The Neglected Tails in Vision-Language Models

Shubham Parashar[*1] Zhiqiu Lin[*2] Tian Liu[*1] Xiangjue Dong[1]
Yanan Li[3] Deva Ramanan[2] James Caverlee[1] Shu Kong[14†]
[1]Texas A&M University [2]Carnegie Mellon University [3]Zhejiang Lab [4]University of Macau

Abstract

Vision-language models (VLMs) excel in zero-shot recognition but their performance varies greatly across different visual concepts. For example, although CLIP achieves impressive accuracy on ImageNet (60-80%), its performance drops below 10% for more than ten concepts like `night snake`*, presumably due to their limited presence in the pretraining data. However, measuring the frequency of concepts in VLMs' large-scale datasets is challenging. We address this by using large language models (LLMs) to count the number of pretraining texts that contain synonyms of these concepts. Our analysis confirms that popular datasets, such as LAION, exhibit a long-tailed concept distribution, yielding biased performance in VLMs. We also find that downstream applications of VLMs, including visual chatbots (e.g., GPT-4V) and text-to-image models (e.g., Stable Diffusion), often fail to recognize or generate images of rare concepts identified by our method. To mitigate the imbalanced performance of zero-shot VLMs, we propose* **RE***trieval-***A***ugmented* **L***earning (REAL). First, instead of prompting VLMs using the original class names, REAL uses their most frequent synonyms found in pretraining texts. This simple change already outperforms costly human-engineered and LLM-enriched prompts over nine benchmark datasets. Second, REAL trains a linear classifier on a small yet balanced set of pretraining data retrieved using concept synonyms. REAL surpasses the previous zero-shot SOTA, using 400× less storage and 10,000× less training time!*

1. Introduction

Vision-language models (VLMs) such as CLIP [35] play a pivotal role in mainstream multimodal systems, including visual chatbots [23, 31] and text-to-image generation [3, 36]. Their efficacy largely stems from their web-scale image-text pretraining datasets like LAION [37, 38] that cover a wide range of visual concepts.

Imbalanced performance of VLMs. Despite their strong capabilities in visual understanding, VLMs often exhibit biased performance in downstream tasks. For instance, in zero-shot visual recognition tasks (which do not use training samples), CLIP [35] achieves up to 80% mean accuracy across 1K semantic classes on ImageNet [10] but less than

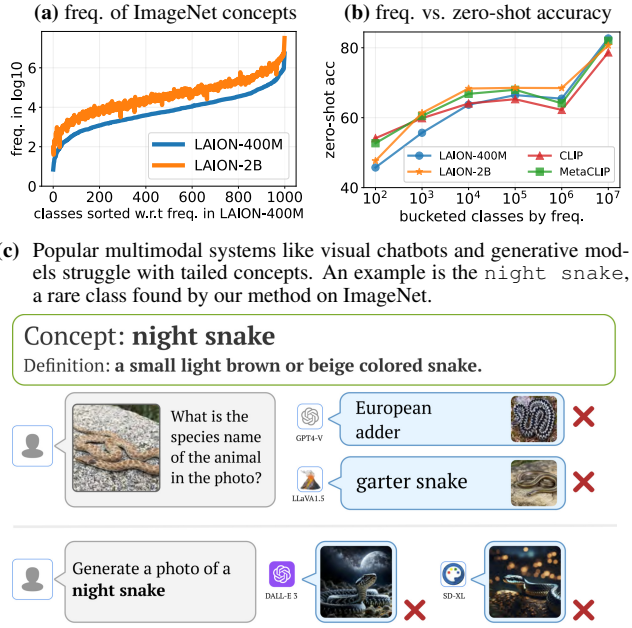

Figure 1. **Vision-language models (VLMs) inherit long tails from their pretraining data. (a)** VLMs' pretraining datasets, such as LAION-400M [37] and LAION-2B [38], exhibit long-tailed distributions for visual concepts defined in downstream tasks like ImageNet [10]. We sort the 1K ImageNet classes according to their frequency in LAION-400M calculated with our concept frequency estimation method (cf. Fig. 2). **(b)** For zero-shot recognition, OpenCLIP models [16] trained on LAION-400M and LAION-2B respectively yield per-class accuracies that strongly correlate with the long-tailed concept frequencies (binned on a log-scale). Interestingly, other VLMs such as CLIP [35] and MetaCLIP [49] (trained on private data) also show similar imbalanced performances, likely because their web-scraped pretraining datasets follow similar long-tailed distributions of the real world. **(c)** Our method helps identify rare concepts, such as the ***night snake***, which is one of the most tailed ImageNet concepts. We show that state-of-the-art multimodal systems, including GPT-4V [50], LLaVA [23], DALL-E 3 [41], and SD-XL [2], all fail to recognize or generate it. The supplement shows more examples.

[*]Co-first authors; [†]corresponding author. Project page at link.

$<10\%$ on specific classes such as `night snake` (Fig. 1b). This motivates us to explore *why* VLMs are imbalanced, a crucial yet ever-neglected issue.

Why do VLMs exhibit imbalanced performance? It is commonly believed that foundational VLMs inherit biases [27] from web-scale pretraining data. However, we find no direct evidence linking VLMs' imbalanced performance to the concept distribution in pretraining data, likely because there is no such tool for measuring *concept frequency* in large multimodal datasets like LAION [37].

Concept frequency estimation. Estimating class frequency in a typical classification dataset is straightforward, where counting class occurrences using annotated labels is sufficient. However, estimating concept frequency in VLMs' pretraining datasets is more complex because their free-form texts (or captions) contain significant *lexical variations*, e.g., a `sneaker` might also be called a `running shoe` or a `trainer`. To address this, we leverage off-the-shelf large language models (LLMs). Fig. 2 illustrates our approach, which begins by asking an LLM (such as ChatGPT [31]) to enumerate all *synonyms* for a specific concept. We then use string matching to find all pretraining texts that contain the concept or its synonyms. However, due to *linguistic ambiguity*, the initially retrieved texts may contain irrelevant phrases, such as "`tiger shark in water`" for the target concept `tiger` (a mammal). We again use an LLM to filter out such irrelevant texts, and conduct human studies to verify the accuracy of our frequency measure. Our method for calculating concept frequency unveils three key insights: (1) it confirms that VLMs' pretraining data is indeed long-tailed (Fig. 1a); (2) it shows VLMs perform better on well-represented concepts and worse on under-represented ones (Fig. 1b); and (3) it explains why recent multimodal systems (e.g., GPT-4Vision and DALL-E 3 in Fig. 1c) struggle with rare concepts. Our analysis also provides technical insights to counteract the bias in VLMs, leading to state-of-the-art zero-shot performance in downstream tasks such as image recognition.

State-of-the-art zero-shot recognition. Motivated by our frequency estimation, we introduce **RE**trieval-**A**ugmented **L**earning (REAL) to mitigate biased performance of zero-shot VLMs. REAL has two variants. First, observing that some synonyms are more frequent in VLM's pretraining data than the original concept names, we propose **REAL-Prompt**. Specifically, we replace given concept names with their *most frequent synonyms*. For example, `cash machine` is replaced with `ATM`, which is ten times more frequent in LAION (Fig. 3). This minor change already surpasses costly human-engineered [35] and LLM-enriched prompts like CuPL [34] (cf. Table 1). Second, inspired by retrieval-augmented strategies [12, 19, 20, 24, 44], we introduce **REAL-Linear**, which reuses relevant pretraining data to better adapt VLMs without using data from downstream tasks. The key idea is to retrieve a small, balanced set of images from pretraining data to train a robust linear classifier [22]. In contrast to prior arts [24, 44] that perform costly *feature-based* retrieval by running VLMs to compute image or text features, our method implements *text-based* retrieval via string matching, achieving a significant boost in efficiency. As a result, our REAL resoundingly outperforms the recent retrieval-augmented SOTA, REACT [24], using 400× less storage and 10,000× less training time (cf. Table 1 and 3)!

Contributions. We summarize our major contributions.

- We propose a method for estimating the frequency of visual concepts in VLMs' large-scale pretraining data. Our analysis, for the first time, exposes long-tailed concept distributions in popular datasets like LAION and reveals systematic failures of VLMs [2, 23, 41, 50] in handling rare concepts.
- We propose REAL to address the biased performance of zero-shot VLMs. REAL establishes a new state-of-the-art in zero-shot recognition through its efficient prompting and retrieval-augmented training strategies.

2. Related Works

Biases in foundation VLMs. Pretrained on large-scale multimodal datasets [4, 8, 45], VLMs often exhibit biases related to gender, race, and geography [27], leading to imbalanced predictions in downstream tasks [1, 29]. Recent studies [39, 47, 52] seek to mitigate imbalanced predictions of VLMs by training on additional data from downstream tasks. Despite these efforts, there is no analysis of the imbalances within the pretraining data itself. Our study presents the first examination of VLMs' pretraining datasets, revealing a long-tailed distribution of concepts that closely correlates with VLMs' imbalanced performance. Our analytical tool also identifies rare concepts that VLMs have insufficiently learned, thereby preventing biases in downstream applications.

Prompting VLMs for zero-shot recognition. VLMs excel in zero-shot recognition tasks, where only the names of target concepts are provided without corresponding training images. CLIP [35] shows that putting given concept names in human-engineered prompt templates, such as "a photo of a {`class`}" and "a demonstration of a {`class`}", often enhances zero-shot recognition. LLM-enriched approaches like DCLIP [28] and CuPL [34] create class-specific prompts by appending rich visual descriptions generated by LLM, for example, "a tiger, which has sharp claws". While most works focus on refining prompt templates [25, 40, 51], they use the provided class names as is. A recent work [32] suggests that prompting with common English names instead of Latin scientific names improves zero-shot recognition of fine-grained species. Differently, our REAL-Prompt replaces given class names

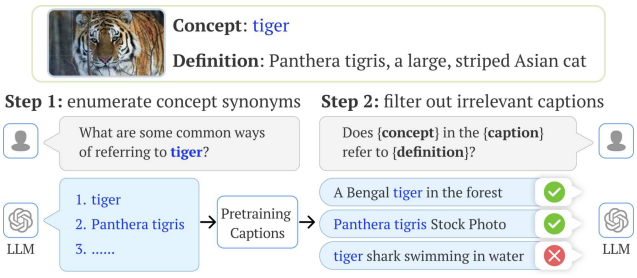

Figure 2. **Using large language models (LLMs) to estimate concept frequency in a VLM's pretraining dataset.** We conduct the frequency estimation using publicly available LAION [37] datasets. First, since a visual concept can be expressed in various ways, we ask an LLM (e.g., ChatGPT [31]) to enumerate all its synonyms to search for potentially relevant pretraining texts. For example, for `tiger`, we retrieve all captions that contain not only "`tiger`" but also its synonyms such as "`Panthera tigris`". Second, we filter out irrelevant captions that do not refer to the target concept by its definition. For example, although "*tiger shark swimming in water*" contains "*tiger*", it actually refers to a type of shark, not the animal `tiger` as defined by "Panthera tigris, a large, striped Asian cat". We conduct the filtering process by querying an LLM Llama-2 [42] (cf. Section 3).

with their most common synonyms found in the pretraining texts. This simple change outperforms existing methods with much less ChatGPT querying costs. Moreover, our approach can be combined with existing prompt templates to further improve performance.

Retrieval-augmented strategy. Introduced in the natural language processing (NLP) literature, this strategy addresses challenging tasks such as knowledge-based question-answering [12, 19] by retrieving relevant facts from an external knowledge source (e.g., the Internet or a pretraining dataset) to ground LLMs on the most accurate, up-to-date information. To improve zero-shot visual recognition, recent works [20, 24, 44] finetune VLMs on images retrieved from VLMs' pretraining datasets. While methods like REACT [24] are effective, they demand significant computing resources, including hundreds of GPU hours and extensive memory for large-batch contrastive finetuning. In contrast, our REAL-Linear uses fast string matching to retrieve data based on concept synonyms, thus avoiding costly VLM-based feature extraction. Moreover, we train a parameter-efficient linear classifier atop the frozen VLM [22], which significantly enhances efficiency. Our method not only sets a new state-of-the-art in zero-shot recognition, but also opens avenues for retrieval-augmented research within a modest computational budget.

3. The Long-Tailed Concept Distribution

This section outlines our approach for estimating concept frequency in VLM's pretraining data and presents our key findings from the analysis.

3.1. Concept Frequency Estimation

The foremost challenge in estimating concept frequency is the sheer size of a VLM's pretraining dataset. For example, the popular open-source LAION-400M [38] dataset (used for training OpenCLIP [16]) takes ∼10TB of physical storage space. Instead, we estimate concept frequency directly from pretraining *texts*, eliminating the need to download images. This allows us to download only the text metadata, requiring only ∼60GB space for LAION-400M. Next, we proceed with the following two steps (cf. Fig. 2).

Step 1: Deriving synonyms for target concepts. A well-known issue in NLP is *lexical variation*, meaning that a concept can be expressed in multiple ways. For example, "`sneaker`" can be referred to as "`running shoes`" or "`athletic footwear`", and "`tiger`" may also be called "`Panthera tigris`". To account for lexical variation, we first derive a list of synonyms[1] for a given visual concept. To do so, we turn to an off-the-shelf LLM (e.g., ChatGPT [31]), by querying a simple question "*What are some common ways of referring to* {**concept**}?". Then, we use string matching to retrieve all pretraining texts containing these synonyms. This matching process is remarkably efficient — it takes only 5 hours to retrieve 400M pretraining texts from the LAION-2B dataset for ImageNet's 1K concepts. Importantly, these derived concept synonyms can also be used to improve downstream zero-shot recognition, which we discuss in Sec. 4.

Step 2: Filtering out irrelevant pretraining texts. Using simple string matching may be inaccurate because it could retrieve irrelevant captions due to *linguistic ambiguities*. For instance, the concept "`tiger`", defined as "*Panthera tigris, a large, striped Asian cat*", might appear in irrelevant contexts in the retrieved captions such as "*tiger shark swimming in water*" and "*Tiger Woods, a famous golf player*". In these retrieved texts, the "tiger" actually refers to a shark species and a celebrity, respectively. To tackle these ambiguities at an affordable cost, we utilize the state-of-the-art open-source LLM Llama-2 [42]. For each retrieved text, we ask:

> Does {`concept`} in the {**caption**} refer to {**definition**}?

In real-world applications, concepts are typically defined in the labeling policies of downstream tasks [6, 11]. In this work, to align with standard benchmarks, we adopt definitions from Elevater [21]. Finally, retrieved captions that are identified as irrelevant to the target concept by the LLM are excluded. We only count the remaining retrieved text to estimate concept frequency.

[1]For simplicity, we use the term "synonyms" in a broader sense to encompass all forms of lexical variation, including but not limited to traditional synonyms, idiomatic expressions, and different phrasings that convey the same or similar meanings.

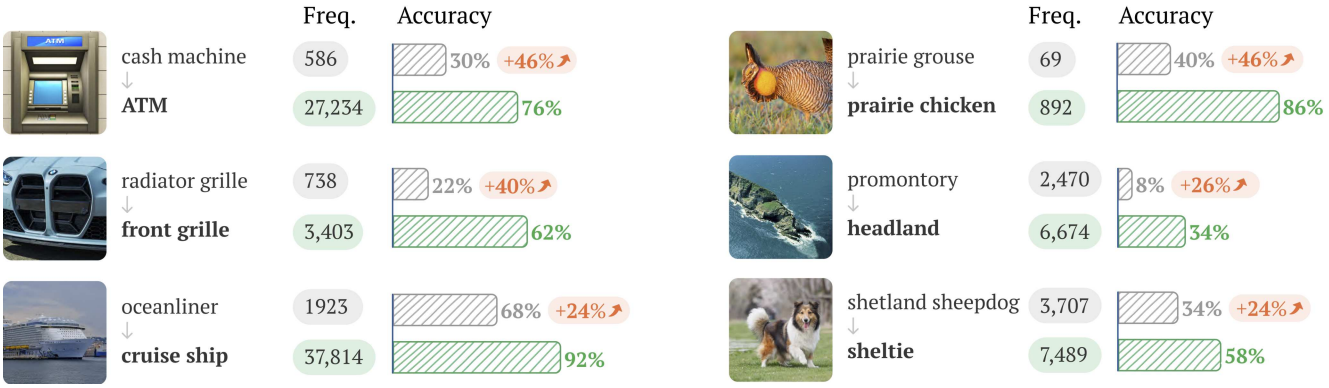

Figure 3. **Demonstration of REAL-Prompt**, which replaces the given concept names (e.g., "cash machine") with their most frequent synonyms (e.g., "ATM") in the prompt template, e.g., "*a photo of* {*concept*}". REAL-Prompt uses an LLM (ChatGPT) to obtain a list of synonyms for a given concept, followed by string matching to identify the most frequently occurring ones in pretraining texts. We demonstrate REAL-Prompt on some ImageNet concepts with their most frequent synonyms, frequencies (in LAION-400M), and per-class accuracies (OpenCLIP ViT-B/32). The simple name change in prompts significantly improves zero-shot recognition. We detail the procedure for REAL-Prompt in Sec. 4.1 and compare against prior works in Table 1.

3.2. Discussions and Remarks

Human-in-the-loop validation. As VLMs' pretraining data does not contain ground-truth concept labels, we perform manual validation to ensure LLM performs well in filtering the irrelevant texts. To do so, we first construct a small validation set by downloading a balanced set of pretraining data (32 image-text pairs per concept). Then, for each concept, we tune the concept definitions for Llama-2 till reaching >85% retrieval precision on the validation set. In particular, since [21] releases multiple definitions per concept, we select the best ones that lead to the highest precision over the validation set. For example, the class `samoyed` in ImageNet refers to a dog breed; we find the definition "*a breed of large white herding dog with a thick coat, native to the Ural Mountains*" to be more precise than others, e.g., "*a member of a people inhabiting northwestern Siberia*". To facilitate future research, we will open-source our code for LLM-based analysis and release all concept synonyms and filtered captions.

The prevalent long tails in VLMs. Our analysis reveals an ever-neglected long-tailed distribution of visual concepts (from standard benchmarks like ImageNet [10]) within widely-used pretraining datasets like LAION-400M and LAION-2B (cf. Fig. 1a). Additionally, we plot per-class zero-shot accuracies of OpenCLIP models (pretrained on LAION), establishing a strong correlation between the long-tailed distribution of concepts and the imbalanced performance of VLMs (cf. Fig. 1b). We also plot the per-class accuracies of CLIP [35] and MetaCLIP [49], which are trained on private datasets. Interestingly, they show similar imbalanced performance across ImageNet's concepts, likely because Internet data follows a similar long-tailed distribution. We observe the same trend across eight more benchmarks (cf. Table 7), e.g., Flowers [30] and Pets [33]. Notably, our frequency estimation method helps find rare (tailed) concepts that challenge popular multimodal systems including the state-of-the-art GPT-4Vision [31] and DALL-E 3 [41] (cf. Fig. 8 and 9). In sum, our analysis shows that long-tailed issues are prevalent in VLMs.

Are long tails inevitable in large-scale data curation? Despite the imbalanced performance of CLIP [35] and MetaCLIP [49], their pretraining datasets are actually created using a "balanced" sampling strategy. Specifically, they define 500K word phrases as web search queries to collect approximately equal numbers of image-text samples for each. However, our analysis indicates that these datasets (which are not fully disclosed to the public) might not be as balanced as intended. We identify key insights into the observed imbalances by examining [49]'s query statistics:

- **Internet data are naturally long-tailed:** Despite balanced sampling in [49], the resulting distribution is still long-tailed. For example, while each search query is capped at 20K images, the average is around ∼1.2K images per query. In fact, we find that over half of the queries, such as "`tailed frog`" (a frog), "`gyromitra`" (a fungus), and "`poke bonnet`" (a traditional hat), have less than 50 samples, likely because these concepts are rare on the web.
- **Limitations of query-based balancing:** Balancing per query does not guarantee a balanced distribution of concepts. For example, [49] inadvertently include overlapping queries such as "`sneaker`" and "`running shoes`", which can lead to the overrepresentation of certain concepts. Moreover, samples retrieved for a single query often contain other concepts. For instance, samples featuring "`keyboard`" may also frequently include "`mouse`" and "`computer`".

As curating a perfectly balanced pretraining dataset is challenging, we recommend that researchers acknowledge the presence of long tails and prioritize addressing VLM's imbalanced performances in downstream applications.

4. Retrieval-Augmented Learning

To address the biased performance of VLMs in zero-shot recognition, we propose **RE**trieval-**A**ugmented **L**earning (REAL), which improves performance without using any data from downstream tasks by retrieving pretraining data relevant to the target concepts. REAL has two variants: REAL-Prompt and REAL-Linear. REAL-Prompt is a novel prompting strategy that replaces the original concept name with its most frequent synonym found in pretraining texts. REAL-Linear retrieves images relevant to the concepts from the pretraining data to form a more balanced subset for training a robust linear classifier. Below we elaborate on these two methods.

4.1. REAL-Prompt for Zero-Shot Prompting

In our analysis, we discover that some synonyms for a concept might appear more frequently in pretraining texts than the concept itself. Therefore, we propose using the most frequent synonym of a concept to construct prompts. Specifically, we utilize an LLM (ChatGPT [31]) to enumerate all synonyms for each concept. Next, we count their individual frequencies in the pretraining texts by string matching. We use the most frequent synonym of each concept in the prompt to construct an off-the-shelf classifier W_{zs} for zero-shot recognition following CLIP [35]. This simple change leads to significantly better zero-shot accuracy than using the original concept names (cf. Fig. 3) released by [35], which are hand-crafted over one year. As shown in Table 1, our approach also outperforms recent LLM-based prompting methods that use additional visual descriptors along with the given concept names (e.g., DCLIP [28] and CuPL [34]).

Synonym filtering using OpenCLIP's text encoder. ChatGPT sometimes generates noisy synonyms. For example, for "`tiger`", it lists "`big cat`" as a synonym, which could be easily confused with another ImageNet class "`tabby cat`". To address this, we use OpenCLIP's text encoder to filter out synonyms that might be confused with other downstream concepts. We retain only those synonyms that have the highest cosine similarity scores with their original class names. This filtering step is fully automated, ensuring a fair comparison with [28, 34] that perform LLM-based prompting without human input. We show that this step is crucial to REAL-Prompt's performance in Table 15.

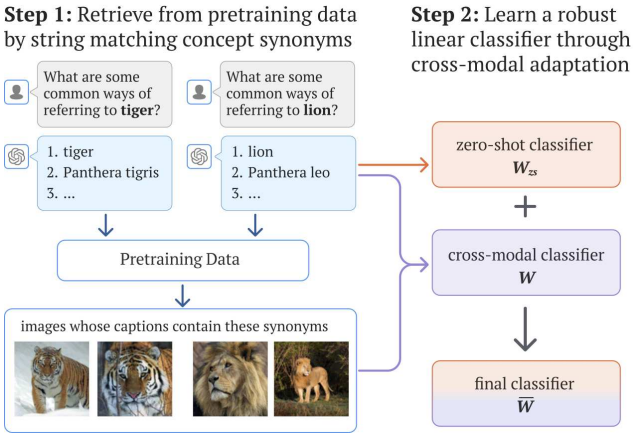

Figure 4. **Flowchart of REAL-Linear.** First, it uses all synonyms of the given concepts to retrieve a class-balanced subset of pretraining images (e.g., 500 images per class from the dataset LAION-400M). Next, it learns a linear classifier W atop the frozen VLM using cross-modal adaptation [22], and then ensembles it with the off-the-shelf classifier W_{zs}, whose weights are text prompt embeddings based on the most frequent synonyms.

4.2. REAL-Linear for Linear Classifier Training

To further improve performance, REAL-Linear finetunes on images retrieved from pretraining data that are relevant to target concepts, as illustrated in Fig. 4.

Step 1: Retrieving data using concept synonyms. For each concept, we retrieve pretraining data (LAION) whose textual captions contain any concept synonyms from REAL-Prompt. Then, we sort the retrieved data by the cosine similarity between their captions and the averaged text features (generated by OpenCLIP's text encoder using all concept synonyms). We select an equal number of the top-ranked images per concept (e.g., 500) to ensure a class-balanced dataset.

Step 2: Training a robust linear classifier. To address the potential domain gap between the pretraining data and the downstream task, we construct a robust linear classifier W using cross-modal adaptation [22]. Concretely, we learn a linear classifier W atop VLM's embeddings of the retrieved images and concept names, and then ensemble it with the zero-shot classifier W_{zs} (i.e., REAL-Prompt): $\bar{W} = W + W_{zs}$.

REAL-Linear's exceptional efficiency. REAL-Linear is significantly more efficient than the state-of-the-art REACT [24] and can be done using only academic-scale computing resources. Unlike REACT, which requires downloading the whole pretraining dataset and running VLMs to extract features for all pretraining images and texts, REAL-Linear processes only pretraining texts via string matching. As a result, it can process all LAION-400M captions in one hour, whereas REACT needs 250 GPU hours just for feature extraction. In addition, REACT's contrastive training

Table 1. **REAL outperforms existing methods on standard zero-shot recognition datasets.** Within the zero-shot prompting paradigm, our REAL-Prompt that prompts with the *most frequent synonyms* of visual concepts (using OpenAI's templates [35]) outperforms existing prompting approaches that adopt the original concept names, such as DCLIP [28] and CuPL [34]. Within the retrieval-augmented learning paradigm (without using any data from downstream tasks), our REAL-Linear retrieves a class-balanced subset of pretraining data (500 examples per concept from LAION-400M), learns a linear classifier ensembled with the zero-shot classifier used by REAL-Prompt. REAL-Linear rivals the recent method REACT [24] (which retrieves 10K examples per concept), and importantly, uses 5% of REACT's retrieved images and 1% of its compute as detailed in Table 3. We highlight the **best accuracy** in bold and underline the second best numbers.

	Method	ImageNet	Flowers	Cars	Aircraft	Pets	Food	DTD	EuroSAT	Avg
Zero-Shot Prompting	prompt template									
	"{concept}"	60.7	63.8	78.1	12.6	83.3	80.1	48.8	28.6	57.0
	"a photo of {concept}"	62.5	66.5	77.2	15.8	84.0	80.3	52.8	36.6	59.5
	OpenAI templates [35]	62.9	68.0	79.2	16.7	86.7	80.9	54.5	51.5	62.6
	DCLIP [28]	62.1	—	—	—	84.6	80.1	51.9	36.8	—
	CuPL [34]	**63.7**	65.8	80.0	17.8	87.4	79.5	59.1	—	—
	REAL-Prompt	63.6	**76.6**	**82.7**	**18.0**	**88.8**	**81.0**	**59.9**	**57.5**	**66.0**
Retrieval Augmented	REACT (10K) [24]									
	Locked-Text	65.7	73.1	**88.5**	24.5	89.2	81.8	49.8	51.1	65.5
	Gated-Image	64.2	72.3	88.1	24.8	89.5	**83.0**	51.4	45.4	64.8
	REAL-Linear (500)	**65.9**	**78.8**	84.4	**29.6**	**89.5**	81.4	**61.5**	**51.5**	**67.8**

demands extensive resources to ensure performance, e.g., large batch size (4,096) and long training (256 GPU hours on 16 V100 GPUs). In contrast, our linear-probing approach trains in minutes on a modest GPU (12GB). Table 3 compares the efficiency between REACT and REAL.

5. Experimental Results

We show the state-of-the-art performance of REAL for zero-shot recognition, outperforming existing prompting and retrieval-augmented methods across standard benchmarks. We ablate the design choices of REAL, revealing technical insights that contribute to its superior performance. We show that REAL can be combined with existing methods for even better results. Moreover, we show that REAL-Prompt improves image generation of rare concepts using text-to-image models like DALL-E 3 and SD-XL.

5.1. Experimental setup

Datasets and metric. We report mean per-class accuracy on standard classification benchmarks, including ImageNet [10], Flowers [30], Cars [18], Aircraft [26], Pets [33], Food [5], DTD [9], EuroSAT [13], and CUB [43]. Moreover, we use variants of ImageNet to study the out-of-distribution (OOD) robustness of our methods, including ImageNet-V2 [17], ImageNet-Adversarial [15], ImageNet-Rendition [14], and ImageNet-Sketch [46]. Table 6 details these datasets.

Compared methods. We compare against state-of-the-art zero-shot recognition methods for VLMs. We report various prompting strategies that directly use the given concept names, including prompt templates such as *"{concept}"*, *"a photo of {concept}"*, and **OpenAI**'s hand-engineered templates [35]. We also compare with LLM-based prompting methods, **DCLIP** [28] and **CuPL** [34], which uses GPT to generate visual descriptions for constructing prompts. Finally, we compare our method with the retrieval-augmented SOTA **REACT** [24] that retrieves data using VLMs' features and performs contrastive finetuning. We report two variants of REACT: Locked-Text and Gated-Image.

Table 2. **REAL boosts both head and tail performance.** We show that REAL-Prompt and REAL-Linear (500 retrieved examples per concept) achieve consistent improvement across all classes over the baseline using OpenAI templates [35]. On each dataset, we define the tail as the 20% least frequent classes and the rest as the head, and report the averaged per-class accuracy over nine standard zero-shot recognition datasets (including CUB [43]). We report detailed improvements on each dataset in Table 8.

	Method	ImageNet		Avg of 9 datasets	
		Head	Tail	Head	Tail
LAION 400M	OpenAI templates	64.8	55.2	65.7	52.5
	REAL-Prompt	65.4$^{+0.6}$	56.2$^{+1.0}$	67.8$^{+2.1}$	56.9$^{+4.4}$
	REAL-Linear	67.8$^{+3.0}$	58.9$^{+3.7}$	72.5$^{+6.8}$	56.0$^{+3.5}$
LAION 2B	OpenAI templates	68.0	61.0	68.6	58.4
	REAL-Prompt	68.2$^{+0.2}$	61.6$^{+0.6}$	69.8$^{+1.2}$	61.8$^{+3.4}$
	REAL-Linear	69.8$^{+1.8}$	64.8$^{+3.8}$	76.2$^{+7.6}$	63.6$^{+5.2}$

Implementation details. In this work, we ablate on a series of OpenCLIP VLMs [7, 16], which are publicly available along with their two pretraining datasets, namely LAION-400M [37] and LAION-2B [38]. We report the performance of OpenCLIP ViT-B/32 architecture in the main paper and show that REAL generalizes to other architectures in Table 13. For our REAL-Linear that learns a linear classifier, we simply use the hyperparameters provided by prior work [22, 48]. We use a single GeForce RTX 2080 Ti (12GB) to train all the models and allocate 50GB of storage to host retrieved data for all the nine benchmark datasets.

5.2. Results

Using frequent synonyms improves zero-shot recognition. Table 1 (top half) compares REAL-Prompt (based on OpenAI's prompt templates) against other prompting-based methods like DCLIP [28] and CuPL [34], which use GPT to generate visual descriptions. REAL-Prompt significantly outperforms them simply by replacing the original concept names with their most frequent synonyms. This highlights the need to reconsider concept names in prompts. REAL-Prompt is also much cheaper because it queries ChatGPT for synonyms, which are shorter than the rich visual descriptions queried by DCLIP and CuPL. For every 1K concepts, DCLIP and CuPL generate 50K ($0.5) and 500K ($5) tokens using ChatGPT, respectively, while REAL-Prompt only requires 10K tokens ($0.1).

REAL-Linear achieves the state-of-the-art. Table 1 (bottom half) compares our REAL-Linear against the retrieval-augmented SOTA REACT [24]. REAL-Linear achieves ∼3% higher accuracy averaged across eight benchmarks than REACT, while using only 500 retrieved images per concept compared to REACT's 10K. Importantly, REAL-Linear is significantly more efficient (cf. Table 3): it requires ∼1% of REACT's computes, making it more accessible to the research community. It also outperforms another recent method NeuralPriming [44] under its experimental setup (cf. Table 10).

REAL improves accuracy on tail classes. Table 2 shows that REAL boosts performance for both tail (least frequent 20%) and head (the rest 80%) classes on ImageNet and all nine datasets. For more specific improvements on each dataset, see Table 8.

REAL benefits existing zero-shot methods. Our methods can be readily applied with existing methods to further improve performance. Table 4 shows that REAL-Prompt's most frequent synonyms can be applied on any prompt templates, including LLM-enriched ones like DCLIP [28] and CuPL [34]. Likewise, REAL-Linear can be applied on top of REACT's finetuned OpenCLIP models, which achieves even better performance as shown in Table 5.

Ablation studies for REAL-Linear. To understand REAL-Linear's superior performance, we conduct several experiments, with key insights summarized below. Table 9 shows that using all concept synonyms, as opposed to just the original concept names, can help retrieve more diverse pretraining data, improving the averaged accuracy by 4%. Table 11 demonstrates that learning the linear classifier with both text and image features using cross-modal WiSE-FT [22, 48] leads to a 6.4% increase compared to linear probing with only image features. Lastly, Table 12 shows that increasing the retrieval size from 100 to 500 per concept yields a modest accuracy improvement of 0.9%. Based on this, we adopt 500 as the standard retrieval size for our experiments.

Table 3. **Compute cost comparison between REACT [24] and our REAL-Linear.** We compare the resources required for each method on the ImageNet experiment. Clearly, REAL-Linear (retrieving 500 pretraining images per concept) uses much less compute than REACT, e.g., retrieving 1000× less images, using 400× less storage and 10,000× less training time.

Stage	Resource	REACT [24]	REAL (500)	Relative Cost
Retrieval	retrieved examples	400M	0.5M	0.1%
	time	200 hrs	6 hrs	3%
	storage	10 TB	25 GB	0.25%
Learning	training images	10M	0.5M	5%
	time	256 hrs	2 mins	0.01%
	# of learned parameters	87M	0.5M	0.6%
	GPU memory	256 GB	2 GB	0.8%

Improving image synthesis using REAL-Prompt. Fig. 5 shows that, while state-of-the-art generative models such as DALL-E 3 [41] and SD-XL [2] may fail to generate correct images for some rare concepts (identified by our frequency estimation on LAION-400M), replacing the rare concepts used in prompts with their most frequent synonyms (found by REAL-Prompt) can help generate more accurate images. More qualitative examples can be found in Fig. 10 and 11.

5.3. Discussions

Broad impacts. Our work has positive societal impacts. Our concept frequency estimation method explains *why* VLMs are biased (or imbalanced) by confirming the long-tailed distribution of concepts in their pretraining data. By identifying concepts that VLMs have insufficiently learned, we can implement targeted measures to prevent unfair or biased predictions related to these concepts.

Limitations and future work. We acknowledge several limitations in our methods. First, while we offer a method to estimate concept frequency, we cannot accurately evaluate its precision and recall due to the absence of ground-truth annotations of the pretraining data. Second, our estimation method relies only on the textual captions which may overlook other visual concepts that are present in the images but not in their captions. Lastly, filtering ambiguous captions using off-the-shelf LLMs for each caption-concept pair is time-consuming. We expect the future work to address these limitations.

6. Conclusions

We investigate the critical yet ever-neglected long-tailed issues of Vision-Language Models (VLMs). We use large language models (LLMs) to estimate concept frequency in VLMs' large-scale multimodal pretraining data, uncovering long-tailed distributions for concepts used in downstream tasks. Crucially, we demonstrate a strong correlation be-

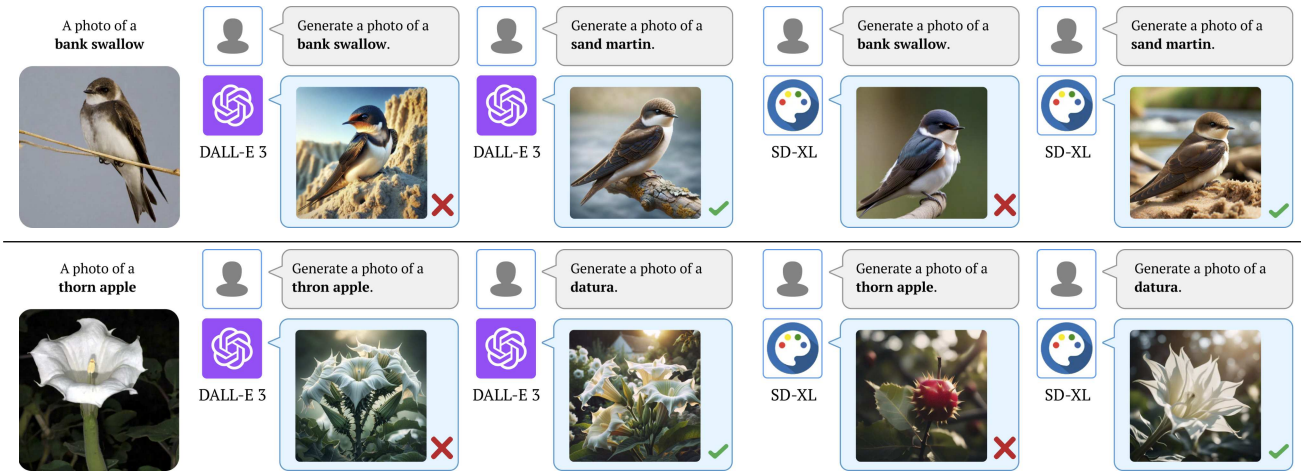

Figure 5. **Improving image generation using REAL-Prompt.** We show two rare concepts identified by our frequency estimation where DALL-E 3 and SD-XL struggle to generate accurate images: `bank swallow` (top, a bird from the CUB dataset) and `thorn apple` (bottom, a flower from the Flowers dataset). Using their original names, both DALL-E 3 and SD-XL incorrectly render the bird's colors. Additionally, DALL-E 3 erroneously adds thorns to the flower, while SD-XL depicts an apple with literal thorns. Instead, using the most frequent synonyms (`sand martin` for `bank swallow`, `datura` for `thorn apple`) as found by REAL-Prompt results in both systems generating accurate images. See more examples in Fig. 10 and 11.

Table 4. **Improvements using REAL-Prompt with existing prompting methods.** Our REAL-Prompt can be combined with existing prompt templates including OpenAI's hand-engineered templates and LLM-enriched templates [35] like DCLIP [28] and CuPL [34]. On datasets such as Flowers, DTD, and EuroSAT, integrating REAL-Prompt results in an accuracy boost of 5∼8%.

Prompting Method	ImageNet	Flowers	Cars	Aircraft	CUB	Pets	Food	DTD	EuroSAT
OpenAI templates [35]	62.9	68.0	79.2	16.7	63.8	86.7	80.9	54.5	51.5
+ REAL-Prompt	63.6$^{+0.7}$	**76.6**$^{+8.6}$	**82.7**$^{+3.5}$	18.0$^{+1.3}$	64.0$^{+0.2}$	**88.8**$^{+2.1}$	**81.0**$^{+0.1}$	**59.9**$^{+5.4}$	**57.5**$^{+6.0}$
DCLIP [28]	62.1	–	–	–	64.5	84.6	80.1	51.4	36.8
+ REAL-Prompt	62.9$^{+0.8}$	–	–	–	**64.7**$^{+0.2}$	88.1$^{+3.5}$	80.0$^{-0.1}$	55.5$^{+4.1}$	36.9$^{+0.1}$
CuPL [34]	63.7	65.8	80.0	17.8	–	87.4	79.5	59.1	–
+ REAL-Prompt	**64.2**$^{+0.5}$	72.3$^{+6.5}$	81.7$^{+1.7}$	**18.3**$^{+0.5}$	–	88.0$^{+0.6}$	79.5$^{+0.0}$	59.3$^{+0.2}$	–

Table 5. **Enhancing REACT's robustness with REAL-Linear.** Our REAL-Linear (using 500 retrieved images per concept), when applied to REACT [24]'s finetuned OpenCLIP models (ViT-B/32 trained on LAION-400M), improves zero-shot accuracy across various challenging ImageNet variants. These variants, including ImageNet-V2 [17], ImageNet-Adversarial [15], ImageNet-Rendition [14], and ImageNet-Sketch [46], are specifically designed to assess model robustness against domain shifts.

Method	ImageNet	→ ImageNet Variants			
		V2 [17]	A [15]	R [14]	S [46]
OpenAI templates [35]	62.9	55.1	21.7	73.5	49.4
REAL-Linear	65.9$^{+3.0}$	57.3$^{+2.2}$	22.7$^{+1.0}$	73.9$^{+0.4}$	50.9$^{+1.5}$
REACT Locked-Text	65.7	57.2	20.3	77.6	54.8
+ REAL-Linear	67.7$^{+2.0}$	59.1$^{+1.9}$	21.3$^{+1.0}$	78.1$^{+0.5}$	55.9$^{+1.1}$
REACT Gated-Image	64.2	56.3	21.1	75.9	52.4
+ REAL-Linear	66.9$^{+2.7}$	59.1$^{+2.8}$	21.7$^{+0.6}$	76.8$^{+0.9}$	54.2$^{+1.8}$

tween the long-tailed concept distributions and VLMs' imbalanced zero-shot performance. To address this imbalance, we propose retrieval-augmented learning (REAL), with two variants: REAL-Prompt and REAL-Linear. REAL-Prompt replaces original class names from downstream tasks with their most common synonyms found in the pretraining texts, outperforming both human-engineered and LLM-enriched prompts. On the other hand, REAL-Linear leverages concepts synonyms to fetch a balanced subset of pretraining data for training a robust linear classifier atop the frozen VLM, surpassing the previous SOTA using 400× less storage and 10,000× less training time. Finally, we highlight that modern text-to-image generators (e.g., DALL-E 3 and SD-XL) often fail to generate images for the rare concepts identified by our frequency estimation method. By applying REAL-Prompt, we demonstrate that using common synonyms helps generate more accurate images.

Acknowledgements

We thank Tiffany Ling for polishing figures. This work was partially supported by the University of Macau (SRG2023-00044-FST), and NSFC (No.62206256). Shubham Parashar acknowledges the support from the CSE Department at Texas A&M University.

References

[1] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv:2108.02818*, 2021. 2

[2] Stability AI. Stable diffusion online, 2023. 1, 2, 7, 14, 17, 18, 19, 20

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Joyce Zhuang, Juntang andLee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiang, and Aditya Ramesh. Improving image generation with better captions. *Note on Dalle-3*, 2023. 1

[4] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv:2110.01963*, 2021. 2

[5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 446–461. Springer, 2014. 6, 11

[6] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019. 3

[7] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2023. 6

[8] Ching-Yao Chuang, Jampani Varun, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint 2302.00070*, 2023. 2

[9] 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. 6, 11

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 4, 6, 11, 17, 18

[11] Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lubing Zhou, Holger Caesar, Oscar Beijbom, and Abhinav Valada. Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. *arXiv preprint arXiv:2109.03805*, 2021. 3

[12] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pretraining. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020. 2, 3

[13] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 6, 11, 13

[14] 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. 6, 8, 11

[15] 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. 6, 8, 11

[16] 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. 1, 3, 6

[17] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019. 6, 8, 11

[18] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization, 2013. 6, 11, 13

[19] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020. 2, 3

[20] Alexander Li, Ellis Brown, Alexei A Efros, and Deepak Pathak. Internet explorer: Targeted representation learning on the open web. In *ICML*, 2023. 2, 3

[21] Chunyuan Li, Haotian Liu, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, and Jianfeng Gao. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Neural Information Processing Systems*, 2022. 3, 4

[22] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19325–19337, 2023. 2, 3, 5, 6, 7, 11, 13, 15

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 2, 13, 17, 18

[24] Haotian Liu, Kilho Son, Jianwei Yang, Ce Liu, Jianfeng Gao, Yong Jae Lee, and Chunyuan Li. Learning customized visual models with retrieval-augmented knowledge. In *CVPR*, 2023. 2, 3, 5, 6, 7, 8, 13

[25] Shihong Liu, Zhiqiu Lin, Samuel Yu, Ryan Lee, Tiffany Ling, Deepak Pathak, and Deva Ramanan. Language models as black-box optimizers for vision-language models. *arXiv preprint arXiv:2309.05950*, 2023. 2

[26] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv:1306.5151*, 2013. 6, 11, 17, 18

[27] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54 (6):1–35, 2021. 2

[28] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv:2210.07183*, 2022. 2, 5, 6, 7, 8, 13, 16

[29] Sachit Menon, Ishaan Preetam Chandratreya, and Carl Vondrick. Task bias in vision-language models. *arXiv:2212.04412*, 2022. 2

[30] 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. 4, 6, 11, 17, 18

[31] OpenAI. Gpt-4 technical report, 2023. 1, 2, 3, 4, 5

[32] Shubham Parashar, Zhiqiu Lin, Yanan Li, and Shu Kong. Prompting scientific names for zero-shot species recognition. In *EMNLP*, 2023. 2

[33] 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. 4, 6, 11

[34] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701, 2023. 2, 5, 6, 7, 8, 13, 16

[35] 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 *ICML*, 2021. 1, 2, 4, 5, 6, 8, 11, 12, 13, 15, 16

[36] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1

[37] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv:2111.02114*, 2021. 1, 2, 3, 6, 12, 13

[38] 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, 3, 6, 12, 14

[39] Jie-Jing Shao, Jiang-Xin Shi, Xiao-Wen Yang, Lan-Zhe Guo, and Yu-Feng Li. Investigating the limitation of clip models: The worst-performing categories. *arXiv:2310.03324*, 2023. 2

[40] Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Zhe Gan, Lijuan Wang, Lu Yuan, Ce Liu, et al. K-lite: Learning transferable visual models with external knowledge. *Advances in Neural Information Processing Systems*, 35:15558–15573, 2022. 2

[41] Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. Improving image captioning with better use of captions. *arXiv:2006.11807*, 2020. 1, 2, 4, 7, 14, 17, 18, 19, 20

[42] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, 2023. 3

[43] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset, 2011. 6, 11

[44] Matthew Wallingford, Vivek Ramanujan, Alex Fang, Aditya Kusupati, Roozbeh Mottaghi, Aniruddha Kembhavi, Ludwig Schmidt, and Ali Farhadi. Neural priming for sample-efficient adaptation. *arXiv:2306.10191*, 2023. 2, 3, 7, 13, 14

[45] Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. Revise: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision*, 130(7):1790–1810, 2022. 2

[46] 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. 6, 8, 11

[47] Xudong Wang, Zhirong Wu, Long Lian, and Stella X Yu. Debiased learning from naturally imbalanced pseudo-labels. In *CVPR*, 2022. 2

[48] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022. 6, 7, 11, 15

[49] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv:2309.16671*, 2023. 1, 4

[50] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9, 2023. 1, 2, 13, 17, 18

[51] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2

[52] Beier Zhu, Kaihua Tang, Qianru Sun, and Hanwang Zhang. Generalized logit adjustment: Calibrating fine-tuned models by removing label bias in foundation models. In *NeurIPS*, 2023. 2