Concept-Aware Batch Sampling Improves Language-Image Pretraining

Adhiraj Ghosh[1] Vishaal Udandarao[1,2*] Thao Nguyen[3*] Matteo Farina[1,4*] Mehdi Cherti[5]
Jenia Jitsev[5] Sewoong Oh[3] Elisa Ricci[4] Ludwig Schmidt[6] Matthias Bethge[1]
[1]Tübingen AI Center, University of Tübingen [2]University of Cambridge [3]University of Washington
[4]University of Trento [5]LAION [6]Stanford University

Abstract

What data should a vision-language model be trained on? To answer this question, many data curation efforts center on the quality of a dataset. However, most of these existing methods are (i) offline, i.e. they produce a static dataset from a set of predetermined filtering criteria, and (ii) concept-agnostic, i.e. they use model-based filters which induce additional data biases. In this work, we go beyond such offline, concept-agnostic methods and advocate for more flexible, task-adaptive online concept-based curation. Our first contribution is DATACONCEPT*, a collection of* 128*M web-crawled image-text pairs annotated with fine-grained details about their concept composition. Building on* DATACONCEPT*, we introduce* ***C****oncept-****A****ware* ***B****atch* ***S****ampling (*CABS*), a simple yet effective batch-sampling framework that flexibly constructs batches on-the-fly based on specific target distributions. We propose two variants: (i) Diversity Maximization (*CABS-*DM) to curate batches with a broad coverage of concepts, and (ii) Frequency Maximization (*CABS-*FM) to curate batches with high object multiplicity. Through extensive evaluations across 28 benchmarks, we demonstrate that* CABS *significantly benefits Language-Image Pretraining (LIP) and yields highly performant models. Overall,* CABS *represents a strong open-source alternative to proprietary online curation algorithms, enabling practitioners to define custom concept distributions that optimize for specific downstream tasks.* DATACONCEPT*, checkpoints and code are available at cabs-vlp.github.io.*

1. Introduction

Web-scale pretraining datasets underlie the impressive generalization capabilities of vision-language models (VLMs). The advent of CLIP [66], trained on 400M image-text pairs, motivated the open development of billion-scale datasets like LAION-5B [71] or DataComp-12.8B [31]. Although dataset size is an influential factor, their *quality* is equally important, if not more [31, 35, 57]. To improve quality, current curation methods range from filtering according to well-defined metrics (*e.g.*, CLIP score) [31] to synthetically augmenting the captions to be more descriptive [49, 58].

However, most of the widely adopted curation strategies (*e.g.*, those benchmarked by DataComp [31]), focus on quality only at the level of individual samples, overlooking the finer, *concept-level distribution*[1] within web-scale datasets. In other words, existing curation methods tend to be *concept-agnostic* (MetaCLIP [87] is a notable exception). Additionally, these methods operate in an *offline* manner, filtering out large portions of data, thus enforcing a *fixed* design choice: once data is discarded, it is difficult, if not impossible to repurpose the resulting subset for other curation strategies. The offline filtering regime also accelerates the depletion of available training samples, creating data scarcity that ultimately imposes a "data wall" on pretraining [60]. Finally, concept-agnostic filtering methods often rely on state-of-the-art, but black-box, models to guide curation. This not only reduces transparency in selection criteria but also risks propagating the model's biases into the curated dataset [34, 40]. In contrast, concept-aware curation provides both transparency and direct control over the composition of the final dataset.

In this work, we depart from such offline sample-level curation protocols, and instead advocate for more flexible *online concept-based curation*. Our rationale is simple: there is no "universal" notion of quality [36, 52], and importantly, as shown in Fig. 1 *(left)*, different downstream evaluations might bias what the optimal concept distribution should look like [3, 55]. Therefore, we aim to show that incorporating concept-level information *during* pretraining, without discarding any data *a priori*, provides a complementary and effective avenue for multimodal data curation. This aligns with recent works advocating for data reuse over filtering [59, 65].

To achieve this goal, we introduce DATACONCEPT: a multimodal pretraining dataset with 128M image-text pairs fully annotated with grounded concept information. In DATACONCEPT, each sample comes with ① semantic concepts, ② bounding boxes, ③ per-concept confidence scores, and

*Equal contribution. Ordered by increasing performance on GSM8K.

[1]We adopt the definition of *concepts* from [78], i.e. objects that can be found in the wild, that we can identify and locate in image samples.

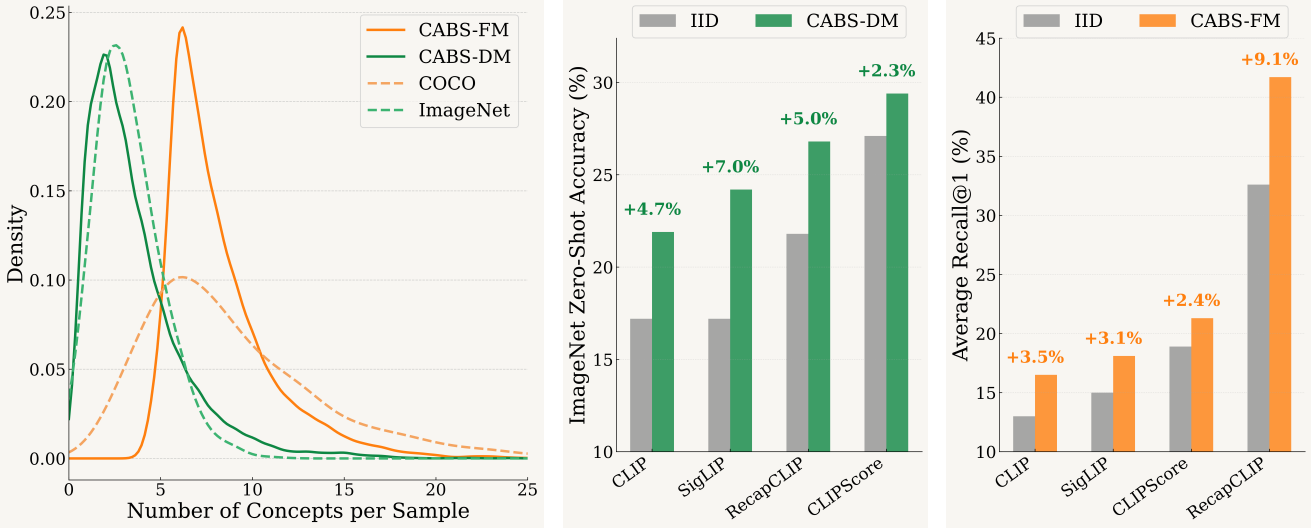

Figure 1. **Task-adaptive, steerable, Concept-Aware Batch Sampling (CABS).** The per-sample concept multiplicities *(left)* of MSCOCO retrieval and ImageNet classification test sets depict their divergent distributional properties. By only modifying a simple scoring function, CABS can flexibly adapt to different target tasks (details in Sec. 3). Both our classification-optimized (CABS-DM, see Sec. 4) and retrieval-optimized (CABS-FM, see Sec. 5) variants outperform IID sampling by large margins, across several experimental configurations.

④ concept-driven synthetic captions. With DATACONCEPT, we ask: *How can we effectively modulate different visual concepts during vision-language pretraining?*

In order to answer this question, we introduce a new training framework: **C**oncept-**A**ware **B**atch-**S**ampling (CABS). In contrast to offline, static curation, we do *not* impose a fixed, predetermined data distribution, but rather enable flexible, task-adaptive control over *online concept-based batch creation*. Our classification-optimized variant, CABS-Diversity Maximization (CABS-DM), selects samples based on *concept-diversity*. This scheme is in line with MetaCLIP's approach and significantly benefits zero-shot classification (see Fig. 1 (middle)), especially over *long-tailed* evaluations. Our second variant, specifically tailored to benefit image-text retrieval tasks (see Fig. 1 (right)), is CABS-Frequency Maximization (CABS-FM). It optimizes for *concept-multiplicity*—selecting samples that encompass a higher number of objects. To our best knowledge, these CABS-variants represent the first reproducible demonstration of task-adaptive online batch sampling. Taken together, our **contributions** are:

1. DATACONCEPT: a new, *concept-centric* pretraining dataset for VLMs comprising 128M samples. Each sample comes with fine-grained concept annotations and a concept-grounded synthetic caption. This helps enable further exploration of concept-centric data curation, a relatively underexplored avenue.
2. CABS: a new framework for vision-language pretraining that involves *online data curation* through *concept-aware batch sampling*. Paired with DATACONCEPT, CABS enables flexible control over the concept distribution of the data used throughout training.
3. Extensive experiments with 28 tasks, 4 visual backbones, and 2 training objectives (CLIP vs SigLIP), demonstrate that CABS variants are highly effective for vision-language pretraining (up to 7% gain on ImageNet zero-shot classification and up to 9.1% gain on image-text retrieval, over strong baselines), while being complementary to existing offline data curation recipes.

2. Concept-Aware Dataset Augmentation

We now introduce DATACONCEPT, our large-scale, *concept-annotated* pool of 128M image-text pairs. We will demonstrate the utility of our annotations by describing how they fit into the CABS framework in the next section (Sec. 3).

Initial pool. We start with DataComp's unfiltered medium pool consisting of 128M image-text pairs [31]. We denote each sample i as $(\mathcal{I}_i, \mathcal{T}_i)$. The standard protocol for downloading the dataset suffers from significant link-rot.[2] Hence, we opt for randomly sampling a 128M subset from Datacomp's XLarge pool (which consists of 12.8B samples).

Building a concept bank. The first step for annotating our pool is determining a *concept bank*, *i.e.*, the set of concepts that we seek to detect and tag. Previous work [78] curated a concept bank but it is rather limited $(4,029)$ due to being constructed from 27 evaluation datasets. For broader coverage, we further source concepts from the class labels used in RAM++ [41], V3Det [81], and OpenImages [47], resulting in $19,261$ concepts, after de-duplication and safety removal (specific details and methods are provided in Appx. A.1).

[2]We successfully downloaded only 79% of the medium-scale pool, as of September 2024.

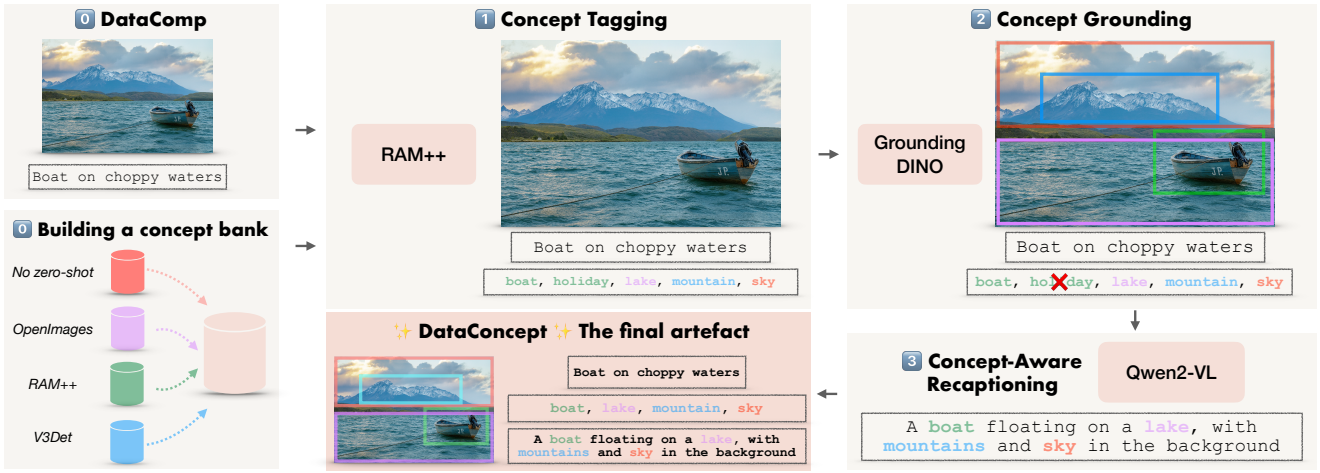

Figure 2. **DataConcept.** We start with images from DataComp [31] and build a concept bank $\mathcal{V}$ by merging, deduplicating, and filtering various concept sources. In ① *First-order tagging*, we assign a preliminary list of concepts (from $\mathcal{V}$) to each sample. ② We then *ground* each concept in the image, removing noise in the initial candidates. ③ Lastly, we use a model to transform alt-texts into *concept-aware captions*.

Concept tagging. Equipped with an expansive concept bank, following [78], we employ the RAM++ model to provide multiple concept tags for each sample in our data pool.

Concept grounding. While RAM++ annotations provide fine-grained concept annotations per sample, we find that *(i)* RAM++ can be miscalibrated in its confidence predictions due to the extreme diversity of our concept bank, and *(ii)* RAM++ only provides a list of concept tags, without localising them in the image, which can lead to incorrect grounding. Thus, we use GroundingDINO [51] to additionally provide concept-specific bounding boxes. To enable precise localization of concepts, we propose two methods: *(i) Confidence seeding*: we feed RAM++ concept tags per sample (only those with at least 0.75 confidence) as seed prompts to GroundingDINO, and *(ii) Resolution ensembling*: we use Weighted Box Fusion [73] to ensemble GroundingDINO predictions over multiple image resolutions of $\{384, 512, 800, 1000\}$. This helps reduce hallucinations without significantly affecting latency. With the two aforementioned steps, we obtain a list of concepts and their corresponding bounding boxes and confidence scores for each sample. Across all samples in the pool, we end up with $12,253$ concepts, i.e. $\mathcal{V}$, the final concept vocabulary for CABS. Each sample i is now tagged with a concept set $\mathcal{C}_i$.

Concept-aware recaptioning. We augment each sample i with a *concept-aware* synthetic caption. Synthetic recaptions have been shown to improve training data quality by reducing noise in alt-texts [27, 28, 58]. We use Qwen2-VL-7B [82] to recaption each image in a *concept-aware manner*: for each sample i, we provide the list of detected concepts $\mathcal{C}_i$ and the original alt-text caption $\mathcal{T}_i$ in the prompt. The resulting generated caption is denoted as $\mathcal{R}_i$.

DataConcept. Our multi-stage pipeline, fully summarised in Fig. 2, yields our final dataset. Each image-text sample in our dataset consists of concept metadata, including concept tags with confidence scores, localised bounding-boxes, and concept-aware synthetic captions. For ease of notation, we denote each sample i as $(\mathcal{I}_i, \mathcal{T}_i, \mathcal{R}_i, \mathcal{C}_i)$.

3. Concept-Aware Batch Sampling

3.1. Formulation

We formalize CABS as a parameterized sampling framework. Given superbatch $\mathcal{B}$ of size B drawn IID from the data-pool, we define a target batch size $b < B$ controlled by filter ratio $f \in [0, 1)$, such that $b = (1 - f)B$. For each sample with concept annotations $\mathcal{C}_i$, CABS computes a score $s_i = h(\mathcal{C}_i; \mathcal{B}, \theta_h)$, where $h(.)$ is a concept-aware heuristic gain function parameterized by θ_h(a set of parameters relevant to the sampling strategy), and selects sub-batch $\mathcal{B}_{\text{sub}} \subset \mathcal{B}$ of size b based on these scores. The target sub-batch is constructed as $\mathcal{B}_{\text{sub}} = \text{TopK}_{i \in \mathcal{B}}(s_i, k=b)$. For example, if the target is IID sampling, $h(i)$ would be set to 1 for all samples in $\mathcal{B}$ and $\theta_h = \varnothing$. Sampling the top-k in this way would be equivalent to IID sampling. By allowing $h(.)$ and θ_h to be flexible, practitioners can flexibly instantiate different batch sampling strategies and induce different concept distributions in $\mathcal{B}_{\text{sub}}$ *on-the-fly* during training. This flexibility is powerful as it enables *task-adaptive* batch curation. We provide PyTorch-style pseudocode for CABS in Alg. 1.

3.2. Task-Adaptivity of CABS

We now demonstrate two specific cases, *zero-shot classification* and *image-text retrieval*, where the flexibility of CABS enables modifying the concept distributions to be task-aware. Prior work [3] argues that classification and retrieval benefit from distinct curation strategies. However,

Algorithm 1 PyTorch-style code for CABS

```
# D=(I,T,C)=super-batch(images,texts,concepts)
# f=filter-ratio
# h=concept-aware heuristic gain function
# theta=parameter for heuristic gain function
def cabs(D, f, h):
  I, T, C = D # unpack super-batch
  B = I.size(0)
  b = (1-f)*B
  # Step1: compute heuristic scores
  scores = []
  for i in range(B):
    s_i = h(C[i], D, theta) # scoring
    scores.append(s_i)
  # Step2: select top-k samples by score
  selected_indices = topk(scores, k=b)
  # Step3: construct target batch
  I_target = I[selected_indices]
  T_target = T[selected_indices]
  return (I_target, T_target)
```

they only perform offline curation and do not disclose details of their methods. This motivates us to develop concrete instantiations of CABS for classification and retrieval.

Zero-shot classification assesses whether a model has learned discriminative features for different classes. Under IID sampling and concept-imbalanced training batches, common concepts are over-represented, resulting in under-optimization for rare concepts, and consequently, poor long-tailed performance [37, 94]. By constructing batches with more uniform distributions, a model would learn stronger representations for rare concepts, yielding improved generalization on long-tailed classification. In contrast, retrieval benchmarks test multi-object compositional understanding, requiring models to align rich textual descriptions to complex visual scenes (images with multiple concepts). By constructing batches enriched with similarly complex samples, each encompassing multiple concepts, models would generalize better to the compositional nature of retrieval. Given this, we develop two CABS algorithms (Tab. 1):

- **Diversity Maximization**: balance the concept distribution, focusing on uniform concept coverage (Sec. 4).
- **Frequency Maximization**: prioritize samples with the highest concept counts (Sec. 5).

Empirical Justification. To validate that classification and retrieval tasks exhibit substantially different concept distributions, we collect 4,096 random samples from MSCOCO (retrieval) and ImageNet (classification) and visualize their per-sample concept counts, following the same protocol used to construct DATACONCEPT. From Fig. 1 (left), we observe that ImageNet images tend to contain single objects, while MSCOCO naturally exhibits multi-object scenes. These characteristics are then approximated by the samples selected by our two CABS variants, further demonstrating the power and flexibility of task-adaptive batch curation. This also highlights the potential of analyzing salient task characteristics and shaping training distributions accordingly [55].

Table 1. **Parameters of CABS variants.** We indicate the heuristic function $h(\cdot)$ and parameters θ_h used for CABS-DM (see Sec. 4) and CABS-FM (see Sec. 5), and if the score is dependent on the current state of the sub-batch.

Method	$h(.)$	θ_h	Dependent?
IID	1	$\varnothing$	✗
CABS-DM	Eq. (1)	t_c	✓
CABS-FM	$\lvert\mathcal{C}_i\rvert$	$\varnothing$	✗

3.3. Experimental Setup

Models. We train a ViT-B-32 [22] CLIP using 224 image-resolution and ViT-B-16 SigLIP [91] at 256 resolution. We further test CABS by training ViT-S-16 CLIP and ViT-SO400M-14 SigLIP [4] models in Appx. D.2.

Data. We experiment with two variants of DATACONCEPT: one using noisy alt-texts ($\mathcal{T}_i$) and another with our *concept-aware re-captions* ($\mathcal{R}_i$). Note that IID sampling with alt-text captions corresponds to DataComp's default setup [31].

Evaluation Benchmarks. Following [79], we consider a diverse pool of 25 classification and 2 image-text retrieval benchmarks, spanning fine-grained, object-centric, and scene-centric categories. Additionally, to assess the effectiveness of our models in long-tailed settings, we evaluate on "Let-It-Wag!" test set from [78].

Training. We fix the training budget to be 128M samples seen; additional findings for higher budgets (1.28B samples seen) are described in Sec. 6. Note that we closely follow the hyperparameters set by DataComp for fair comparison, including a batch-size of 4096. The sample-level concepts $\mathcal{C}_i$ are used only for batch curation and do not contribute to the contrastive objective. Unless specified, we set the filter ratio to f=0.8, sampling from superbatches of size B=20,480. We show performance for other filter ratios in Appx. F.

Baselines. We compare CABS performance with two popular online batch sampling methods, GRIT-VLP [11] and MAFA [12]. Both GRIT-VLP and MAFA sample hard negatives based on embedding similarity. The key difference lies in how these similarities are computed: GRIT uses the *current* model's embeddings, while MAFA relies on those from a *pretrained* model. MAFA used BLIP for this purpose, but its smaller training budget makes comparisons unfair. To ensure parity, we instead pretrain CLIP and SigLIP on 128M samples and use their embeddings to compute MAFA similarities. Additionally, we note that, although JEST [24] and ACID [79] are also relevant baselines, they are proprietary algorithms with no public implementation.

4. CABS with Diversity Maximization

4.1. Formulation

As motivated in Sec. 3.1, zero-shot classification tasks benefit from balanced concept-level supervision across batches.

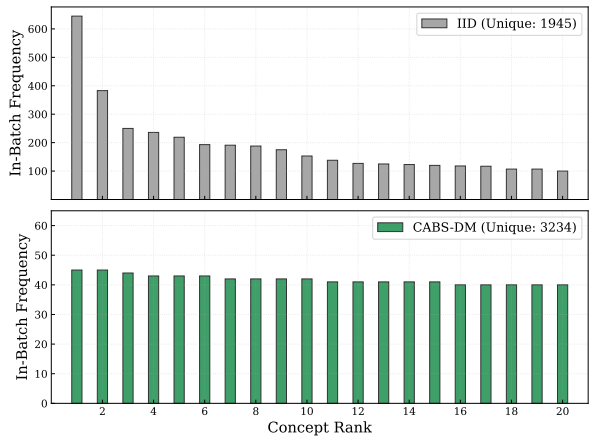

Figure 3. **Sub-batch compositions.** CABS-DM induces a near-uniform concept frequency distribution, de-biasing the distributional skew induced by IID-sampling. **Unique** indicates total unique concepts in the sub-batch—CABS-DM incorporates nearly double the concepts in the curated sub-batch, compared to IID.

Given the general formulation detailed previously, we instantiate CABS with diversity maximization (CABS-DM) and its corresponding heuristic function h_{DM}, which scores samples iteratively such that the top-k-filtered batch approximates a uniform concept distribution. For a superbatch $\mathcal{B}$, CABS-DM assigns higher scores to samples containing under-represented concepts in $\mathcal{B}_{\text{sub}}$ and selects the top $b = (1-f)B$ samples until the frequency of each concept reaches an upper bound t_c, a tunable hyperparameter.

CABS-DM constructs a sub-batch by iteratively selecting samples that maximize a gain function $h_{DM}(i)$ and updating the sub-batch concept count n_c for all concepts $c \in \mathcal{C}_i$. This process continues until the desired batch size for training is obtained, which is vastly different from an IID-sampled batch, as illustrated in Fig. 3. An average CABS-DM sub-batch contains $1.5\times$ more concepts than an IID-sampled batch, in addition to exhibiting a mostly flat concept distribution. This helps increase diversity at the batch level. CABS-DM includes the following components:

Pooling Concepts and Target Count. For each $\mathcal{B}$, we first compute the global frequency $\mathcal{F}_c$ of each concept. We next fix the target count t_c for concept c, *i.e.* the maximum number of times c should appear in the sub-batch, to enforce approximate uniformity (following our prior notation, θ_h consists of t_c in this case). In a simplified setting, if each sample comprises 1 concept, $\sum_c t_c \approx b = (1-f)B$.

Gain Function. The sample-level gain function is based on the current state of the sub-batch's concept distribution. Given sample i with concept set $\mathcal{C}_i$, we define the gain as

$$h_{\text{DM}}(i) = \frac{1}{|\mathcal{C}_i|} \sum_{c \in \mathcal{C}_i} \begin{cases} \dfrac{t_c - n_c}{t_c} + \dfrac{1}{\mathcal{F}_c}, & \text{if } n_c < t_c, \\ 0, & \text{if } n_c \geq t_c. \end{cases} \quad (1)$$

Table 2. **CABS-DM improves over IID.** Our method substantially outperforms IID sampling across settings. Importantly, gains from CABS-DM extend to the long-tailed "Let-It-Wag!" test set too.

Method	Cap.	Zero-shot Classification				Let-it-Wag!	Avg (Clf)
		IN-Val	IN-shift	Obj	Scene		
ViT-B-32-CLIP							
IID [31]	alt	17.3	15.2	32.3	36.4	5.1	28.2
CABS-DM	alt	**21.9**	**18.6**	**34.5**	**38.0**	**7.5**	**30.7**
IID [31]	recap	21.7	20.8	36.4	**43.1**	5.9	33.0
CABS-DM	recap	**26.7**	**25.4**	**39.6**	42.8	**7.1**	**35.5**
ViT-B-16-SigLIP-256							
IID [31]	alt	17.2	15.3	29.6	35.9	5.2	26.4
CABS-DM	alt	**24.1**	**20.8**	**33.5**	**39.6**	**7.0**	**30.9**
IID [31]	recap	28.8	27.4	41.5	48.9	6.6	38.6
CABS-DM	recap	**34.7**	**32.3**	**43.2**	**50.6**	**7.6**	**41.1**

Each concept contributes two components: a `balance gain` $((t_c - n_c)/t_c)$ that prioritises under-represented concepts and a `rarity bonus` $(1/\mathcal{F}_c)$ that upweights long-tailed concepts. The rarity bonus ensures rare concepts are incorporated into the sub-batch earlier during greedy selection thereby reducing the total number of samples that must be parsed. At each step, we sort all remaining super-batch samples by this score, deterministically select sample $i^\star = \arg\max_i h_{DM}(i)$, append $i^\star$ to the sub-batch, and update $n_c \leftarrow n_c + 1$ for all $c \in \mathcal{C}_{i^\star}$. If concept c exceeds t_c, all remaining samples containing c are rendered invalid. Scores s_i are then generated for all relevant samples remaining in $\mathcal{B}$ using h_{DM} and the sample with the highest score is incorporated into the sub-batch for the next iteration.

Sample Selection. CABS-DM *proceeds through a sequence of greedy maximizations to yield a balanced and diverse sub-batch*. At every iteration, it deterministically selects the sample with the highest gain, conditioned on the current sub-batch composition, without randomness, akin to an Expectation-Maximization alternating optimization between sample selection and score update. Benefits of h_{DM} include (i) reproducibility across runs for the same super-batch due to deterministic selection, and (ii) gain terms jointly enforce uniform concept coverage and higher batch diversity. Importantly, while h_{DM} is deterministic given a fixed super-batch, sample selection varies across training steps due to stochastic super-batch draws. This variability is desirable, as it distinguishes CABS-DM from offline curation and promotes diversity across training. We provide PyTorch-style pseudocode for CABS-DM in Appx. C.1.

4.2. Improvements on Zero-shot Classification

We now comprehensively evaluate the effectiveness of CABS-DM against standard IID sampling for multimodal pretraining. As shown in Tab. 2, CABS-DM consistently delivers improvements across four different test settings. On ImageNet, CABS-DM yields substantial gains over IID sampling, with an absolute improvement of $+5.0\%$ for CLIP ViT-

Table 3. **CABS-DM beats MetaCLIP-style curation**. Despite having similar curation objectives, we show our online concept-balanced batch sampling significantly outperforms offline curation.

Method	Zero-shot Classification				Let-it-Wag!	Avg (Clf)
	IN-Val	IN-shift	Obj	Scene		
ViT-B-32-CLIP						
IID [31]	17.3	15.2	32.3	36.4	5.1	28.2
MetaCLIP [87]	18.2	16.9	30.3	32.9	5.3	26.9
CABS-DM	**21.9**	**18.6**	**34.5**	**38.0**	**7.5**	**30.7**
ViT-B-16-SigLIP-256						
IID [31]	17.2	15.3	29.6	35.9	5.2	26.4
MetaCLIP [87]	20.3	18.9	30.7	35.3	5.3	28.0
CABS-DM	**24.1**	**20.8**	**33.5**	**39.6**	**7.0**	**30.9**

B-32 and +6.9% for SigLIP B-16-256. Similar trends are observed across the broader suite of benchmarks and model variants (refer to Appx.D.2), where CABS-DM boosts average accuracy. Beyond standard benchmarks, CABS-DM also enhances long-tailed recognition on Let-It-Wag! [78], with boosts of $1.0 - 2.4\%$. This demonstrates CABS-DM's ability to improve both general and long-tailed performance. We also observe strong gains on autoregressive VQA and image captioning results (Appx. H).

Notably, we also observe consistent improvements from using our concept-aware re-captions compared to alt-texts, even with standard IID sampling. With CLIP-ViT-B/32, our re-captions lead to a +4.3% boost on ImageNet, and +4.8% for zero-shot classification. For SigLIP-ViT-B/16, the accuracy gains are as large as +11.6% and +12.2%. These results quantify the benefits of both DATACONCEPT and CABS, showcasing that *concept-aware recaptions and task-aware online curation provide the strongest gains*.

4.3. Improvements over State-of-the-art Methods

MetaCLIP. We compare CABS-DM with MetaCLIP [87], an offline curation method that aims at concept-balanced curation by first collecting 500,000 queries from WordNet synsets and Wikipedia titles, followed by matching these queries to a pool of image–text pairs via substring search in alt-texts, capping each query at 20,000 samples. To provide a fair baseline, we re-implement MetaCLIP curation based on image content, using our concept vocabulary $\mathcal{V}$ as the query pool and approximating the concept threshold based on the desired curated dataset size. To align with CABS-DM at $f = 0.8$ (where the full dataset is repeated $5\times$ to match our 128M samples-seen regime), we construct a 25.6M MetaCLIP-subset and train with $5\times$ repeats. This is achieved by using a per-concept threshold of 70,000.

Tab. 3 shows comparisons with CABS-DM, MetaCLIP and IID sampling, with CABS-DM substantially outperforming MetaCLIP in zero-shot classification (+3.8%/ +2.9% gains on ImageNet and the average classification set respectively) as well as long-tailed evaluations, highlighting the performance boosts achieved with online batch curation.

Table 4. **CABS-DM outperforms SOTA open-source batch sampling methods.** With both CLIP-ViT-B/32 and SigLIP-ViT-B/16, CABS-DM provides significant benefits to LIP compared to GRIT-VLP and MAFA, making it more suitable for modern LIP.

Method	Zero-shot Classification				Let-it-Wag!	Avg (Clf)
	IN-Val	IN-shift	Obj	Scene		
ViT-B-32-CLIP						
IID	17.3	15.2	32.3	36.4	5.1	28.2
GRIT-VLP [11]	17.6	15.0	31.7	35.6	6.3	27.5
MAFA [12]	17.0	15.0	32.2	35.9	5.6	27.9
CABS-DM	**21.9**	**18.6**	**34.5**	**38.0**	**7.5**	**30.7**
ViT-B-16-SigLIP-256						
IID	17.2	15.3	29.6	35.9	5.2	26.4
GRIT-VLP [11]	17.3	15.1	30.7	37.3	5.0	27.2
MAFA [12]	17.2	15.2	30.7	36.2	5.1	27.1
CABS-DM	**24.1**	**20.8**	**33.5**	**39.6**	**7.0**	**30.9**

Online Batch Sampling. After demonstrating benefits of online sampling compared to offline curation, we next compare CABS-DM to other online approaches such as GRIT-VLP [11] and MAFA [12]. Tab. 4 highlights that both methods lag behind CABS-DM. We note GRIT and MAFA also struggle to outperform the IID baseline (with CLIP), but offer modest improvements with SigLIP. These observations are in line with recent works suggesting that SigLIP models benefit more from active batch sampling [24, 79]. With SigLIP-ViT-B/16, CABS-DM improvements are up to +6.8% on ImageNet and +3.7% on average.

5. CABS with Frequency Maximization

5.1. Formulation

As described previously in Sec. 3.1, we next focus on retrieval. Retrieval benchmarks like MSCOCO and Flickr often consist of images with multiple objects and complex scenes (Fig. 1), necessitating changes to the design of scoring function compared to CABS-DM. This leads us to instantiate CABS with frequency maximization (CABS-FM) and its corresponding heuristic function h_{FM}, which scores samples based on concept count. As a result, filtered sub-batches contain samples from super-batch $\mathcal{B}$ with maximal object multiplicity, exhibiting higher scene complexity overall.

Gain Function. We define a simple sample-level gain function $h_{FM}(i) = |\mathcal{C}_i|$. CABS-FM scores every $i \in \mathcal{B}$ by $h_{FM}(i)$, sorts samples by this value, constructs a top-k sub-batch $\mathcal{B}_{\text{sub}} = \text{TopK}_{i \in \mathcal{B}}(|\mathcal{C}_i|, k = b)$ (PyTorch-style pseudocode can be found in the Appx. C.2). h_{FM} thus provides the model with the most concept-dense sub-batch.

5.2. Experiment Results

Improvements on Image-Text Retrieval. Following our previous CABS-DM evaluation protocol, we test CABS-FM across the full model suite using alt-text and concept-aware re-captions. As shown in Tab. 5, CABS-FM consistently outperforms IID sampling across all configurations, yielding

Table 5. **CABS-FM improves over IID.** Performance on both Flickr and MSCOCO significantly improved, demonstrating that concept multiplicity curation indeed benefits retrieval.

Method	Captions	COCO	Flickr	Avg(Ret)
ViT-B-32-CLIP				
IID	alt	9.7	16.2	12.9
CABS-FM	alt	**11.0**	**21.9**	**16.4**
IID	recap	24.0	41.3	32.6
CABS-FM	recap	**30.4**	**52.9**	**41.6**
ViT-B-16-SigLIP-256				
IID	alt	11.1	18.9	15.0
CABS-FM	alt	**12.3**	**23.9**	**18.1**
IID	recap	37.1	57.0	47.0
CABS-FM	recap	**39.7**	**63.5**	**51.6**

Table 6. **CABS-FM outperforms state-of-the-art.** Arriving at the same conclusions as CABS-DM, we show significant benefits in using CABS compared to other online batch sampling methods.

Method	COCO	Flickr	Avg(Ret)
ViT-B-32-CLIP			
IID	9.7	16.2	12.9
GRIT-VLP [11]	9.6	15.6	12.6
MAFA [12]	9.6	15.5	12.5
CABS-FM	**11.0**	**21.9**	**16.5**
ViT-B-16-SigLIP-256			
IID	11.1	18.9	15.0
GRIT-VLP [11]	11.6	19.6	15.6
MAFA [12]	10.5	19.4	14.9
CABS-FM	**12.3**	**23.9**	**18.1**

gains of +3.5% and +3.1% for ViT-B-32-CLIP and ViT-B-16-SigLIP-256 (alt-text), averaged over MSCOCO and Flickr. These improvements further widen to +9.0% and +4.6% when training on the re-captions.

Online Batch Sampling Methods. In Tab. 6, we find that CABS-FM outperforms GRIT-VLP and MAFA. Similar to the classification case, both baselines fail to surpass IID sampling for ViT-B-32-CLIP and offer only modest improvements for ViT-B-16-SigLIP-256. In contrast, CABS-FM offers large boosts, improving over GRIT-VLP by +3.9% (ViT-B-32-CLIP) and +2.5% (ViT-B-16-SigLIP-256).

6. Data- / Compute-Constrained Experiments

Having explored the efficacy of our CABS variants across both classification and image-text retrieval tasks, in this section, we now study the benefits of CABS along another axis: *data-* vs *compute-constrained* pretraining.

Definition. Let C denote the target compute (FLOPs), $\mathcal{D}$ the pretraining dataset, and $C_{\mathcal{D}}$ the required compute for one epoch over $\mathcal{D}$. If $C \leq C_{\mathcal{D}}$, then training is compute-constrained, *i.e.*, the compute budget is insufficient to consume all the data. If $C > C_{\mathcal{D}}$, then training is data-constrained, *i.e.*, samples must be repeated.

Experimental Design. Due to the sampling mechanism of CABS, going from a larger superbatch to a training subbatch, all the experiments in Secs. 4.2, 4.3 and 5.2 operate under a data-constrained setting for both CABS variants. This occurs since a fraction f=0.8 of samples are filtered out online during training, making the *effective* samples-seen-per-epoch for CABS 5× less than IID, which instead operates with C=$C_{\mathcal{D}}$. Following common practices in pretraining, we increase the constraints further with two experiments: ① *less data, but higher quality*, where we keep the 128M sample budget, but filter DataConcept via CLIP Score [71]. We keep the top 30% samples as in [31], thereby reducing the starting dataset to ∼38M samples[3]. To prevent high repeat rates, we set f=0.5, yielding 6.67× worst-case repeats for CABS, which are comparable to the 5× worst-case repeats induced by f=0.8 in Sec. 4. Note that IID sampling yields 3.33 worst-case repeats after CLIP-Score filtering. ② *long training*, where we do not filter, but rather increase the training budget to 1B samples seen, matching the *large* scale of DataComp [31].

Table 7. **CABS-DM is compatible with CLIPScore filtering.** Although CABS-DM leads to more repeats, which yield diminishing returns on already curated data [35], we generally improve over IID even with 2× more repeats across model architectures.

Method	Zero-shot Classification				Let-it-Wag!	Avg (Clf)
	IN-Val	IN-shift	Obj	Scene		
ViT-B-32-CLIP						
IID	27.3	23.0	39.8	43.1	10.7	35.7
CABS-DM	**30.1**	**25.6**	**41.8**	**44.8**	**12.7**	**37.8**
ViT-B-16-SigLIP-256						
IID	34.7	29.5	46.2	**48.9**	11.9	42.0
CABS-DM	**37.5**	**32.2**	46.2	48.5	**12.6**	**42.7**

Table 8. **CABS-FM is also compatible with CLIPScore filtering.** Despite the same repeat protocol as CABS-DM, we show unanimous performance gains across all benchmarks and models tested.

Method	COCO	Flickr	Avg(Ret)
ViT-B-32-CLIP			
IID	13.8	24.1	18.9
CABS-FM	**15.9**	**26.5**	**21.2**
ViT-B-16-SigLIP-256			
IID	18.7	34.7	26.7
CABS-FM	**20.1**	**36.3**	**28.2**

Less data, but higher quality. In this regime, both CABS variants remain effective even with CLIP-score-filtered data (see Tab. 7 for CABS-DM and Tab. 8 for CABS-FM). Notably, while repeating curated data has been shown to yield diminishing returns [35], CABS still trumps IID sampling despite using a 2× more data repeat rate.

Long Training. Next, we study the training dynamics under the regime where we train both IID and CABS variants with a CLIP ViT-B/32 backbone for 1.28B samples seen. As

[3]We use OpenAI's CLIP ViT-L/14 model for scoring cosine similarities.

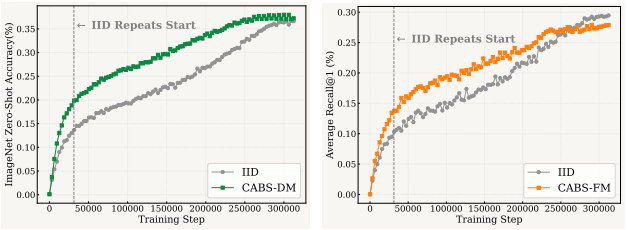

Figure 4. **CABS with longer training (1.28B samples seen).** Both CABS-DM and CABS-FM show significant boost over IID for ViT-B-32-CLIP in both compute-constrained and data-constrained regimes, the grey dashed line being the point where compute-constraint shift to data-constraint in an IID sampling regime.

illustrated in Fig. 4, we find that as long as IID training is compute-constrained (dashed gray line), CABS significantly outperforms the vanilla IID recipe, displaying impressive $3.2\times$ and $2\times$ compute multipliers. The performance gains only slightly diminish when training is far into the data-constrained regime, with CABS yielding a worst-case of 50 repeats and IID yielding only 10. However, the overall performance is still quite strong compared to the IID baseline. These experiments confirm that our CABS method is fully compatible with ① *smaller, highly curated datasets, and* ② *pre-training on web-crawled corpora for multiple epochs*.

7. Related Work

Sampling Approaches for Training Multimodal Models. Training web-scale foundation models typically uses uniform, IID mini-batch sampling, which assigns equal weights to each sample in the training set. However, in multimodal corpora, examples differ drastically in quality [31, 71, 86], are possibly redundant [1, 2, 23, 74, 83], and exhibit skewed, long-tailed distributions across concepts [63, 78]. Moreover, for contrastive objectives like CLIP [66], batch composition heavily shapes the learning process. In this context, uniform sampling is not neutral: it can overexpose trivial or spurious correlations and under-represent rare but informative cases. Hence, several recent approaches try to apply better batch sampling schemes to ensure more effective cross-modal learning. Early works like RHO-Loss [54] and Bad-Students [25] move away from IID sampling, but select data samples independently without considering the overall batch composition. This issue is addressed by methods such as GRIT-VLP [11], MAFA [12], JEST [24], B3 [75], Falcon [43] and ACID [79]. Our paper builds on this line of work by incorporating concept diversity into the training batch construction, an aspect missing from previous works.
Analyzing Concepts in Multimodal Datasets. Understanding the composition of multimodal datasets is important for building better batch sampling methods. Early image-text datasets like CC-3M [72], CC-12M [13] and YFCC-100M [76] partially characterize their inherent concept distributions using metadata from the web sources where images are scraped from. The WebLI [15] dataset (used for training models like PaliGemma [8] and SigLIP [91]) was annotated using OCR models to detect objects in images. However, due to the scale of compute required for annotating recent open datasets like LAION-5B [71] and DataComp-1B [31], very few works have studied their concept distribution. Udandarao et al. [78] tag each sample in LAION-400M with its constituent concepts by using a pretrained image-tagging model [41] and text search. Other works have proposed improving concept coverage in various ways, e.g. considering multilingual data [59] or recaptioning [89]. Our DataConcept also augments samples with fine-grained concept annotations and is designed specifically to enable explicit control over online, concept-based batch construction.

8. Conclusion

We investigate the role of incorporating concept-level information during vision-language pretraining, which is relatively underexplored by prior data-centric work. To this end, we introduce DataConcept, a large-scale, fully annotated pretraining dataset designed to expose concept-level annotations, and CABS, a flexible framework leveraging this information to perform online, concept-aware batch sampling during pretraining. Our extensive evaluations demonstrate the benefits of CABS over IID and other curation strategies (including existing batch sampling algorithms) across both classification and retrieval tasks, highlighting its versatility. By making DataConcept and CABS publicly available, we hope to motivate future work to incorporate concept-awareness into their data pipelines for building better VLMs.
Limitations. One disadvantage of CABS is the cost of concept annotations. However, this cost is amortizable as the annotated data can be re-used for training different models to do well on different tasks. It is also worth noting that the runtime of CABS increases as we increase the filtering ratio f from the superbatch. Our experiments show that CABS can still offer performance benefits at low filtering ratios, where the runtime overhead is more manageable. Besides, we have not experimented with more complex multimodal architectures or large-scale training runs that mirror current state-of-the-art training setups.
Future Work. Our proposed framework motivates several directions to study concept-centric data curation further. One avenue could be applying CABS to fine-tuning data. In addition, future work could look into other score functions that will work well with a wide range of tasks, balancing both retrieval and classification performance. This balance could potentially be achieved through curriculum learning as well. In our experiments, we pick a score function at the start and apply it to all samples across all superbatches. One could study how to best update the score function throughout the course of training, e.g. by first prioritizing single-object images and then moving on to selecting complex scenes.

Acknowledgements

The authors thank (in alphabetical order): Sebastian Dziadzio, Andreas Hochlehnert, Benno Krojer, Hilde Kuehne, Ameya Prabhu, Thaddäus Wiedemer, Konstantin Wilkin and Jiajun Zhang for helpful feedback at different stages of this project. AG gratefully acknowledges LAION and the Gauss Centre for Supercomputing e.V. for funding this work by providing computing time on the JUWELS Booster at Jülich Supercomputing Centre (JSC). AG receives funding from the European Union's Horizon Europe research and innovation program under ELLIOT - Grant Agreement No 101214398. AG and VU thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for support. VU was supported by a Google PhD fellowship in Machine Intelligence. MF acknowledges travel support from ELIAS (GA no 101120237). MB acknowledge financial support by the Federal Ministry of Education and Research (BMBF), FKZ: 011524085B and Open Philanthropy Foundation funded by the Good Ventures Foundation.

References

[1] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023. 8

[2] Amro Abbas, Evgenia Rusak, Kushal Tirumala, Wieland Brendel, Kamalika Chaudhuri, and Ari S Morcos. Effective pruning of web-scale datasets based on complexity of concept clusters. *arXiv preprint arXiv:2401.04578*, 2024. 8

[3] Amro Abbas, Josh Wills, Haoli Yin, Paul Burstein, Ning Cao, Aldo Carranza, Alvin Deng, Priya Goyal, Pratyush Maini, Joshua McGrath, Fan Pan, Jack Urbanek, Vineeth Kada, Muhammed Razzak, Vishwa Shah, Vishruth Veerendranath, Bogdan Gaza, Ari Morcos, and Matthew Leavitt. DatologyAI Technical Deep-Dive: Image-Text Data Curation at the Billion-Sample Scale. Technical report, DatologyAI, 2024. 1, 3, 36

[4] Ibrahim M Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. Getting vit in shape: Scaling laws for compute-optimal model design. *Advances in Neural Information Processing Systems*, 36:16406–16425, 2023. 4

[5] Anas Awadalla, Le Xue, Manli Shu, An Yan, Jun Wang, Senthil Purushwalkam, Sheng Shen, Hannah Lee, Oscar Lo, Jae Sung Park, et al. Blip3-kale: Knowledge augmented large-scale dense captions. *arXiv preprint arXiv:2411.07461*, 2024. 27

[6] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019. 36

[7] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 39

[8] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 8

[9] Cody Blakeney, Mansheej Paul, Brett W Larsen, Sean Owen, and Jonathan Frankle. Does your data spark joy? performance gains from domain upsampling at the end of training. *arXiv preprint arXiv:2406.03476*, 2024. 39

[10] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014. 36

[11] Jaeseok Byun, Taebaek Hwang, Jianlong Fu, and Taesup Moon. Grit-vlp: Grouped mini-batch sampling for efficient vision and language pre-training. In *European Conference on Computer Vision*, pages 395–412. Springer, 2022. 4, 6, 7, 8

[12] Jaeseok Byun, Dohoon Kim, and Taesup Moon. Mafa: Managing false negatives for vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27314–27324, 2024. 4, 6, 7, 8

[13] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021. 8

[14] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 36

[15] Xi Chen, Xiao Wang, Soravit Changpinyo, Anthony J Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 8

[16] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 36

[17] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018. 36

[18] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 36

[19] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011. 36

[20] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and

pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 91–104, 2025. 27

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 36

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[23] Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. What's in my big data? *arXiv preprint arXiv:2310.20707*, 2023. 8

[24] Talfan Evans, Nikhil Parthasarathy, Hamza Merzic, and Olivier Henaff. Data curation via joint example selection further accelerates multimodal learning. *Advances in Neural Information Processing Systems*, 37:141240–141260, 2024. 4, 6, 8

[25] Talfan Evans, Shreya Pathak, Hamza Merzic, Jonathan Schwarz, Ryutaro Tanno, and Olivier J Henaff. Bad students make great teachers: Active learning accelerates large-scale visual understanding. In *European Conference on Computer Vision*, pages 264–280. Springer, 2024. 8

[26] Mark Everingham. The pascal visual object classes challenge 2007. In *http://www. pascal-network. org/challenges/VOC/voc2007/workshop/index. html*, 2009. 36

[27] Fartash Faghri, Pavan Kumar Anasosalu Vasu, Cem Koc, Vaishaal Shankar, Alexander Toshev, Oncel Tuzel, and Hadi Pouransari. Mobileclip2: Improving multi-modal reinforced training. *arXiv preprint arXiv:2508.20691*, 2025. 3

[28] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36: 35544–35575, 2023. 3

[29] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 36

[30] Steven Feng, Shrimai Prabhumoye, Kezhi Kong, Dan Su, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Maximize your data's potential: Enhancing llm accuracy with two-phase pretraining. *arXiv preprint arXiv:2412.15285*, 2024. 39

[31] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 1, 2, 3, 4, 5, 6, 7, 8, 35, 36

[32] William Gaviria Rojas, Sudnya Diamos, Keertan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. *Advances in Neural Information Processing Systems*, 35:12979–12990, 2022. 36

[33] Adhiraj Ghosh, Sebastian Dziadzio, Ameya Prabhu, Vishaal Udandarao, Samuel Albanie, and Matthias Bethge. Onebench to test them all: Sample-level benchmarking over open-ended capabilities. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32445–32481, 2025. 22, 43

[34] Leander Girrbach, Stephan Alaniz, Genevieve Smith, Trevor Darrell, and Zeynep Akata. Person-centric annotations of laion-400m: Auditing bias and its transfer to models. *arXiv preprint arXiv:2510.03721*, 2025. 1

[35] Sachin Goyal, Pratyush Maini, Zachary C Lipton, Aditi Raghunathan, and J Zico Kolter. Scaling laws for data filtering–data curation cannot be compute agnostic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22702–22711, 2024. 1, 7

[36] Suchin Gururangan, Dallas Card, Sarah Dreier, Emily Gade, Leroy Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A Smith. Whose language counts as high quality? measuring language ideologies in text data selection. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 2562–2580, 2022. 1

[37] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009. 4

[38] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021. 36

[39] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021. 36

[40] Rachel Hong, William Agnew, Tadayoshi Kohno, and Jamie Morgenstern. Who's in and who's out? a case study of multimodal clip-filtering in datacomp. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–17, 2024. 1

[41] Xinyu Huang, Yi-Jie Huang, Youcai Zhang, Weiwei Tian, Rui Feng, Yuejie Zhang, Yanchun Xie, Yaqian Li, and Lei Zhang. Open-set image tagging with multi-grained text supervision. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 4117–4126, 2025. 2, 8, 15, 16

[42] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 35

[43] Myunsoo Kim, Seong-Woong Shim, and Byung-Jun Lee. Falcon: False-negative aware learning of contrastive negatives in vision-language pretraining. *arXiv preprint arXiv:2505.11192*, 2025. 8

[44] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani,

Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021. 36

[45] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 36

[46] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 36

[47] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020. 2, 15

[48] Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, et al. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Advances in Neural Information Processing Systems*, 35:9287–9301, 2022. 22

[49] Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, et al. What if we recaption billions of web images with llama-3? *arXiv preprint arXiv:2406.08478*, 2024. 1, 18

[50] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 42

[51] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 3, 18

[52] Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276, 2024. 1

[53] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 36

[54] Sören Mindermann, Jan M Brauner, Muhammed T Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, et al. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *International Conference on Machine Learning*, pages 15630–15649. PMLR, 2022. 8

[55] David Mizrahi, Anders Boesen Lindbo Larsen, Jesse Allardice, Suzie Petryk, Yuri Gorokhov, Jeffrey Li, Alex Fang, Josh Gardner, Tom Gunter, and Afshin Dehghan. Language models improve when pretraining data matches target tasks. *arXiv preprint arXiv:2507.12466*, 2025. 1, 4

[56] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *18th international conference on pattern recognition (ICPR'06)*, pages 850–855. IEEE, 2006. 21

[57] Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of clip. *Advances in Neural Information Processing Systems*, 35:21455–21469, 2022. 1

[58] Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. Improving multimodal datasets with image captioning. *Advances in neural information processing systems*, 36:22047–22069, 2023. 1, 3, 18, 28

[59] Thao Nguyen, Matthew Wallingford, Sebastin Santy, Wei-Chiu Ma, Sewoong Oh, Ludwig Schmidt, Pang Wei W Koh, and Ranjay Krishna. Multilingual diversity improves vision-language representations. *Advances in Neural Information Processing Systems*, 37:91430–91459, 2024. 1, 8

[60] Thao Nguyen, Yang Li, Olga Golovneva, Luke Zettlemoyer, Sewoong Oh, Ludwig Schmidt, and Xian Li. Recycling the web: A method to enhance pre-training data quality and quantity for language models. *arXiv preprint arXiv:2506.04689*, 2025. 1

[61] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 36

[62] Chengcheng Ning, Huajun Zhou, Yan Song, and Jinhui Tang. Inception single shot multibox detector for object detection. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 549–554. IEEE, 2017. 21

[63] Shubham Parashar, Zhiqiu Lin, Tian Liu, Xiangjue Dong, Yanan Li, Deva Ramanan, James Caverlee, and Shu Kong. The neglected tails in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12988–12997, 2024. 8, 24

[64] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 36

[65] Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Steiner, Xiaohua Zhai, and Ibrahim M Alabdulmohsin. No filter: Cultural and socioeconomic diversity in contrastive vision-language models. *Advances in Neural Information Processing Systems*, 37:106474–106496, 2024. 1

[66] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 8, 36

[67] Vikram V Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. Geode: a geographically diverse evaluation

dataset for object recognition. *Advances in Neural Information Processing Systems*, 36:66127–66137, 2023. 36

[68] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 36

[69] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019. 15

[70] Karsten Roth, Vishaal Udandarao, Sebastian Dziadzio, Ameya Prabhu, Mehdi Cherti, Oriol Vinyals, Olivier Hénaff, Samuel Albanie, Matthias Bethge, and Zeynep Akata. A practitioner's guide to continual multimodal pretraining. *arXiv preprint arXiv:2408.14471*, 2024. 39

[71] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 1, 7, 8

[72] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018. 8

[73] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107: 104117, 2021. 3, 18, 20, 21

[74] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022. 8

[75] Raghuveer Thirukovalluru, Rui Meng, Ye Liu, Mingyi Su, Ping Nie, Semih Yavuz, Yingbo Zhou, Wenhu Chen, Bhuwan Dhingra, et al. Breaking the batch barrier (b3) of contrastive learning via smart batch mining. *arXiv preprint arXiv:2505.11293*, 2025. 8

[76] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 8, 36

[77] Lukas Tuggener, Raphael Emberger, Adhiraj Ghosh, Pascal Sager, Yvan Putra Satyawan, Javier Montoya, Simon Goldschagg, Florian Seibold, Urs Gut, Philipp Ackermann, et al. Real world music object recognition. *Transactions of the International Society for Music Information Retrieval*, 7(1): 1–14, 2024. 18

[78] Vishaal Udandarao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. No" zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1, 2, 3, 4, 6, 8, 15, 16, 17, 24, 26, 36

[79] Vishaal Udandarao, Nikhil Parthasarathy, Muhammad Ferjad Naeem, Talfan Evans, Samuel Albanie, Federico Tombari, Yongqin Xian, Alessio Tonioni, and Olivier J Hénaff. Active data curation effectively distills large-scale multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14422–14437, 2025. 4, 6, 8, 36

[80] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems*, 32, 2019. 36

[81] Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. V3det: Vast vocabulary visual detection dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19844–19854, 2023. 2, 15

[82] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3, 27, 28

[83] Ryan Webster, Julien Rabin, Loic Simon, and Frederic Jurie. On the de-duplication of laion-2b. *arXiv preprint arXiv:2303.12733*, 2023. 8

[84] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6609–6618, 2019. 15

[85] Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, 119(1):3–22, 2016. 36

[86] Hu Xu, Saining Xie, Po-Yao Huang, Licheng Yu, Russell Howes, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Cit: Curation in training for effective vision-language data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15180–15189, 2023. 8

[87] Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 6, 48

[88] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014. 36

[89] Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu. Capsfusion: Rethinking image-text data at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14022–14032, 2024. 8

[90] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga,

Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. The visual task adaptation benchmark. 2019. 36

[91] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 4, 8

[92] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1724–1732, 2024. 15, 16

[93] Yang Zhang, Amr Mohamed, Hadi Abdine, Guokan Shang, and Michalis Vazirgiannis. Beyond random sampling: Efficient language model pretraining via curriculum learning. *arXiv preprint arXiv:2506.11300*, 2025. 39

[94] Peilin Zhao and Tong Zhang. Accelerating minibatch stochastic gradient descent using stratified sampling. *arXiv preprint arXiv:1405.3080*, 2014. 4

[95] Xiangyu Zhao, Yicheng Chen, Shilin Xu, Xiangtai Li, Xinjiang Wang, Yining Li, and Haian Huang. An open and comprehensive pipeline for unified object grounding and detection. *arXiv preprint arXiv:2401.02361*, 2024. 22