No "Zero-Shot" Without Exponential Data: Pretraining Concept Frequency Determines Multimodal Model Performance

Vishaal Udandarao[1,2*] Ameya Prabhu[1,3*] Adhiraj Ghosh[1] Yash Sharma[1]
Philip H.S. Torr[3] Adel Bibi[3] Samuel Albanie[2†] Matthias Bethge[1†]

[1]Tübingen AI Center, University of Tübingen [2]University of Cambridge
[3]University of Oxford

github.com/bethgelab/frequency_determines_performance
huggingface.co/datasets/bethgelab/let-it-wag

Abstract

Web-crawled pretraining datasets underlie the impressive "zero-shot" evaluation performance of multimodal models, such as CLIP for classification/retrieval and Stable-Diffusion for image generation. However, it is unclear how meaningful the notion of "zero-shot" *generalization* is for such multimodal models, as it is not known to what extent their pretraining datasets encompass the downstream concepts targeted for during "zero-shot" evaluation. In this work, we ask: *How is the performance of multimodal models on downstream concepts influenced by the frequency of these concepts in their pretraining datasets?*

We comprehensively investigate this question across 34 models and five standard pretraining datasets (CC-3M, CC-12M, YFCC-15M, LAION-400M, LAION-Aesthetics), generating over 300GB of data artifacts. We consistently find that, far from exhibiting "zero-shot" generalization, multimodal models require exponentially more data to achieve linear improvements in downstream "zero-shot" performance, following a sample inefficient log-linear scaling trend. This trend persists even when controlling for sample-level similarity between pretraining and downstream datasets [79], and testing on purely synthetic data distributions [51]. Furthermore, upon benchmarking models on long-tailed data sampled based on our analysis, we demonstrate that multimodal models across the board perform poorly. We contribute this long-tail test set as the *Let it Wag!* benchmark to further research in this direction. Taken together, our study reveals an exponential need for training data which implies that the key to "zero-shot" generalization capabilities under large-scale training paradigms remains to be found.

1 Introduction

Multimodal models like CLIP [91] and Stable Diffusion [96] have revolutionized performance on downstream tasks—CLIP is now the de-facto standard for "zero-shot" image recognition [133, 72, 126, 48, 132] and image-text retrieval [46, 64, 24, 117, 129], while Stable Diffusion is now the de-facto standard for "zero-shot" text-to-image (T2I) generation [93, 17, 96, 41]. In this work, we investigate this empirical success through the lens of zero-shot generalization [69], which refers to the ability of the model to apply its learned knowledge to new unseen concepts. Accordingly, we ask: *Are current multimodal models truly capable of "zero-shot" generalization?*

To address this, we conducted a comparative analysis involving two main factors: (1) the performance of models across various downstream tasks and (2) the frequency of test concepts within their pretraining datasets. We compiled a comprehensive list of $4,029$ concepts[1] from 27 downstream tasks spanning classification, retrieval, and image generation, assessing the performance against these concepts. Our analysis spanned

*equal contribution and †equal advising, order decided by a coin flip

[1]class categories for classification tasks, objects in the text captions for retrieval tasks, and objects in the text prompts for generation tasks, see Sec. 2 for more details on how we define concepts.

five large-scale pretraining datasets with different scales, data curation methods and sources (CC-3M [107], CC-12M [27], YFCC-15M [113], LAION-Aesthetics [103], LAION-400M [102]), and evaluated the performance of 10 CLIP models and 24 T2I models, spanning different architectures and parameter scales. We consistently find across all our experiments that, across concepts, the frequency of a concept in the pretraining dataset is *a strong predictor* of the model's performance on test examples containing that concept. Notably, ***model performance scales linearly as the concept frequency in pretraining data grows exponentially*** *i.e.*, we observe a consistent log-linear scaling trend. We find that this log-linear trend is robust to controlling for correlated factors (similar samples in pretraining and test data [79]) and testing across different concept distributions along with samples generated entirely synthetically [51].

Our findings indicate that the impressive empirical performance of multimodal models like CLIP and Stable Diffusion can be largely attributed to the presence of test concepts within their vast pretraining datasets, thus their reported empirical performance does not constitute "zero-shot" generalization. Quite the contrary, these models require exponentially more data on a concept to linearly improve their performance on tasks pertaining to that concept, highlighting extreme sample inefficiency.

In our analysis, we additionally document the distribution of concepts encountered in pretraining data and find that:

- **Concept Distribution:** Across all pretraining datasets, the distribution of concepts is long-tailed (see Fig. 5 in Sec. 5), which indicates that a large fraction of concepts are rare. However, given the extreme sample inefficiency observed, what is rare is not properly learned during multimodal pretraining.
- **Concept Correlation across Pretraining Datasets:** The distribution of concepts across different pretraining datasets are strongly correlated (see Tab. 4 in Sec. 5), which suggests web crawls yield surprisingly similar concept distributions across different pretraining data curation strategies, necessitating explicit rebalancing efforts [11, 125].
- **Image-Text Misalignment between Concepts in Pretraining Data:** Concepts often appear in one modality but not the other, which implies significant misalignment (see Tab. 3 in Sec. 5). Our released data artifacts can help image-text alignment efforts at scale by precisely indicating the examples in which modalities misalign. Note that the log-linear trend across both modalities is robust to this misalignment.

To provide a simple benchmark for generalization performance for multimodal models, which controls for the concept frequency in the training set, we introduce a new long-tailed test dataset called "*Let It Wag!*". Current models trained on both openly available datasets (*e.g.*, LAION-2B [103], DataComp-1B [46]) and closed-source datasets (*e.g.*, OpenAI-WIT [91], WebLI [29]) have significant drops in performance, providing evidence that our observations may also transfer to closed-source datasets. We publicly release all our data artifacts (over 300GB), amortising the cost of analyzing the pretraining datasets of multimodal foundation models for a more data-centric understanding of the properties of multimodal models in the future.

Several prior works [91, 46, 82, 42, 83, 74] have investigated the role of pretraining data in affecting performance. Mayilvahanan et al. [79] showed that CLIP's performance is correlated with the similarity between training and test datasets. In other studies on specific areas like question-answering [62] and numerical reasoning [94] in large language models, high train-test set similarity did not fully account for observed performance levels [127]. Our comprehensive analysis of several pretraining image-text datasets significantly adds to this line of work, by (1) showing that concept frequency determines zero-shot performance and (2) pinpointing the exponential need for training data as a fundamental issue for current large-scale multimodal models. We conclude that the key to "zero-shot" generalization capabilities under large-scale training paradigms remains to be found.

2 Concepts in Pretraining Data and Quantifying Frequency

In this section, we outline our methodology for obtaining concept frequencies within pretraining datasets. We first define our concepts of interest, then describe algorithms for extracting their frequencies from images

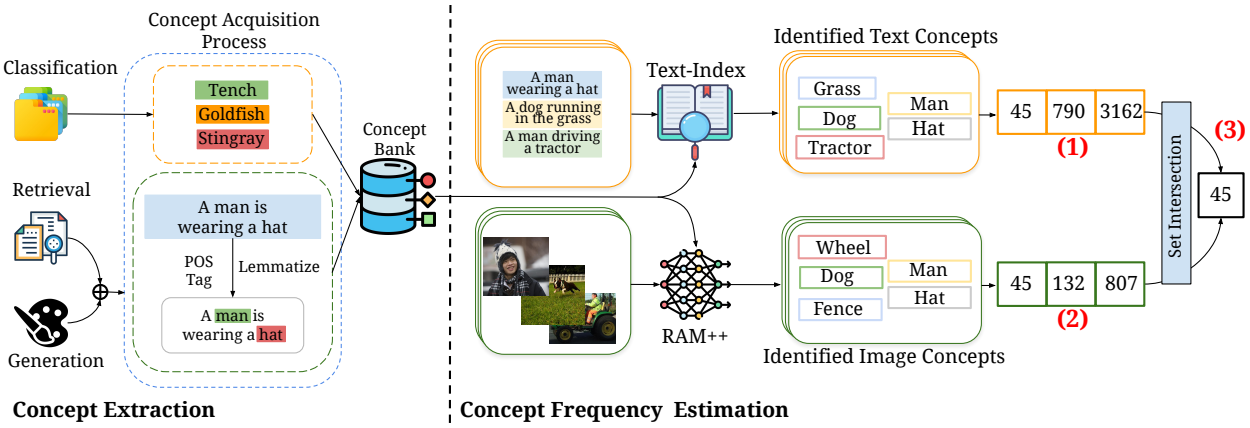


Figure 1: **Concept Extraction and Frequency Estimation Pipeline.** (*left*) We compile 4,029 concepts from 17 classification, 2 retrieval, and 8 image generation prompt datasets. (*right*) We construct efficient indices for both text-search (using standard unigram indexing (1)) and image-search (using RAM++ [59] (2)); intersecting hits from both gives us (3) the image-text matched frequencies per concept.

and text captions of pretraining datasets. Finally, we discuss how to aggregate them to calculate matched image-text concept frequencies. For a schematic overview of our methods, see Fig. 1.

Defining Concepts. We define "concepts" as the specific objects or class categories we seek to analyze in the pretraining datasets. For zero-shot classification tasks, these concepts are the class names, such as the 1,000 classes in ImageNet [35] (*e.g.*, "tench", "goldfish", "stingray"). For image-text retrieval and image generation tasks, concepts are identified as all nouns present in the test set captions or generation prompts, respectively. For example, in the caption, "A man is wearing a hat", we extract "man" and "hat" as relevant concepts. We additionally filter out nouns that are present in less than five downstream evaluation samples to remove ambiguous or irrelevant concepts. Across all our experiments, we collate a list of 4,029 concepts sourced from 17 classification, 2 retrieval, and 8 image generation downstream datasets (see Tab. 1 for details).

Concept Frequency from Text Captions. To enable efficient concept searches, we pre-index all captions from the pretraining datasets, *i.e.*, construct a mapping from concepts to captions. We first use part-of-speech tagging to isolate common and proper nouns and subsequently lemmatize them to standardize word forms [65] with SpaCy [58] . These lemmatized nouns are then cataloged in inverted unigram dictionaries, with each noun being the key and all the indices in the pretraining data samples containing that noun being its values. To determine the frequency of a concept, particularly those composed of multiple words, we examine the concept's individual unigrams within these dictionaries. For multi-word expressions, by intersecting the lists of sample indices corresponding to each unigram, we identify the samples that contain all parts of the concept. The frequency of the concept in the text captions is the count of these intersecting sample indices. Our frequency estimation algorithm hence allows scalable $\mathcal{O}(1)$ search with respect to the number of captions for any given concept in the pretraining dataset captions.

Concept Frequency from Images. Unlike text captions, we do not have a finite vocabulary for pre-indexing pretraining images, and thus cannot perform $\mathcal{O}(1)$ concept lookup. Instead, we collect all the 4,029 downstream concepts and verify their presence in images using a pretrained image tagging model. We tested various open-vocabulary object detectors, image-text matching models and multi-tagging models. We found that RAM++ [59]—an open-set tagging model that tags images based on a predefined list of concepts in a multi-label manner—performs the best. This approach generates a list of pretraining images, each tagged with whether the downstream concepts are present or not, from which we can compute concept frequencies. We provide qualitative examples along with design choice ablations in Appx. F.

Image-Text Matched Concept Frequencies. Finally, we combine the frequencies obtained from both text and image searches to calculate *matched image-text frequencies*. This involves identifying pretraining

Table 1: Pretraining and downstream datasets used in Image-Text (CLIP) experiments.

Dataset Type	Datasets
Pretraining	CC-3M [107] CC-12M [27] YFCC-15M [113] LAION-400M [102]
Classification-Eval	ImageNet [35] SUN397 [123] UCF101 [108] Caltech101 [44] EuroSAT [55] CUB [121] Caltech256 [49] Flowers102 [84] DTD [31] Birdsnap [15] Food101 [20] Stanford-Cars [66] FGVCAircraft [77] Oxford-Pets [87] Country211 [91] CIFAR-10 [67] CIFAR100 [67]
Retrieval-Eval	Flickr-1K [128] COCO-5K [73]

samples where both the image and its associated caption correspond to the concept. By intersecting the lists from our image and text searches, we determine the count of samples that align in both modalities, offering a comprehensive view of concept representation across the dataset. We note that this step is necessary as we observed significant image-text misalignment between concepts in the pretraining datasets (see Tab. 3), hence captions may not reflect what is present in the image and vice-versa. This behaviour has also been alluded to in prior work investigating pretraining data curation strategies [76, 75, 124, 83]. We provide more detailed analysis on image-text misalignment in Sec. 5.

3 Comparing Pretraining Frequency & "Zero-Shot" Performance

Having obtained frequency estimates for our downstream concepts, we now establish the relationship between image-text matched pretraining concept frequencies and zero-shot performance across classification, retrieval, and generation tasks. We first detail our experimental approach and then discuss key results.

3.1 Experimental Setup

We analyze two classes of multimodal models: Image-Text and Text-to-Image. For both, we detail the pretraining and testing datasets, along with their associated evaluation parameters.

3.1.1 Image-Text (CLIP) Models

Datasets. Our evaluation consists of 4 pretraining datasets, 2 downstream retrieval datasets, and 17 downstream classification datasets, presented in Tab. 1, covering a broad spectrum of objects, scenes, and fine-grained distinctions.

Models. We test CLIP [91] models with both ResNet [53] and Vision Transformer [36] architecture, with ViT-B-16 [81] and RN50 [48, 82] trained on CC-3M and CC-12M, ViT-B-16, RN50, and RN101 [61] trained on YFCC-15M, and ViT-B-16, ViT-B-32, and ViT-L-14 trained on LAION400M [102]. We follow `open_clip` [61], `slip` [81] and `cyclip` [48] for all implementation details.

Prompting. For zero-shot classification, we experiment with three prompting strategies: {classname} only, "A photo of a {classname}" and prompt-ensembles [91], which averages over 80 different prompt variations of {classname}. For retrieval, we use the image or the caption as input corresponding to I2T (image-to-text) or T2I (text-to-image) retrieval respectively.

Metrics. We compute mean zero-shot classification accuracy for classification tasks [91]. For retrieval, we assess performance using traditional metrics for both text-to-image and image-to-text retrieval tasks [91] (Recall@1, Recall@5, Recall@10).

3.1.2 Text-to-Image Models

Datasets. Our pretraining dataset is LAION-Aesthetics [103], with downstream evaluations done on subsampled versions of eight datasets as released by HEIM [71]: CUB200 [121], Daily-DALLE [33],

Table 2: Models used in text-to-image (T2I) experiments.

Category	Models
Models	M-Vader [14] DeepFloyd-IF-M [9] DeepFloyd-IF-L [9] DeepFloyd-IF-XL [9] GigaGAN [63] DALL·E Mini [34] DALL.E Mega [34] Promptist+SD-v1.4 [52] Dreamlike-Diffusion-v1.0 [2] Dreamlike Photoreal v2.0 [3] OpenJourney-v1 [4] OpenJourney-v2 [5] SD-Safe-Max [96] SD-Safe-Medium [96] SD-Safe-Strong [96] SD-Safe-Weak [96] SD-v1.4 [96] SD-v1.5 [96] SD-v2-Base [96] SD-v2-1-base [96] Vintedois-Diffusion-v0.1 [7] minDALL.E [97] Lexica-SD-v1.5 [1] Redshift-Diffusion [6]

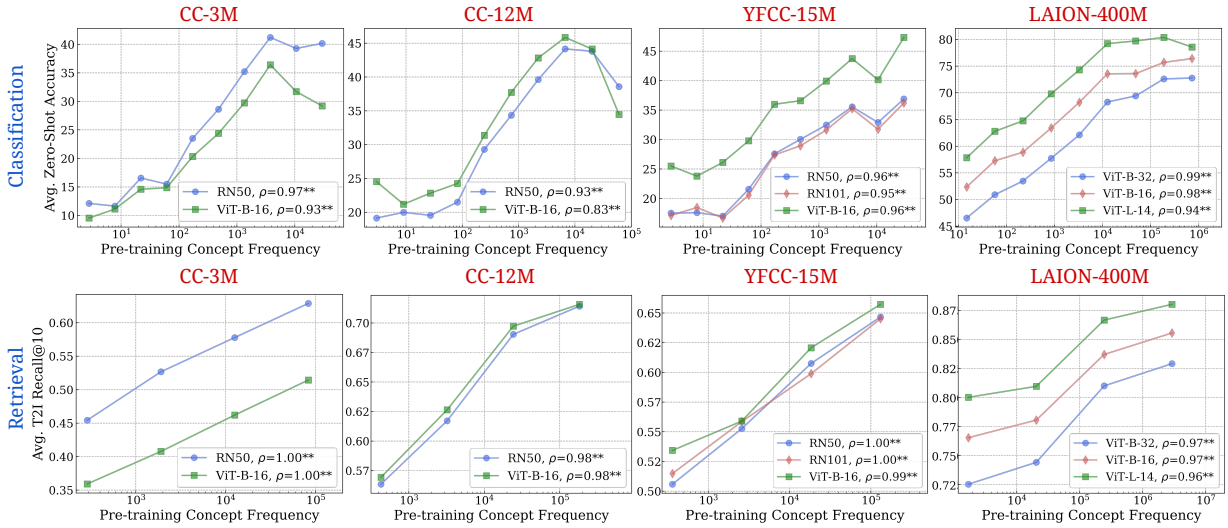


Figure 2: **Log-linear relationships between concept frequency and CLIP zero-shot performance.** Across all tested architectures (RN50, RN101, ViT-B-32, ViT-B-16, ViT-L-14) and pretraining datasets (CC-3M, CC-12M, YFCC-15M, LAION-400M), we observe a consistent linear relationship between CLIP's zero-shot performance on a concept and the log-scaled concept pretraining frequency. This trend holds for both zero-shot classification (results averaged across 17 datasets) and image-text retrieval (results averaged across 2 datasets). ** indicates that the result is significant ($p < 0.05$ with a two-tailed t-test.), and thus we show pearson correlation (ρ) as well.

Detection [30], Parti-Prompts [130], DrawBench [98], COCO-Base [73], Relational Understanding [32] and Winoground [114]. Please refer to HEIM [71] for more details on the evaluation datasets used.

Models. We evaluate 24 T2I models, detailed in Tab. 2. Their sizes range from 0.4B parameters (DeepFloyd-IF-M [9] and DALL·E Mini [34]) to 4.3B parameters (DeepFloyd-IF-XL [9]). We include various Stable Diffusion models [96] as well as variants tuned for specific visual styles [6, 4, 5].

Prompting. Text prompts from the evaluation datasets are used directly to generate images, with 4 image samples generated for each prompt.

Metrics. Evaluation consists of image-text alignment and aesthetic scores. For automated metrics [71], we use expected and max CLIP-score [57] to measure image-text alignment along with expected and max aesthetics-score [102] to measure aesthetics. To verify reliability of the automated metrics, we compare them with human-rated scores (measured on a 5-point grading scale) for both image-text alignment and aesthetics [71]. To supplement the human-rated scores provided by HEIM [71], we confirm our findings by performing a small-scale human evaluation as well (see Appx. C).

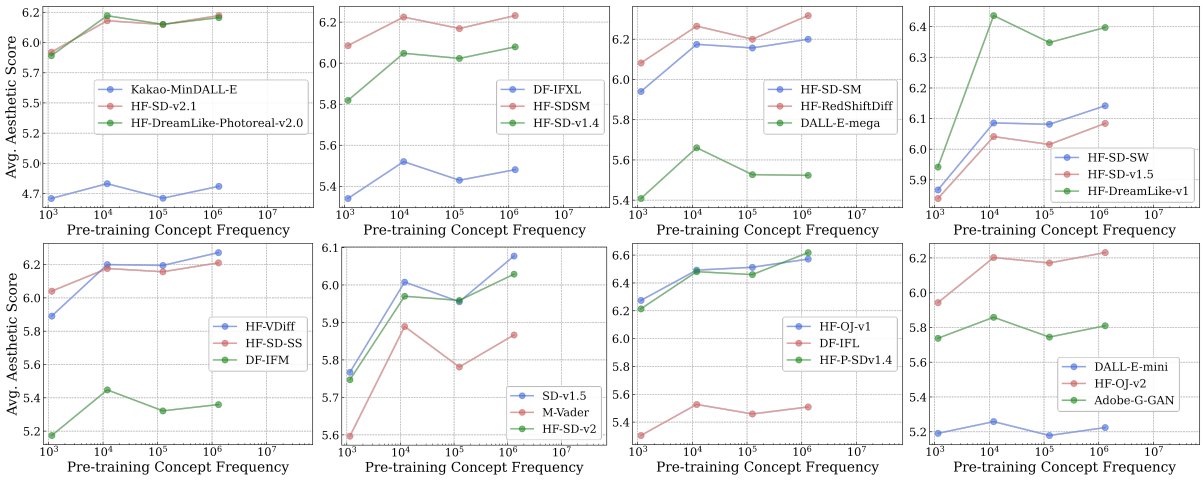


Figure 3: **Log-linear relationships between concept frequency and T2I aesthetic scores.** Across all tested T2I models pretrained on the LAION-Aesthetics dataset, we observe a consistent linear relationship between zero-shot performance on a concept and the log-scaled concept pretraining frequency.

3.2 Result: Pretraining Frequency is Predictive of "Zero-Shot" Performance

We now probe the impact of concept frequency in pretraining datasets on the zero-shot performance of image-text models. We utilize the matched image-text concept frequencies for estimating frequency of concepts during pretraining. Our findings, illustrated comprehensively in Figs. 2 and 3, demonstrate the effect concept frequency has on model performance across various tasks and model types.

Understanding the Plots. The plots in the main paper present text-image (CLIP) models' zero-shot classification results using accuracy and text-to-image retrieval performance using Recall@10. Similarly, we present T2I generative models' performance on image generation tasks using the expected aesthetics score. For the other aforementioned metrics for retrieval as well as other automated generation metrics along with human-rated scores, we find that they show similar trends, and we provide them for reference in Apps. B and C. For clarity, the data presentation is simplified from scatter plots to a cohesive line similar to work from Kandpal et al. [62] and Razeghi et al. [94]. The x-axis is log-scaled, and performance metrics are averaged within bins along this axis for ease-of-visualization of the log-linear correlation. We removed bins containing very few concepts per bin by standard IQR removal [122] following Kandpal et al. [62]. We additionally compute the pearson correlation ρ for each line and provide significance results based on a two-tailed t-test [110].

Key Finding: Log-linear scaling between concept frequency and zero-shot performance. Across all 16 plots, we observe a clear log-linear relationship between concept frequency and zero-shot performance. Note that these plots vary in (i) discriminative vs. generative model types, (ii) classification vs. retrieval tasks, (iii) model architecture and parameter scales, (iv) pretraining datasets with different curation methods and scales, (v) different evaluation metrics, (vi) different prompting strategies for zero-shot classification, and (vii) concept frequencies isolated only from image or text domains (additional experiments which show variation along (v) are presented in Apps. B and C, across (vi) are presented in Appx. A, and across (vii) are presented in Appx. D). The observed log-linear scaling trend persists *across all seven presented dimensions*. Thus, our results clearly reveal data hungry learning, *i.e*, a lack in current multimodal models' ability to learn concepts from pretraining datasets in a sample-efficient manner.

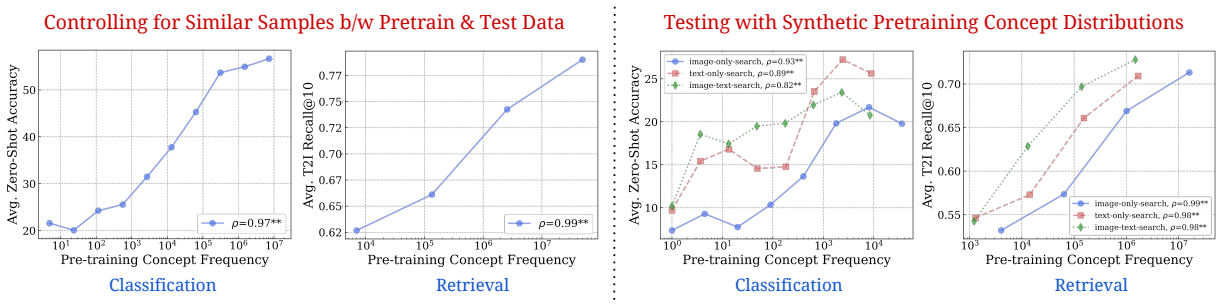


Figure 4: **Stress-testing the log-linear scaling trends.** We provide further evidence that the log-linear relationship between performance and concept frequency holds across different scenarios: (*left*) we control for the effect of "similarity" between the downstream test sets and pretraining datasets, and (*right*) we conduct experiments on an entirely synthetic pretraining distribution with no real-world text-captions or images.

4 Stress-Testing the Concept Frequency-Performance Scaling Trend

In this section, we seek to isolate the effect of concept frequency on zero-shot performance by controlling a widely known influential factor [127, 79]: similarity in distribution between pretraining and downstream test data. Additionally, we aim to validate our hypothesis further by examining the relationship between concept frequency and downstream performance on models trained on pretraining data with synthetically controlled concept distributions, images and captions.

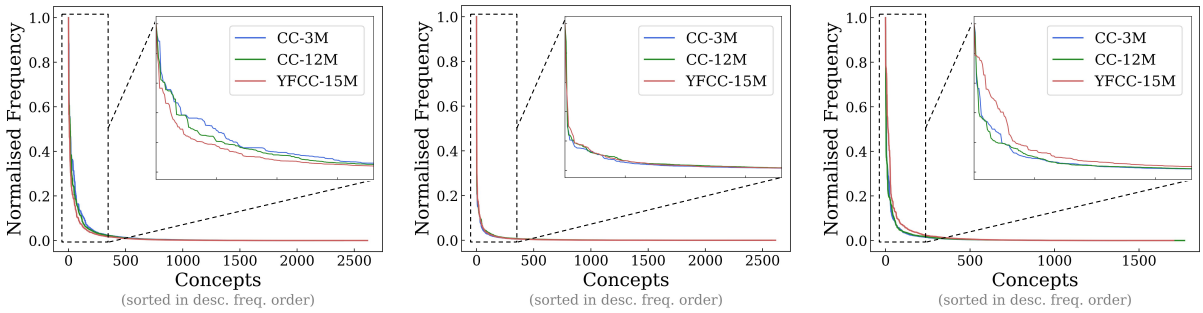
4.1 Controlling for Similar Samples in Pretraining and Downstream Data

Motivation. Prior work has suggested that sample-level similarity between pretraining and downstream datasets impacts model performance [62, 79, 127, 94]. This leaves open the possibility that our frequency-performance results are simply an artifact of this factor, *i.e.*, as concept frequency increases, it is likely that the pretraining dataset also contains more similar samples to the test sets. We hence investigate whether concept frequency remains predictive of downstream performance after controlling for sample-level similarity.

Setup. We use the LAION-200M [10] dataset for this experiment. We first verified that a CLIP-ViT-B-32 model trained on LAION-200M dataset (used to study sample similarity in prior work [79]) exhibits a similar log-linear trend between concept frequency and zero-shot performance. Then, we use the `near_pruning` method from Mayilvahanan et al. [79] to eliminate 50 million samples most similar to the test sets from the pretraining LAION-200M dataset. We provide details for this in Appx. E.1. This removes the most similar samples between pretraining and test sets. We verify that this procedure influences the performance of the model drastically in performance across our aggregate classification and retrieval tasks respectively, replicating the findings of Mayilvahanan et al. [79].

Key Finding: *Concept Frequency still Predictive of Performance.* We repeat our analysis on models trained with this controlled pretraining dataset with 150M samples, and report results on the same downstream classification and retrieval datasets in Fig. 4 (left). Despite the removal of the most similar samples between pretraining and test sets, we still consistently observe a clear log-linear relationship between pretraining frequency of test set concepts and zero-shot performance.

Conclusion. This analysis reaffirms that, despite removing pretraining samples closely related to the test sets, the log-linear relationship between concept frequency and zero-shot performance persists. Note that this is despite substantial decreases in absolute performance, highlighting the robustness of concept frequency as a performance indicator.



(a) Text search counts (b) Image search counts (c) Image-text search counts

Figure 5: **Concept distribution of pre-training datasets is highly long-tailed.** We showcase the distribution of pre-training frequencies of all concepts aggregated across all our downstream classification datasets. Across all three pre-training datasets, we observe very heavy tails. We normalize the concept frequencies and remove concepts with 0 counts for improved readability.

4.2 Testing Generalization to Purely Synthetic Concept and Data Distributions

Motivation. Sampling across real-world data might not result in significant differences in concept distribution, as we will later show in Sec. 5. Hence, we repeat our analysis on a synthetic dataset designed with an explicitly different concept distribution [51]. This evaluation aims to understand if pretraining concept frequency remains a significant performance predictor within a synthetic concept distribution, generalizing even on models pretrained on entirely synthetic images and captions.

Setup. The SynthCI-30M dataset [51] introduces a novel concept distribution, generating 30 million synthetic image-text pairs. Utilizing the publicly available data and models from this benchmark, we explore the relationship between concept frequency and model performance in this synthetic data regime.

Key Finding: *Concept Frequency is still Predictive of Performance.* We report results on models trained with their controlled dataset in Fig. 4 (right). We still consistently observe a clear log-linear relationship between concept frequency and zero-shot performance.

Conclusion. This consistency highlights that concept frequency is a robust indicator of model performance, extending even to entirely synthetically constructed datasets and pretraining concept distributions.

5 Additional Insights from Pretraining Concept Frequencies

We now present notable observations concerning the distribution of downstream concept frequencies across text, image, and text-image matched modalities in pretraining datasets.

Finding 1: *Pretraining Datasets Exhibit Long-tailed Concept Distribution.* Our analysis in Fig. 5 reveals an extremely long-tailed distribution of concept frequencies in pretraining datasets, with over two-thirds of concepts occurring at almost negligible frequencies relative to the size of the datasets. Our observations support the findings of past work that have noted the long-tailed distribution of large-scale language datasets [25, 88, 136]. As we observed with the log-linear trend, this distribution directly reflects disparities in performance.

Finding 2: *Misalignment Between Concepts in Image-Text Pairs.* We investigated the alignment of concepts within paired pretraining image-text data. Perfect image-text alignment is defined as every image-text pair containing the same concepts. Previous studies have qualitatively discussed the problem of misalignment in large image-text datasets [75, 124, 76]. Our analysis enables us to quantify this *misalignment degree*—for each image-text pair in the pretraining dataset, we find the concepts that are matched to the image and the text caption independently. If there are no intersecting concepts from the independent image

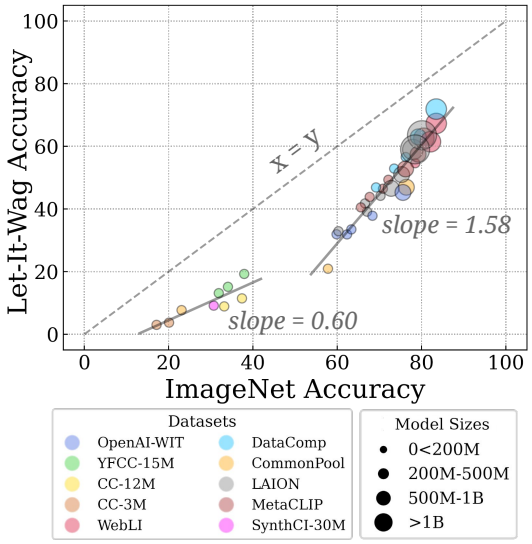


Figure 6: **Large-drops in accuracy on "*Let It Wag!*".** Across all 40 tested models, we note large performance drops compared to ImageNet. Further, we note that the gap in performance seems to be decreasing for higher-capacity models as demonstrated by the large positive slope (1.58) for the larger models.

Dataset/ Misalignment	Number of Misaligned pairs	Misalignment Degree (%)
CC3M	557,683	16.81%
CC12M	2,143,784	17.25%
YFCC15M	5,409,248	36.48%
LAION-A	23,104,076	14.34%
LAION400M	21,996,097	5.31%

Table 3: For each pretraining dataset, we present the number of misaligned image-text pairs and the *misalignment degree*: the fraction of misalignment pairs in the dataset.

Correlations	CC3M	CC12M	YFCC15M	L400M
CC3M	1.00	0.79	0.96	0.63
CC12M	–	1.00	0.97	0.74
YFCC15M	–	–	1.00	0.76
L400M	–	–	–	1.00

Table 4: We compute correlation in concept frequency across pretraining datasets. Despite significant differences in scale and curation, we consistently observe strong correlation.

and text hits, we count that pair as misaligned (detailed algorithm provided in Appx. G). Tab. 3 shows the high degree of misalignment in all image-text pairs. To the best of our knowledge, this is the first attempt to explicitly quantify the degree of misalignment in pretraining image-text datasets. We release the precise misaligned image-text pairs in the pretraining datasets to enable better data curation.

Finding 3: ***Concept Frequencies Across Datasets are Correlated.*** Despite vast differences in the size (ranging from 3M to 400M samples) and curation strategies of the datasets analyzed, we discovered a surprisingly high correlation in concept frequencies across them, as presented in Tab. 4. This consistency suggests that the internet, as the common source of these datasets, naturally exhibits a long-tailed distribution, influencing any dataset derived from it to also display similar long-tailed behavior. This result inspired the "*Let It Wag!*" dataset.

6 Testing the Tail: *Let It Wag!*

Motivation. From the previous sections, we have identified a consistent long-tailed concept distribution, highlighting the scarcity of certain concepts on the web. This observation forms the basis of our hypothesis that models are likely to underperform when tested against data distributions that are heavily long-tailed. To test this, we carefully curate 290 concepts that were identified as the least frequent across all pretraining datasets. This includes concepts like an `A310` aircraft, a `wormsnake`, and a `tropical kingbird`. We then use these concepts to create a classification test set, "*Let It Wag!*".

Dataset Details. The *"Let It Wag!"* classification dataset comprises 130K test samples downloaded from the web using the method of Prabhu et al. [90]. The test samples are evenly distributed across 290 categories that represent long-tailed concepts. From the list of curated concepts, we download test set images, deduplicate them, remove outliers, and finally manually clean and hand-verify the labels.

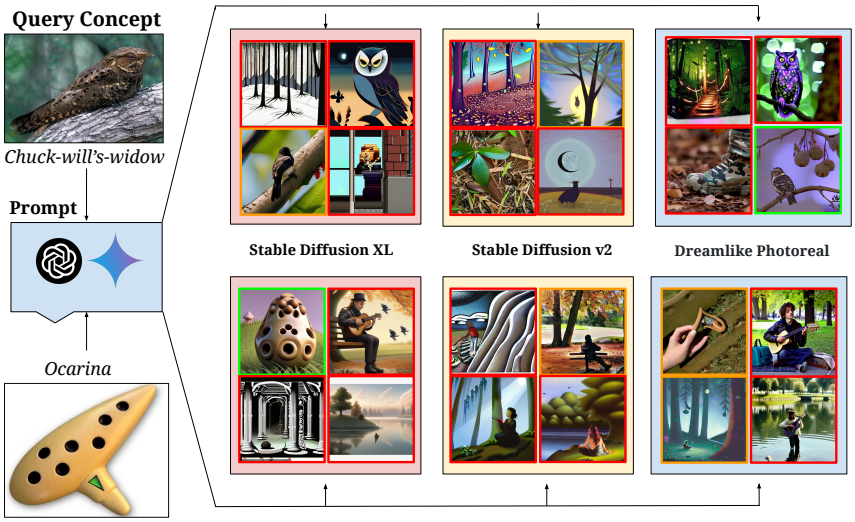


Figure 7: **Qualitative results on the *"Let It Wag!"* dataset categories demonstrate failure cases of state-of-the-art T2I models on long-tailed concepts.** In our experiments, we create 4 text prompts for each category using Gemini [112] and GPT4 [12] which are fed to 3 Stable Diffusion [96] models. Generation with red border is incorrect, with green border is correct and with yellow border is ambiguous. We observe that despite advances in high-fidelity image generation, there is scope for improvement for such concepts.

Analysis Details. We ran both classification and image generation experiments on *"Let It Wag!"*. For classification, we assessed the performance of 40 text-image (CLIP) models on the *"Let It Wag!"* classification dataset, using an ensemble of 80 prompts from Radford et al. [91]. For the generative task, we utilized SD-XL [89] , SD-v2 [96], and Dreamlike-Photoreal-v2.0 [3] to generate images for the long-tailed concepts. For each model, we ran 50 diffusion steps, maintaining default settings for all other parameters.

Text-Image Classification Results. We showcase the results of our long-tailed classification task in Fig. 6—we plot results of all models on both "*Let It Wag!*" (y-axis) and ImageNet (x-axis). We observe that all models underperform by large margins on the long-tailed *"Let It Wag!"* dataset (upto 20% lower absolute accuracies compared to ImageNet). This performance drop-off generalises across all model scales and 10 different pretraining data distributions, reinforcing the notion that all web-sourced pretraining datasets are inherently constrained to be long-tailed. With that said, note that the higher capacity models (fitted line with slope=1.58 in Fig. 6) seem to be closing the gap to ImageNet performance, indicating improved performance on the long-tailed concepts.

T2I Generation Results. We provide a qualitative analysis on image generation for assessing T2I models on rare concepts in Fig. 7. For diversity, we generate prompts using Gemini [112] (top row of generated images) and GPT4 [12] (bottom row of generated images). Green borders represent correct generations, red borders represent incorrect generations and yellow borders represent ambiguous generation. While descriptive prompting generally aids in improving the quality of generated images [52], we still observe T2I models failing to comprehend and accurately represent many concepts in our *"Let It Wag!"* dataset. Some failure cases involve misrepresenting activities (such as `Pizza Tossing` or `Cricket Bowling` as shown in Fig. 24), generating the wrong concept (`Chuck-will's-widow` as shown in Fig. 7 top), as well as not comprehending the concept at all (`Ocarina` in Fig. 7 bottom). We can see that Stable Diffusion models are prone to the long tail qualitatively—we also provide quantitative results in Appx. H.1.

Conclusion. Across both the classification and generation experiments, we have showcased that current multimodal models predictably underperform, regardless of their model scale or pretraining datasets. This suggests a need for better strategies for sample-efficient learning on the long-tail.

7 Related Work

Effect of Pre-training Data on Downstream Data. Several data-centric prior works [91, 46, 82, 42, 83, 74, 124, 125, 135, 109, 78, 92, 99, 100, 38, 26, 95] have highlighted the importance of pretraining data in affecting performance. Fang et al. [42] robustly demonstrated that pretraining data diversity is the key property underlying CLIP's strong out-of-distribution generalisation behaviour. Similarly, Berlot-Attwell et al. [16] showed that attribute diversity is crucial for compositional generalization [60], namely systematicity [45]. Nguyen et al. [82] extended the Fang et al. [42] analysis to show that differences in data distributions can predictably change model performance, and that this behaviour can lead to effective data mixing strategies at pretraining time. Mayilvahanan et al. [79] complemented this research direction by showing that CLIP's performance is correlated with the similarity between training and test datasets. Udandarao et al. [118] further showed that the frequency of certain visual data-types in the LAION-2B dataset was roughly correlated to the performance of CLIP models in identifying visual data-types. Our findings further pinpoint that the frequency of concept occurrences is a key indicator of performance. This complements existing research in specific areas like question-answering [62] and numerical reasoning [94] in large language models, where high train-test set similarity does not fully account for observed performance levels [127]. Concurrent to our work, Parashar et al. [86] also explore the problem of long-tailed concepts in the LAION-2B dataset and how it affects performance of CLIP models, supporting our findings. In contrast to their work, we look at count separately in image and text modalities, as well as across pretraining sets, and do a number of control experiments to thoroughly test the robustness of our result. Finally, our demonstration that the long tail yields a log-linear trend explicitly indicates exponential sample inefficiency in large-scale pretrained models.

Data-centric analyses. Our work also adds to the plethora of work that aims to understand and explore the composition of large-scale datasets, and uses data as a medium for improving downstream tasks. Prior work has noted the importance of data for improving model performance on a generalised set of tasks [46, 11, 40, 13, 106]. For instance, several works utilise retrieved and synthetic data for adapting foundation models on a broad set of downstream tasks [119, 54, 115, 21, 101, 134, 90]. Maini et al. [76] observed the existence of "text-centric" clusters in LAION-2B and measured its impact on downstream performance. Other work has seeked to target the misalignment problem that we quantified in Tab. 3 by explicit recaptioning of pretraining datasets [68, 28, 120, 131, 83, 17]. Further, studies have also shown that by better data pruning strategies, neural scaling laws can be made more efficient than a power-law [109, 10]. Prior work has also showcased that large-scale datasets suffer from extreme redundancy in concepts, and high degrees of toxic and biased content [39, 116]. Further research has showcased the downstream effects that such biases during pretraining induce in state-of-the art models [19, 104, 18, 47]. Our work tackles the issue of long-tailed concepts in pretraining datasets, and shows that this is an important research direction to focus efforts on.

8 Conclusions and Open Problems

In this work, we delved into the five pretraining datasets of 34 multimodal vision-language models, analyzing the distribution and composition of concepts within, generating over 300GB of data artifacts that we publicly release. Our findings reveal that across concepts, significant improvements in zero-shot performance require exponentially more data, following a log-linear scaling trend. This pattern persists despite controlling for similarities between pretraining and downstream datasets or even when testing models on entirely synthetic data distributions. Further, all tested models consistently underperformed on the *"Let it Wag!"* dataset, which we systematically constructed from our findings to test for long-tail concepts. This underlines a critical reassessment of what "zero-shot" generalization entails for multimodal models, highlighting the limitations in their current generalization capabilities. We highlight a few exciting avenues for future research to bridge these gaps or obtain further insights:

Understanding Image-Text Misalignments. One can explore the origins of misalignments between images and texts, such as the limitations of exact matching for concept identification in captions, inaccuracies from the RAM++ tagging model, or captions that are either too noisy or irrelevant.