675
DFN* [9]		1.28B@112 + 128M@224	68.7	42.4	151
DynamiCS (Ours)		0.64B@112 + 128M@224	69.2	46.5	95
DynamiCS (Ours)		1.28B@112 + 128M@224	**71.3**	**50.2**	163

6.1 DynamiCS *vs.* Random Sampling

As shown in Table 3, DynamiCS outperforms random pruning by 5.2% and 4.7% on ImageNet-1K when pre-training on LAION-400M and DataComp, respectively. The improvements on the long-tail benchmark are larger (10.2% and 11.0%), indicating that DynamiCS improves coverage of long-tail concepts while also delivering clear improvements on ImageNet-1K.

6.2 Zero-shot Classification on ImageNet-1K

Cost-saving Baselines on LAION-400M: DynamiCS outperforms cost-saving baselines, RECLIP [15], FLIP [19], CLIPA [18] as demonstrated in Table 4, with around 50% of the samples seen and 60% GPU hours used by baselines. Specifically, DynamiCS-1.28B outperforms CLIPA by 1.8%, RECLIP by 2.1% and FLIP by 4.1% on ImageNet-1K while using only 50% of the total samples seen,

primarily due to improved performance on long-tail concepts (See Appendix). Notably, our dynamic sampling plays a key role in reducing training cost while preserving model quality.

Dual-purpose Filtering-based Baselines on DataComp: DynamiCS also outperforms different dual-purpose filtering-based baselines. We first reproduce the result for DFN [9] on the filtered DataComp-Large dataset, **DFN*** in Table 4. We then apply DynamiCS on the same filtered dataset. DynamiCS-0.64B improves over DFN* by 0.5% on ImageNet-1K while using only 50% of the training samples and 62% of the GPU hours. DynamiCS-1.28B also outperforms HQ-CLIP by 0.7% with about 40% of the samples and about 6% of the GPU hours. In addition, DynamiCS-0.64B surpasses the Captioning baseline by 9.4% on ImageNet-1K. DynamiCS achieves comparable or better ImageNet-1K performance than WhatIf [17] (recaptioning on DataComp-1B), while requiring substantially less computing resources.

6.3 Zero-shot Classification on Long-tail Dataset

To further investigate whether upsampling long-tail concepts can improve the performance of low-frequency classes. We further evaluate DynamiCS on the long tail dataset, which is *Let It Wag!* [39], as shown in Table 4 demonstrates that DynamiCS can substantially enhance the long-tail performance, thereby contributing to the model's overall performance.

DynamiCS-1.28B pre-trained on LAION-400M outperforms the baseline model RECLIP by 6.1% and CLIPA by 5.7% on the *Let It Wag!* dataset, while maintaining comparable performance on ImageNet and using only about 50% of the total samples seen and 60% GPU hours. Furthermore, DynamiCS-0.64B pre-trained on DataComp-DFN substantially outperforms the DataComp (Image-based $\cap$ CLIP score) filtering methods that are pre-trained on the DataComp dataset by 12.8%. DynamiCS-0.64B also outperforms HQ-CLIP by 8.3% on the long-tail dataset with only 20% samples seen and 3.6% GPU hours. Overall, DynamiCS boosts long-tail performance at far lower training cost, outperforming large-scale models and baselines by a wide margin.

6.4 Data Scaling

DynamiCS has shown a substantial improvement over random pruning, cost-saving baselines, and dual-purpose baselines on both ImageNet-1K and *Let It Wag!*, while using about 60% of the GPU hours. We further scale DynamiCS from 0.64B to 2.56B samples seen on LAION-400M and DataComp. As shown in Table 5, on LAION-400M, ImageNet-1K accuracy increases from 61.5% to 67.5%, and *Let It Wag!* accuracy increases from 37.3% to 45.5%. On DataComp, ImageNet-1K accuracy improves from 69.2% to 72.6%, and *Let It Wag!* improves from 46.5% to 52.0%. Overall, the scaling experiments show a clear and consistent improvement as the number of training samples increases.

Table 5: Data scaling results for DynamiCS. Zero-shot top-1 accuracy on ImageNet-1K and *Let It Wag!* when scaling the number of training samples from 0.64 billion to 2.56 billion samples seen on LAION-400M and DataComp-DFN.

Models	Dataset (Data Size)	Samples Seen@Resolution	ImageNet-1K	*Let It Wag!*	GPU hours
DynamiCS (Ours)	LAION-400M (298M)	0.64B@112 + 128M@224	61.5	37.3	95
		1.28B@112 + 128M@224	65.0	42.1	163
		2.56B@112 + 128M@224	67.5	45.5	299
	DataComp-DFN (130M)	0.64B@112 + 128M@224	69.2	46.5	95
		1.28B@112 + 128M@224	71.3	50.2	163
		2.56B@112 + 128M@224	**72.6**	**52.0**	299

Table 6: Zero-shot top-1 classification on ImageNet-1K and *Let It Wag!* with the full training CLIP. All models use the ViT-B/16 image encoder.

Models	Dataset (Data Size)	Samples Seen@Resolution	Tokens	ImageNet-1K	*Let It Wag!*	GPU hours
OpenAI-WIT [32]	—	12.8B@224	274	68.3	37.9	10700
MetaCLIP-400M [45]	(400M)	12.8B@224	274	70.8	46.5	≈ 10700
OpenCLIP [4]	LAION-400M	12.8B@224	274	67.1	39.1	10736
LaCLIP [8]	(400M)	12.8B@224	274	69.4	48.4	≈ 10700
DynamiCS (Ours)	LAION-400M (298M)	2.56B@112 + 128M@224	81	67.5	45.5	299
OpenVision [16]	Recap-DataComp-1B (1.4B)	12.8B@160 + 1.024B@224 + 256M@336	180	73.9	—	—
DynamiCS (Ours)	DataComp-DFN (130M)	1.28B@112 + 128M@224	81	71.3	50.2	163
		2.56B@112 + 128M@224	81	72.6	52.0	299

6.5 Comparison with Full-scale Training CLIP Baselines

As shown in Table 6, we compare DynamiCS with the full-scale training CLIP baselines, which are pre-trained with 224×224 image resolution and 12.8 billion samples seen. Our DynamiCS-2.56B has the best performance on both ImageNet-1K and *Let It Wag!* datasets. Surprisingly, DynamiCS pre-trained on DataComp with 2.56B samples seen, with only about 3% of the training cost, outperforms OpenAI-WIT CLIP, OpenCLIP, and MetaCLIP by 4.3%, 5.5%, 1.8% on ImageNet-1K and by 14.1%, 12.9%, 5.5% on the *Let It Wag!* dataset. Moreover, DynamiCS-1.28B pre-trained on DataComp with 1.28B samples already surpasses full-training baselines using only 163 GPU hours.

6.6 Zero-shot Robustness

We evaluate DynamiCS on 6 robustness datasets following the evaluation of CLIP [32] and OpenCLIP [4] as shown in Table 7. On LAION-400M, DynamiCS-1.28B outperforms CLIPA by 1.1% and RECLIP by 1.5%, while using only 50% of the samples seen by RECLIP. It also improves over the random baseline by 4.6%. DynamiCS-2.56B remains below full-training CLIP, with a gap of 0.4%.

On DataComp, DynamiCS performs well compared with CLIP-score-based filtering methods. DynamiCS-0.64B already outperforms DataComp by 4.9% while using only 50% of the samples and 8% of the GPU hours. DynamiCS-1.28B improves over the DFN baseline by 2.2% and over HQ-CLIP by 1.0%. Both

Table 7: Zero-shot robustness evaluation of DynamiCS and other methods on the different robustness datasets.

Models	Dataset (Data Size)	Samples Seen@Resolution	IN-A	IN-O	IN-R	IN-S	IN-V2	ON	Avg.	GPU hours
OpenAI-WIT [32]	—	12.8B@224	50.0	42.3	77.7	48.2	55.3	61.9	55.9	10700
MetaCLIP-400M [45]	(400M)	12.8B@224	–	–	–	–	–	–	57.5	≈ 10700
OpenCLIP [4]	LAION-400M (400M)	12.8B@224	33.2	50.8	77.9	52.4	50.8	59.6	54.1	10736
RECLIP* [15]	LAION-400M (298M)	2.56B@112 + 128M@224	26.1	53.5	72.8	48.1	55.2	47.2	50.5	280
CLIPA* [18]		2.56B@112 + 128M@224	26.7	54.1	73.3	48.6	55.5	47.2	50.9	269
Random		1.28B@112 + 128M@224	21.6	53.7	69.3	44.9	52.0	43.1	47.4	151
DynamiCS (Ours)		1.28B@112 + 128M@224	28.9	56.1	73.2	49.2	56.4	47.9	52.0	163
DynamiCS (Ours)		2.56B@112 + 128M@224	30.9	53.5	76.3	51.7	59.2	50.7	53.7	299
DataComp [10]	DataComp-Large (1.28B)	1.28B@224	25.5	49.6	71.8	49.8	55.1	53.1	50.8	≈ 1070
DFN [9]	DataComp-DFN (130M)	1.28B@224	–	–	–	–	–	–	54.0	≈ 1070
HQ-CLIP [43]		3.20B@224	39.1	43.0	80.1	57.3	63.1	60.6	57.2	≈ 2675
DFN* [9]		1.28B@112 + 128M@224	31.4	59.7	74.5	53.9	60.9	55.5	56.0	151
Random		0.64B@112 + 128M@224	23.0	59.5	70.3	49.3	56.4	51.4	51.7	90
DynamiCS (Ours)		0.64B@112 + 128M@224	30.4	62.3	73.2	53.2	60.7	54.6	55.7	95
DynamiCS (Ours)		1.28B@112 + 128M@224	35.7	58.4	76.6	55.9	64.0	58.4	58.2	163
DynamiCS (Ours)		2.56B@112 + 128M@224	37.2	57.6	79.0	57.4	64.1	59.3	59.1	299

Table 8: Zero-shot image-text retrieval evaluation of DynamiCS and other methods on the COCO [22] and Flickr30k [47]. Report the average recall@1 of the image-to-text and text-to-image retrieval results.

Models	Dataset (Data Size)	Samples Seen@Resolution	COCO	Flickr30k	GPU hours
OpenAI-WIT [32]	—	12.8B@224	42.8	72.2	10700
MetaCLIP-400M [45]	(400M)	12.8B@224	48.2	76.7	≈ 10700
OpenCLIP [11]	LAION-400M (400M)	12.8B@224	46.9	74.6	10736
FLIP [19]		2.56B@224 + 128M@224	–	–	—
RECLIP* [15]	LAION-400M (298M)	2.56B@112 + 128M@224	44.4	72.4	280
CLIPA* [18]		2.56B@112 + 128M@224	44.4	71.7	269
Random		1.28B@112 + 128M@224	42.7	69.4	151
DynamiCS (Ours)		1.28B@112 + 128M@224	45.1	72.7	163
DynamiCS (Ours)		2.56B@112 + 128M@224	46.5	73.9	299
DataComp [10]	DataComp-Large (1.28B)	1.28B@224	39.8	63.6	≈ 1070
HQ-CLIP [43]	DataComp-DFN (130M)	3.20B@224	52.2	77.9	≈ 2675
DFN* [9]		1.28B@112 + 128M@224	44.3	68.1	151
Random		0.64B@112 + 128M@224	39.6	62.3	90
DynamiCS (Ours)		0.64B@112 + 128M@224	42.5	66.6	95
DynamiCS (Ours)		1.28B@112 + 128M@224	44.2	69.3	163
DynamiCS (Ours)		2.56B@112 + 128M@224	46.2	71.3	299

DynamiCS-1.28B and DynamiCS-2.56B also outperform fully trained OpenAI-WIT, MetaCLIP, and OpenCLIP. In particular, DynamiCS-2.56B exceeds full training on OpenAI-WIT by 3.2%, MetaCLIP by 1.6%, and OpenCLIP by 5.0%, while using only 3% of the GPU hours. Overall, DynamiCS achieves comparable robustness results while using substantially fewer GPU hours, demonstrating strong robustness and improved data efficiency.

Table 9: Zero-shot top-1 classification accuracy across 22 benchmark classification datasets.

Models	Dataset (Data Size)	Samples Seen @Resolution	Food-101	CIFAR-10	CIFAR-100	CUB200	SUN397	Cars	Aircraft	VOC2007	DTD	OxfordPets	Caltech-101	Flowers102	MNIST	STL10	EuroSAT	Resisc45	GTSRB	KITTI	Country211	PCAM	CLEVR	SST2	GPU hours
OpenAI-WIT [32]	—	12.8B@224	88.7	90.8	67.0	—	64.8	58.2	24.2	78.3	45.0	88.9	89.0	—	51.4	98.3	55.9	60.7	43.4	26.4	22.8	50.7	21.2	50.7	10700
MetaCLIP-400M [45]	(400M)	12.8B@224	87.3	90.1	66.6	—	66.8	74.2	28.4	72.2	55.9	90.4	93.4	72.3	47.9	97.2	55.7	66.2	43.8	24.2	22.6	62.0	30.1	62.0	≈ 10700
OpenCLIP [4]	LAION-400M (298M)	12.8B@224	86.1	91.7	71.2	—	69.6	—	17.7	76.8	51.3	89.2	91.3	—	66.2	—	50.2	58.5	43.5	18.1	18.1	59.6	28.7	54.4	10736
RECLIP* [15]		2.56B@112 + 128M@224	81.3	92.1	71.8	55.1	66.0	80.7	13.0	75.3	49.4	85.3	82.1	64.6	65.7	96.3	45.5	55.5	33.5	21.7	14.4	48.9	17.7	49.8	280
CLIPA* [18]		2.56B@112 + 128M@224	81.5	91.8	70.3	57.2	65.8	81.1	14.3	76.6	46.8	85.9	82.8	61.7	41.8	96.3	55.9	55.5	38.7	21.2	14.5	50.7	18.8	53.0	269
DynamiCS (Ours)		2.56B@112 + 128M@224	84.0	92.9	75.0	68.9	67.5	75.7	16.8	76.8	52.9	87.7	83.7	70.4	53.1	96.3	42.9	59.1	44.1	15.6	16.2	48.2	14.0	52.9	299
DataComp [10]	DataComp (1.28B)	1.28B@224	83.1	93.8	75.4	47.0	64.3	77.2	10.0	80.9	46.9	83.5	89.7	64.0	54.0	95.8	50.1	52.7	43.4	40.1	14.3	49.7	23.1	52.9	≈ 1070
HQ-CLIP [43]	DataComp-DFN (130M)	3.20B@224	87.8	96.2	81.0	—	69.7	85.3	11.3	78.8	51.5	89.5	93.1	69.0	77.7	98.1	47.6	60.6	54.4	43.0	15.9	47.5	27.5	51.7	≈ 2675
DFN* [9]		1.28B@112 + 128M@224	85.8	95.1	79.4	58.1	66.1	86.6	14.2	76.7	46.3	89.9	84.8	73.2	64.2	96.8	46.2	48.6	42.2	35.7	12.3	52.6	17.7	49.5	151
Random		0.64B@112 + 128M@224	82.7	93.7	77.3	50.7	62.9	83.7	10.1	76.9	40.9	86.7	84.0	70.6	53.3	95.7	47.3	42.7	22.4	30.8	10.8	57.7	14.9	50.4	90
DynamiCS (Ours)		0.64B@112 + 128M@224	84.6	94.0	79.5	65.2	65.4	74.7	14.2	76.4	42.0	88.4	84.5	75.9	41.2	96.0	45.8	44.8	25.5	21.1	12.1	61.6	14.7	46.6	95
DynamiCS (Ours)		1.28B@112 + 128M@224	86.9	95.2	81.3	70.8	66.7	78.3	16.5	76.5	45.4	89.9	84.3	83.5	53.2	96.7	48.9	55.5	31.1	11.7	13.5	42.1	22.2	50.1	163
DynamiCS (Ours)		2.56B@112 + 128M@224	86.7	96.2	81.8	71.3	68.1	80.8	19.2	74.6	49.2	90.7	84.4	79.0	66.8	97.3	44.2	56.7	41.5	31.5	14.4	56.5	27.2	48.9	299

6.7 Image-Text Retrieval

Table 8 reports zero-shot retrieval results on COCO [22] and Flickr30k [47]. Overall, DynamiCS consistently outperforms RECLIP, CLIPA, and the Random baseline on both LAION-400M and DataComp. On LAION-400M, DynamiCS-2.56B exceeds fully trained OpenAI-WIT, but remains below OpenCLIP and MetaCLIP. In contrast to the zero-shot classification results, pre-training on DataComp does not provide a clear advantage over LAION-400M for image–text retrieval. HQ-CLIP outperforms DynamiCS, likely because its generated captions are longer and more detailed, which better match the needs of retrieval tasks. DynamiCS can also be applied to HQ-CLIP's synthetic data to further improve retrieval performance, since HQ-CLIP does not explicitly balance the data.

6.8 Zero-shot Classification on Other Datasets

DynamiCS consistently outperforms random pruning and achieves performance comparable to RECLIP and DFN, as shown in Table 9. Interestingly, compared with these baselines, DynamiCS shows a clear advantage on fine-grained recognition tasks such as CUB-200-2011 [41], which contains 200 bird species primarily from North America. On LAION-400M, DynamiCS-2.56B outperforms RECLIP by 13.8% and CLIPA by 11.7%. On DataComp, DynamiCS-0.64B exceeds the random baseline by 14.5%, and DynamiCS-1.28B surpasses DFN* by 12.7%.

On the other fine-grained Flowers102 dataset [26] with 102 flower categories. On LAION-400M, DynamiCS again improves effectiveness substantially, outperforming RECLIP by 5.8% and CLIPA by 8.7%. On DataComp, DynamiCS-0.64B outperforms random by 5.3% and DynamiCS-1.28B outperforms DFN* by 10.3%. DynamiCS-1.28B also improves substantially over the results of models filtered and pre-trained on the DataComp dataset, by 19.5%. We leave further exploration to future work.

Table 10: Zero-shot top-1 classification accuracy with ViT-L/16 encoder on ImageNet-1K. We pre-train the model at an image resolution of 112×112 with 16 text tokens (using syntax masking) for 1.28B samples seen, and then fine-tune it in small steps, reducing the total samples seen to 50% of that of CLIPA.

Models	Dataset (Data Size)	Samples Seen @Resolution	ViT-L/16 ImageNet-1K	ViT-L/16 *Let It Wag!*	GPU hours
CLIPA [18]	LAION-400M (400M)	2.56B@112 + 128M@224	68.8	–	> 625
CLIPA + DynamiCS (Ours)	LAION-400M (298M)	1.28B@112 + 128M@224	70.3	44.5	300

6.9 Larger Vision Transformer

To demonstrate scalability, we scale the model from ViT-B/16 to ViT-L/16, which has about 2.9× more parameters. As shown in Table 10, DynamiCS maintains strong performance and outperforms CLIPA by 1.5% on ImageNet, while using only 50% of the samples seen by CLIPA.

7 Conclusion and Outlook

This paper has proposed a dual-purpose data sampling approach to reduce the training costs of VLMs, important for practical applications, while at the same time balancing the availability of semantic data that is key to performance. Extensive experiments demonstrate that DynamiCS achieves better full-scale effectiveness at much lower cost while outperforming other low-cost pre-training alternatives. Its effectiveness on both LAION-400M and DataComp-DFN suggests it can extend to other imbalanced datasets to reduce training cost.

DynamiCS is based on two established insights. First, we establish that data sampling should involve both down- and upsampling, to maintain VLMs' performance on the long tail. The cost savings achieved with dynamic sampling give us the freedom to integrate upsampling into our approach. The novelty of DynamiCS is that it breaks with the "aim for even" philosophy, which cuts off the fat head in an attempt to flatten the distribution, and instead pursues an "aim for utility" philosophy. Second, dynamic sampling makes an important contribution to efficient VLM pre-training: it is very effective at reducing the cost of VLMs, and we expect it to be an element in future methods as well.

Acknowledgements

This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-17261. This work was supported by the CiCS project of the research programme Gravitation (024.006.037).

References

1. Abbas, A., Tirumala, K., Simig, D., Ganguli, S., Morcos, A.S.: SemDeDup: Data-efficient learning at web-scale through semantic deduplication. arXiv preprint arXiv:2303.09540 (2023)
2. Abbas, A.K.M., Rusak, E., Tirumala, K., Brendel, W., Chaudhuri, K., Morcos, A.S.: Effective pruning of web-scale datasets based on complexity of concept clusters. In: International Conference on Learning Representations (2024)
3. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
4. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)
6. Desai, K., Johnson, J.: VirTex: Learning visual representations from textual annotations. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
8. Fan, L., Krishnan, D., Isola, P., Katabi, D., Tian, Y.: Improving CLIP training with language rewrites. In: Advances in Neural Information Processing Systems (2023)
9. Fang, A., Jose, A.M., Jain, A., Schmidt, L., Toshev, A.T., Shankar, V.: Data filtering networks. In: International Conference on Learning Representations (2024)
10. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., Orgad, E., Entezari, R., Daras, G., Pratt, S.M., Ramanujan, V., Bitton, Y., Marathe, K., Mussmann, S., Vencu, R., Cherti, M., Krishna, R., Koh, P.W., Saukh, O., Ratner, A., Song, S., Hajishirzi, H., Farhadi, A., Beaumont, R., Oh, S., Dimakis, A., Jitsev, J., Carmon, Y., Shankar, V., Schmidt, L.: DataComp: In search of the next generation of multimodal datasets. In: Advances in Neural Information Processing Systems, Datasets and Benchmarks Track (2023)
11. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: OpenCLIP (2021)
12. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning (2021)
13. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020)
14. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning (2022)

15. Li, R., Kim, D., Bhanu, B., Kuo, W.: RECLIP: Resource-efficient CLIP by training with small images. Transactions on Machine Learning Research (2023)
16. Li, X., Liu, Y., Tu, H., Xie, C.: OpenVision: A fully-open, cost-effective family of advanced vision encoders for multimodal learning. In: IEEE/CVF International Conference on Computer Vision (2025)
17. Li, X., Tu, H., Hui, M., Wang, Z., Zhao, B., Xiao, J., Ren, S., Mei, J., Liu, Q., Zheng, H., Zhou, Y., Xie, C.: What if we recaption billions of web images with LLaMA-3? In: International Conference on Machine Learning (2025)
18. Li, X., Wang, Z., Xie, C.: An inverse scaling law for CLIP training. In: Advances in Neural Information Processing Systems (2023)
19. Li, Y., Fan, H., Hu, R., Feichtenhofer, C., He, K.: Scaling language-image pre-training via masking. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
20. Liang, M., Larson, M.: Centered masking for language-image pre-training. In: Machine Learning and Knowledge Discovery in Databases. Research Track and Demo Track (2024)
21. Liang, M., Larson, M.: Frequency is what you need: Considering word frequency when text masking benefits vision-language model pre-training. In: IEEE/CVF Winter Conference on Applications of Computer Vision (2026)
22. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision (2014)
23. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems (2023)
24. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems (2013)
25. Nguyen, T., Gadre, S.Y., Ilharco, G., Oh, S., Schmidt, L.: Improving multimodal datasets with image captioning. In: Advances in Neural Information Processing Systems (2023)
26. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Indian Conference on Computer Vision, Graphics and Image Processing (2008)
27. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
28. OpenAI, Achiam, J., Adler, S., et al., S.A.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2024)
29. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
30. Parashar, S., Lin, Z., Liu, T., Dong, X., Li, Y., Ramanan, D., Caverlee, J., Kong, S.: The neglected tails in vision-language models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
31. Pei, G., Chen, T., Wang, Y., Cai, X., Shu, X., Zhou, T., Yao, Y.: Seeing what matters: Empowering CLIP with patch generation-to-selection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)

32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021)
33. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125 (2022)
34. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C.W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S.R., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5B: An open large-scale dataset for training next generation image-text models. In: Advances in Neural Information Processing Systems, Datasets and Benchmarks Track (2022)
35. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
36. Shao, Y., Li, L., Fei, Z., Yan, H., Lin, D., Qiu, X.: Balanced data sampling for language model training with clustering. In: Findings of the Association for Computational Linguistics: ACL 2024 (2024)
37. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Annual Meeting of the Association for Computational Linguistics (2018)
38. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: YFCC100M: The new data in multimedia research. Communications of the ACM (2016)
39. Udandarao, V., Prabhu, A., Ghosh, A., Sharma, Y., Torr, P., Bibi, A., Albanie, S., Bethge, M.: No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. In: Advances in Neural Information Processing Systems (2024)
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (2017)
41. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 dataset. Tech. rep., California Institute of Technology (2011)
42. Webster, R., Rabin, J., Simon, L., Jurie, F.: On the de-duplication of LAION-2B. arXiv preprint arXiv:2303.12733 (2023)
43. Wei, Z., Wang, G., Ma, X., Mei, K., Chen, H., Jin, Y., Rao, F.: HQ-CLIP: Leveraging large vision-language models to create high-quality image-text datasets and CLIP models. In: IEEE/CVF International Conference on Computer Vision (2025)
44. Wen, X., Zhao, B., Chen, Y., Pang, J., QI, X.: What makes CLIP more robust to long-tailed pre-training data? a controlled study for transferable insights. In: Advances in Neural Information Processing Systems (2024)
45. Xu, H., Xie, S., Tan, X., Huang, P.Y., Howes, R., Sharma, V., Li, S.W., Ghosh, G., Zettlemoyer, L., Feichtenhofer, C.: Demystifying CLIP data. In: International Conference on Learning Representations (2024)
46. Yang, Y., Huang, W., Wei, Y., Peng, H., Jiang, X., Jiang, H., Wei, F., Wang, Y., Hu, H., Qiu, L., et al.: Attentive mask CLIP. In: IEEE/CVF International Conference on Computer Vision (2023)
47. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics (2014)

48. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: IEEE/CVF International Conference on Computer Vision (2023)
49. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference (2022)