African or European Swallow? Benchmarking Large Vision-Language Models for Fine-Grained Object Classification

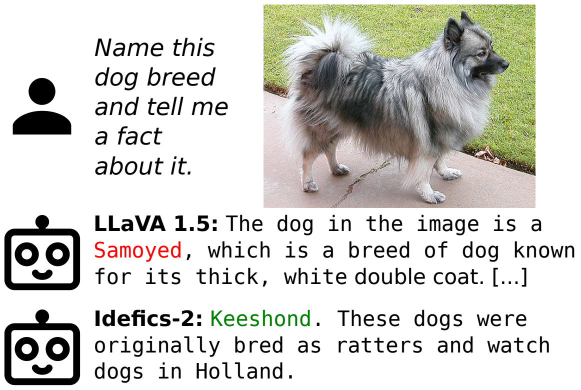
Anonymous ACL submission

Abstract

Recent Large Vision-Language Models (LVLMs) demonstrate impressive abilities on numerous image understanding and reasoning tasks. The task of fine-grained object classification (e.g., distinction between *animal species*), however, has been probed insufficiently, despite its downstream importance. We fill this evaluation gap by creating `FOCI` (**F**ine-grained **O**bject **C**lass**I**fication), a difficult multiple-choice benchmark for fine-grained object classification, from existing object classification datasets: (1) multiple-choice avoids ambiguous answers associated with casting classification as open-ended QA task; (2) we retain classification difficulty by mining negative labels with a CLIP model. `FOCI` complements five popular classification datasets with four domain-specific subsets from ImageNet-21k. We benchmark 12 public LVLMs on `FOCI` and show that it tests for a *complementary skill* to established image understanding and reasoning benchmarks. Crucially, CLIP models exhibit dramatically better performance than LVLMs. Since the image encoders of LVLMs come from these CLIP models, this points to inadequate alignment for fine-grained object distinction between the encoder and the LLM and warrants (pre)training data with more fine-grained annotation. We release our code at `ANONYMIZED`.

1 Introduction

Large Vision Language Models (LVLMs)—Large Language Model (LLM) that have been adapted to process images as input alongside text—have shown impressive performance on a wide range vision-language tasks (Li et al., 2023b; Liu et al., 2023b; OpenAI, 2023a; Anil et al., 2023). LVLMs are mutually compared using a range of benchmarks that test for various image understanding and reasoning skills, such as existence and counting of objects, localization, comparison between objects or identifying object attributes (Goyal et al., 2017; Hudson and Manning, 2019; Liu et al., 2023c).



Name this dog breed and tell me a fact about it.

LLaVA 1.5: `The dog in the image is a Samoyed, which is a breed of dog known for its thick, white double coat. [...]`

Idefics-2: `Keeshond. These dogs were originally bred as ratters and watch dogs in Holland.`

Figure 1: The importance of object recognition: LLaVA 1.5 fails to identify the dog breed. Idefics-2 correctly recognizes it and gives a correct fact as a result.

LVLMs are, however, barely ever tested for fine-grained object classification—the ability to correctly recognize different animals, plants, or man-made objects—which is, we argue, an important skill that complements general image understanding.[1] Besides it being an end-task in itself, e.g., to answer questions such as "*What is this flower called?*", it is often implicitly needed in information-seeking situations, where the success depends on the models' ability to correctly and precisely identify an object (Hu et al., 2023; Chen et al., 2023; Mensink et al., 2023). As illustrated in Figure 1, only one of the LVLMs correctly identifies the dog breed in the image and can follow up with relevant information. Note that this is different from general L(V)LM hallucination (Zhang et al., 2023b), where models 'invent' incorrect information. Instead, the generated content is correct for the object, but the object is misclassified: the information about *Samoyed* by LLaVa 1.5 is correct, but the dog in the image is a *Keeshond*.

To fill this gap in LVLM evaluation, we create

[1] For simplicity, we use 'object' to refer to both living entities like animals as well as to inanimate objects.

a comprehensive benchmark dubbed FOCI (**F**ine-grained **O**bject **C**lass**I**fication) that tests models' fine-grained object recognition over a wide range of object categories. Our key contribution is a well-defined task formulation that avoids pitfalls of prior work: We argue that an open question answering (QA) formulation (i.e., answer the question "What is this"?), as done, e.g., by Xu et al. (2023a), is an ill-defined task for two reasons. **1)** the *complete* set of admissible answers is *not* provided (e.g., admissible answers for the dog in Figure 1 include *Keeshond*, *Dutch Barge Dog*, and *Wolfspitz*). For objects with only a few synonym labels, one can provide all answer options but this does not scale to hundreds or thousands of objects. Constrained decoding to only the admissible labels is computationally expensive for large label sets (Chen et al., 2022). **2)** The expected taxonomy level of the answer is not specified. For the given example, *dog*, *Spitz*, and *Keeshond* are all ontologically correct answers; but recognizing a *Keeshond* is much more difficult than recognizing a *dog*. To address the above shortcomings, we formulate object classification as a multiple-choice problem To avoid that the reduction to only a handful candidate answers renders the task trivial, we use a CLIP model (Radford et al., 2021a) in a zero-shot configuration to mine difficult choices from the pool of class labels. We assemble FOCI from 5 popular classification datasets for different domains (flowers, cars, food, aircraft, pets) and additionally create 4 domain subsets from ImageNet-21k (Deng et al., 2009) for *animals*, *plants*, *food*, and *man-made* objects.

We extensively evaluate 12 publicly available LVLMs on FOCI and find that many of them like the popular LLaVA 1.5 struggle with fine-grained object classification. We observe that models with similar performance on established benchmarks can yield quite different and uncorrelated results on FOCI, highlighting that fine-grained object classification is indeed a *distinct skill* for LVLMs, and that FOCI should thus complement existing image understanding and reasoning benchmarks. Comparing the models further, we observe that the scale of their (pre-)training data seems to impact their performance on FOCI significantly more than for image understanding tasks. A comparison with the underlying CLIP models used as the LVLMs' image encoders shows that the encoder's zero-shot accuracy provides an upper bound for the LVLM, with the LVLM performance lagging drastically behind. This suggests that the alignment between the image encoder and LLM in LVLMs seems to be insufficiently semantically fine-grained. We finally perform controlled experiments to isolate the modeling and training decisions that impact the models' performance in FOCI. As is the case with other benchmarks, both larger LLMs and stronger image encoders improve results. Most importantly, incorporating captions into the training data that explicitly name the downstream objects helps with classification. Similarly, including fine-grained classification objectives to the training mix can improve models' FOCI performance.

2 Related Work

Large Vision-Language Models. LVLMs align pre-trained image encoders (generally a Vision Transformer (ViT) (Dosovitskiy et al., 2021) from CLIP (Radford et al., 2021a)) to a Large Language Model (LLM), yielding an LLM that can work with images as input besides text (Chen et al., 2022; Alayrac et al., 2022; Li et al., 2023b; Dai et al., 2023; Liu et al., 2023b,a; Bai et al., 2023; Laurençon et al., 2023; Chu et al., 2023; Zhang et al., 2023a). LVLMs are commonly trained in two stages: first, an alignment module between the image encoder and the LLM—a shallow feed-forward network (Liu et al., 2023b,a) or more complex modules like a resampler (Alayrac et al., 2022; Li et al., 2023b)—that projects image tokens into the LLM input embedding space is trained using image-caption pairs. In the second stage, the model is trained for general-purpose inference on a mix of tasks, e.g., visual Q&A (Goyal et al., 2017; Hudson and Manning, 2019) and (visual) chat instruction data (Chiang et al., 2023; Liu et al., 2023b). While the second stage is fairly similar across the recent models, the first stage is where training greatly varies: on the low end, models are trained with less than a million examples (Liu et al., 2023a, 2024); on the high end, over a billion image-text pairs are used (Bai et al., 2023; Dong et al., 2024; Laurençon et al., 2024). Despite differences in data size, models on both ends of the spectrum can achieve competitive results on popular benchmarks. In this work, we show that better visio-linguistic alignment in the first training stage substantially boosts fine-grained object classification abilities.

Benchmarking LVLMs. Most existing benchmarks, e.g., VQAv2 (Goyal et al., 2017), GQA (Hudson and Manning, 2019), MME (Fu et al.,

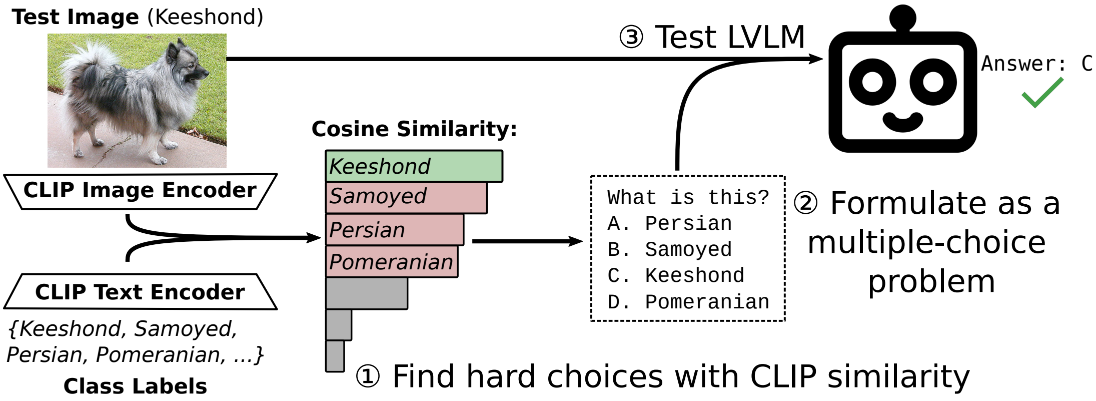


Figure 2: Testing LVLMs on object classification through multiple-choice: (1) We compute the CLIP cosine similarity between a test image and class labels; we select the correct label and the three most similar (wrong) labels to (2) formulate a multiple-choice problem, which (3) is given to the LVLM who has to predict the correct choice.

2023), MMBench (Liu et al., 2023c), or Seed-Bench (Li et al., 2023a), test LVLMs for image understanding and reasoning capabilities such as recognition of color and other attributes, object counting, recognizing object position and orientation and similar. Other benchmarks like MMMU (Yue et al., 2023) test world knowledge and reasoning capabilities in different domains. Although (fine-grained) object classification is a prominent end-task in itself and relevant in conversational applications, it is barely considered in LVLM evaluation protocols. The work that addresses the task is limited. *(i)* Models with in-context learning capabilities are evaluated on few-shot object classification but the models do not classify images in isolation and instead compare the target image with labeled in-context examples (Tsimpoukelli et al., 2021; Alayrac et al., 2022). *(ii)* Pali (Chen et al., 2022) was evaluated on ImageNet (Deng et al., 2009) by scoring every class labels, which is computationally expensive. *(iii)* LVLM-e-Hub (Xu et al., 2023a) includes some image classification datasets (like ImageNet) but they formulate it as open-ended QA task with ambiguity over expected answers, which leads to low accuracy scores for all models. *(iv)* In knowledge-intensive VQA, models have to recognize the correct object (e.g., a specific building) to answer correctly; objects are recognized either implicitly (the QA model needs to know which object it is to answer correctly) or explicitly when a knowledge base is used to retrieve relevant information (Hu et al., 2023; Chen et al., 2023; Mensink et al., 2023). In contrast to these efforts, we propose a standardized evaluation of LVLMs for (fine-grained) object classification by converting image classification datasets into difficult multi-choice tasks.

3 Multiple-Choice Image Classification

Image classification is a fundamental problem in computer vision with a plethora of datasets available. In this work, we focus on *fine-grained object classification* where models have to differentiate between several objects belonging to a specific domain, e.g., animal species or car models. We leverage existing datasets as resources for annotated data and frame object classification as a multiple-choice task with well-defined answer candidates.

Why Multiple-Choice? The standard formulation of object classification tasks for LVLMs is via question answering, with open-ended answer generation Xu et al. (2023b). This formulation, we argue, represents an ill-posed problem for two main reasons: (1) the expected level of granularity in the object taxonomy that is expected as the answer is not defined, and is difficult to define in general (e.g., for the image from Figure 1, *dog*, *Spitz*, or *Keeshond* are all correct labels); (2) the set of admissible answers in existing datasets is *not complete*: most objects have multiple synonymous labels, all of which constitute a correct answer (e.g., *Keeshond*, *Dutch Barge Dog*, and *Wolfspitz*), but only subsets of those are provided as admissible labeles in existing datasets. Providing complete synonym sets and specifying the expected level of granularity of the answer is, in the general case, infeasible. Instead, we propose to formulate fine-grained object classification as a multi-choice task, where the models are provided with a set of candidate answers from which the correct answer is to be selected; this way the expected (i.e., correct) output is well-defined.

Mining Hard Choices. To maintain difficulty despite the reduction to only a small set of candidate labels, we mine for each example image *difficult* incorrect labels from all class labels used in the concrete image classification dataset. We argue that a reduction to the most likely incorrect classes retains the task difficulty as even in classification over large class sets (e.g., thousands of classes), models easily discern between unrelated classes and most errors stem from close classes anyways (e.g., in the Oxford-Pets dataset, which covers 37 cat and dog breeds, cat breeds are irrelevant for dog images). We use a CLIP model for mining difficult candidates: for every example image, we select the three most similar (incorrect) class labels as negative choices. We rank the dataset classes for an image using the standard CLIP zero-shot setup: the text encoder embeds all class labels, the image encoder embeds the image, and the class labels are ranked in decreasing order of cosine similarity of their respective text embeddings with the image embedding. We avoid biasing the choice selection towards any concrete LVLM in our evaluation by selecting `OpenCLIP ViT-L/14` (Ilharco et al., 2021): its image encoder has not been used by any of the LVLMs. Figure 2 illustrates both the process of mining negatives for an image and testing an LVLM on the resulting set of candidate choices.

`FOCI` (Fine-grained Object ClassIfication). We collate our `FOCI` benchmark from diverse existing datasets, selecting in all cases four candidate choices for each image (i.e., the correct label and three most similar negatives). We complement (1) established datasets commonly used for evaluating CLIP models (Radford et al., 2021a; Ilharco et al., 2021) with (2) additional challenging larger-scale datasets that we derive from ImageNet-21k (Deng et al., 2009). For the former, we select the following five datasets: **FGVC-Aircraft** (Maji et al., 2013) contains images of 100 different aircraft types; **Flowers102** (Nilsback and Zisserman, 2008) contains images of 102 different flower species; **Food101** (Bossard et al., 2014) covers 101 dishes; **Oxford-Pet** (Parkhi et al., 2012) contains images of 37 cat and dog breeds. **Stanford-Cars** (Krause et al., 2013) covers 196 car models.

As some of the above datasets are not particularly challenging for existing CLIP models in zero-shot evaluations, we additionally construct four new challenging datasets from ImageNet-21k (**IN-21k**). We first merge ImageNet-COG (Sariyildiz et al., 2021) (5k classes) and ImageNet-1k (**IN-1k**), for a total of 6k classes that are all leaf nodes in the WordNet (Miller, 1994) taxonomy: this means that no two labels stand in the *is-a* relation and there cannot be multiple correct answers stemming from different taxonomy levels (e.g., *dog* and *Pomeranian*). Next, we group the classes according to their WordNet lexicographer file names, and create a dataset for each of the four most represented ones: **Animal** (1322 classes), **Plant** (957 classes), **Food** (563 classes), and **Artifact** (man-made objects, 2631 classes). We prepend **IN-** (ImageNet-) in our experiment to mark these datasets.

Model	**#P**	**Pretrain**	**Task Mix**
Idefics-1 (Laurençon et al., 2023)	9B	350M	1M
Idefics-2 (Laurençon et al., 2024)	8B	1.5B	?
BLIP2 Flan-T5-XL (Li et al., 2023b)	4B	130M	—
InstructBLIP Flan-T5-XL (Dai et al., 2023)	4B	130M	1M
InstructBLIP Vicuna (Dai et al., 2023)	8B	130M	1M
InternLM XComposer 2 (Dong et al., 2024)	7B	>1B	600M
LLaVA 1.5 (Liu et al., 2023a)	7B	560k	660k
LLaVA-Next (Mistral) (Liu et al., 2024)	7B	560k	760k
MobileVLM V2 (Chu et al., 2024)	7B	1.2M	2.4M
Pali-Gemma[1]	3B	>1M	?
Phi-3-Vision (Abdin et al., 2024)	4B	>10M	>1M
Qwen-VL-Chat (Bai et al., 2023)	10B	1.4B	50M

Table 1: The 12 tested public LVLMs. We provide parameters count (#P; LLM + image encoder parameters) and the dataset size (in images) used during the pretraining and task mix training phase. For some fields, we put a conservative estimate or '?' if no estimate is possible. [1]Model Card, tech report pending at time of writing.

One could, in principle, add more object types and domains to the evaluation: our goal was to include a reasonably diverse set of domains, from which, when put together in a benchmark, one could reliably extrapolate general fine-grained object recognition abilities of LVLMs. For further analysis, in Appendix D we additionally evaluate LVLMs on more general (i.e., not domain-specific) object classification under different image distribution shifts (using ImageNet-1k) and for geographic distribution shifts with common objects photographed in different regions of the world, using GeoDE (Ramaswamy et al., 2023).

4 Evaluating Public LVLMs

We evaluate 12 diverse and publicly available LVLMs on `FOCI`. We then analyze how the performance of LVLMs relates to the results of their underlying CLIP image encoders.

Model and Inference Details.. Our selected models span a variety of architectures and training paradigms. Table 1 summarizes key information (the number of parameters and the size of the training data) for each model. Due to our hardware constraints, we benchmark models with LLMs having ≤7B parameters. At inference time, we provide the LVLMs with the image and the four candidate choices. The choices are in random order to avoid model-specific preferences for answer positions (Liu et al., 2023c)); the model provides as output one of the choices, which is compared with the ground truth label: we then report the performance in terms of accuracy. See Appendix A for further details on models, the inference setup, and datasets.

4.1 Results

FOCI vs. Other Benchmarks. Table 2 displays the results for the 12 benchmarked LVLMs on FOCI. We first compare the models' performance and relative ranking on FOCI with their results on popular image understanding benchmarks (we show the models' performance on GQA (Hudson and Manning, 2019), MMBench (Liu et al., 2023c), and MMMU (Yue et al., 2023) in Table 5 in the Appendix C). Model's results on FOCI are much less correlated with their respective results on other benchmarks: better results on GQA, MMBench, or MMMU do not necessarily imply better results for fine-grained object classification and vice versa. Qwen-VL, for example, is amongst the best-performing models in object classification in FOCI, but is fares much worse on the standard benchmarks, where several yield better results. On the other hand, Phi-3-Vision has among the best results on the standard benchmarks but exhibits only average performance on FOCI. These results indicate that fine-grained object classification is a skill that is complementary to what other image understanding and reasoning benchmarks test and as such should be added to LVLM evaluation protocols.

Training Data. One important factor for strong object recognition on FOCI seems to be the amount of image-text data used for (pre-)training the alignment component of the LVLM in the first training phase (see §2). On the common understanding benchmarks, models like LLaVA 1.5, and LLaVA-Next show strong results despite being pretrained with <1M image-text pairs. However, the two best models on FOCI, Idefics-2 and Qwen-VL, are both pretrained on ∼1.5B images and drastically outperform the LLaVA models. This suggests that object classification requires larger-scale training for a much more fine-grained alignment between the image encoder and LLM, compared to what is needed in general for image understanding. We isolate the effect of the alignment training data (in a smaller-scale setup) in §5. The results for InstructBLIP are somewhat inconclusive: with Flan-T5-XL as LLM, it exhibits good FOCI performance, but with Vicuna (and otherwise identical training) the results are substantially worse. This would suggest that, other than the scale of the alignment training, the LLM itself plays an important role.

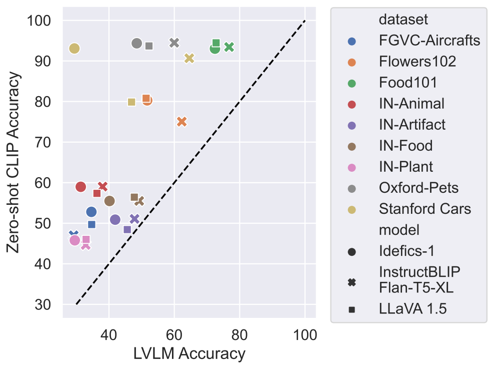


Figure 3: We plot the LVLM accuracy against the CLIP zero-shot accuracy (using the 4 multiple-choice options for CLIP for a fair comparison) of the underlying CLIP image encoder used by the LVLM.

Other Factors. Very high image resolution, which is highly beneficial for OCR-heavy tasks like chart understanding (Liu et al., 2024), does not seem to be relevant for fine-grained object classification. This stems from the comparison between LLaVA 1.5 and LLaVA-Next, where the latter's main difference w.r.t. the former is training with (and inference on) images of higher resolution. This is unsurprising as images in object classification datasets typically contain large centered objects, making larger resolution unnecessary for solving the task. The LLM and image encoder are likely also major factors for the ultimate performance but we cannot isolate them in this observational analysis; instead, we consider them in controlled experiments in §5.

4.2 LVLM vs. Its Corresponding CLIP

Several of the tested LVLMs keep their underlying CLIP image encoder frozen throughout training. This means that the cross-modal alignment between the CLIP's image encoder and its text encoder is untouched, allowing us to compare the performance of these LVLMs directly against the CLIP models from which they take the image encoder.

Specifically, we consider three LVLMs with their corresponding CLIP models: Idefics-1, which uses OpenCLIP ViT-H/14 (Ilharco et al., 2021),

Model	IN-Food	IN-Artifact	IN-Animal	IN-Plant	Aircraft	Flowers102	Food101	O.-Pet	S.-Cars	∅
Idefics-1	40.18	41.90	31.37	29.55	34.62	51.70	72.44	48.51	29.42	42.19
Idefics-2	**56.38**	**52.56**	46.50	**41.47**	**56.23**	72.78	**89.70**	81.28	**80.25**	**64.13**
BLIP-2 Flan-T5-XL	51.47	47.41	39.22	32.59	32.94	64.32	82.51	65.00	67.68	53.68
InstructBLIP Flan-T5-XL	49.25	47.83	38.07	32.88	29.19	62.29	76.77	59.99	64.58	51.21
InstructBLIP Vicuna	43.94	42.39	37.32	30.04	31.68	50.90	63.47	54.92	48.25	44.77
InternLM XComposer 2	50.43	47.84	38.98	33.23	40.53	54.25	79.30	63.23	53.89	51.30
LLaVA 1.5	47.76	45.61	36.32	33.00	34.71	51.37	72.80	52.25	46.92	46.75
LLaVA-Next	46.32	45.54	35.51	31.86	32.49	43.91	71.30	53.72	49.48	45.57
MobileVLM v2	46.50	44.58	37.60	33.75	35.01	54.89	74.38	53.69	46.29	47.41
Pali-Gemma	54.25	48.79	42.28	37.04	39.87	69.64	82.36	75.42	64.64	57.14
Phi-3-Vision	46.66	42.75	35.11	31.27	42.33	51.59	69.98	56.36	54.50	47.84
Qwen-VL-Chat	52.36	50.95	**48.45**	40.09	45.96	**75.95**	83.92	**87.82**	76.23	62.41

Table 2: Accuracy on FOCI: on individual datasets and average (∅), for the 12 tested public LVLMs.

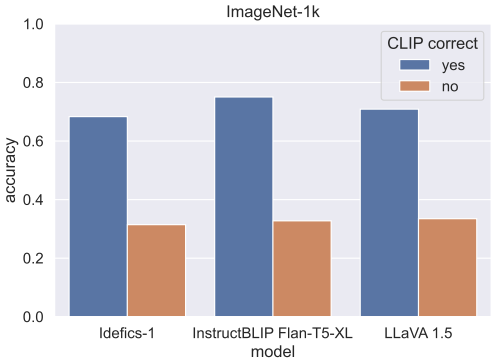


Figure 4: Accuracy of three LVLMs on ImageNet-1k, for example subsets on which the zero-shot classification with the corresponding CLIP model is (in)correct.

LLaVA 1.5, which uses OpenAI ViT-L/14 (Radford et al., 2021b), and InstructBLIP Flan-T5 with EVA-1 ViT-g/14 (Fang et al., 2022).

CLIP Zero-Shot Classification as Upper Bound. The image and text encoder of a CLIP model were trained jointly on huge datasets; in contrast, the alignment of the CLIP's image encoder to the LLM is learned with comparatively less image-text data (e.g., InstructBLIP is pre-trained with 100M samples while EVA-1 was trained with 11B samples). We compare in Figure 3 the LVLM performance against the zero-shot classification accuracy of the corresponding CLIP model (for a fair comparison, CLIP only considers the same 4 labels as LVLM does in multiple-choice formulation). We observe that the LVLM performance is indeed consistently lower than that of the corresponding CLIP model. However, while the CLIP zero-shot classification accuracy seems to be an upper bound for the LVLM, the gaps vary substantially across the FOCI datasets: from <10% on IN-Artifact to 40-50% on Oxford-Pets. These results indicate that, while the alignment between the image encoder and LLM is undertrained in general, there are also drastic differences in the quality of alignment for different types of objects (i.e., domains). For certain domains (e.g., Oxford-Pets) the LLM seems to struggle to process the image features, despite the CLIP image encoder encoding sufficient information (as evidenced by the much better corresponding CLIP performance).

CLIP wrong $\implies$ LVLM wrong? We analyze the predictions of LVLMs on instances that the corresponding CLIP model misclassifies to measure whether those classification errors propagate to the LVLM: in other words, if the CLIP model is wrong, is the LVLM using its image encoder also bound to misclassify the image? Figure 4 summarizes the results of this analysis on ImageNet-1k (in our multi-choice formulation) for three LVLMs; for the FOCI datasets we provide the same analysis in Figure 7 in the Appendix. We observe that LVLM accuracy plummets on examples on which the corresponding CLIP fails: in fact, for instances that CLIP cannot correctly classify, the performance of the corresponding LVLM gets close to random (25%) for all three LVLMs in the analysis. These observations—that CLIP performance is an upper-bound for LVLM accuracy and that its errors propagate to the LVLM—highlight that the selection of an image encoder is a key design decision for LVLMs performance and suggest that future improvements in image encoding are likely to also propagate to LVLM object recognition capabilities.

5 Controlled Experiments

We next perform a set of controlled experiments to disentangle the effects of individual LVLM design choices on (fine-grained) object classification. Our analysis encompasses three main factors: (1) the LLM size, (2) the image encoder, and (3) targeted changes to the training data. For (2) and (3), we train LVLMs following the LLaVA 1.5 recipe with StableLM 2 Zephyr 1.6B (Bellagente et al., 2024) as LLM and OpenAI CLIP-L/14-224 as the image

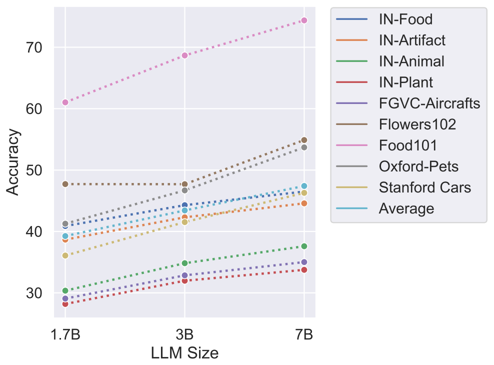


Figure 5: Results with MobileVLM v2 over its three LLM sizes with otherwise identical training.

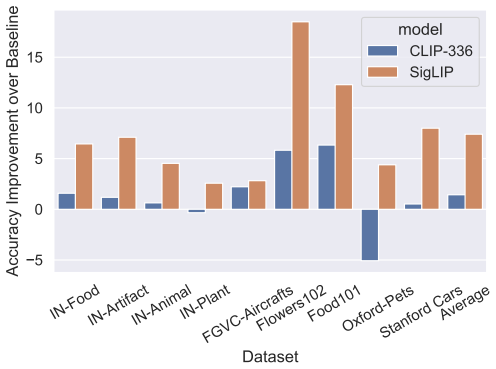


Figure 6: Improvements over our baseline when changing the `OpenAI ViT-L/14-224` image encoder to a higher resolution (336) or to `SigLIP SO400-224`.

encoder (see the Appendix B for training details).

LLM Size. Larger LLMs generally make for better LVLMs, yielding better benchmark performance due to (*inter alia*) improved reasoning capabilities (Liu et al., 2023a; Karamcheti et al., 2024; Chu et al., 2024). Our multiple-choice object classification is not difficult from a reasoning or language-understanding perspective, but it requires familiarity with thousands of objects, which may be beyond the knowledge stored in smaller LLMs. For this analysis, we turn to the MobileVLM v2 model series (Chu et al., 2024): with models trained on top of 1.7B, 3B, and 7B LLM backbones and otherwise identical architecture (image encoder and alignment module) and training procedure (data and training protocol for both the LLMs and subsequent LVLMs), we can isolate the effect of LLM size. Figure 5 summarizes the results. Expectedly, the performance on all `FOCI` datasets consistently improves with increased LLM size: we believe that this is because smaller LLMs simply encode less world knowledge and have semantically poorer representations for (fine-grained) objects.

Model	**IN-1k**	**Train Half**	**Test Half**	**`FOCI`**
`Baseline`	53.12	53.71	52.52	41.19
`No Pretrain`	51.94	51.56	52.32	38.71
`Synthetic`	54.46	55.12	53.80	41.48
`Template`	54.81	58.82	50.80	40.69
`QA Task`	57.40	59.89	54.91	43.64

Table 3: Results for experiments with changes to the training data on: ImageNet-1k overall (IN-1k) and broken down for the training half and the held-out test half, and the average results over the 9 `FOCI` datasets.

Image Encoder. Following the observation that the quality of the CLIP image encoder may cap the LVLMs' performance (Figure 3), we investigate the effect that LVLM's image encoder has on fine-grained object recognition. Our "baseline" LVLM aligns the `OpenAI CLIP-L/14-224` (CLIP-224 for short) image encoder with the LLM. We then create two other LVLMs by changing the image encoder with: (1) `OpenAI CLIP-L/14-336` (CLIP-336 for short), which takes images of *larger resolution* and (2) `SigLIP SO400M-224` (SigLIP for short) (Zhai et al., 2023) as a 'better' image encoder, boasting substantially higher benchmark results on image processing benchmarks. Figure 6 summarizes the results. On one hand, encoding images in higher resolution (with CLIP-336, i.e., increasing from 224px to 336px) leads to only a marginal ∼1 accuracy point gain, averaged over all `FOCI` datasets. The effect seems to depend on the object type: we see gains of over 5 points on Flowers102 & Food102 but also a 5-point drop on Oxford-Pets. The SigLIP encoder, on the other hand, greatly improves the baseline performance across the board. The absolute gains of the SigLIP-based LVLM over the baseline LVLM (CLIP-224 encoder) are, however, not proportionate to gains that the corresponding SigLIP CLIP model yields over CLIP-224 in zero-shot object classification. For example, while SigLIP beats CLIP-224 by 27% on FGVC-Aircraft,[2] the SigLIP-based LVLM beats the CLIP-224-based LVLM on the same dataset by only 3%; inversely, on Food101, SigLIP has only a 2% edge in CLIP comparison, but yields 12% better performance in LVLM comparison.

Training Data. The two LVLMs trained with most data, Idefics-2 and Qwen-VL (>1.5B images in total over both training stages) demonstrated the best performance on `FOCI` (Table 2). As this scale

[2] Taken from: openclip_classification_results.csv

of training is beyond the (computational) budget of most practitioners, we set to quantify the FOCI gains from adding training data at smaller data scales, concretely at the data budget of LLaVA 1.5 (ca. 1.2M images in total, see Table 1).

Changes to Pretraining. We hypothesize that a larger pretraining corpus benefits the LVLM due to having more of the objects named explicitly in the corresponding captions. We test this explicitly by replacing 25% of the LLaVA 560k pretrainng images (with captions) with images from the ImageNet-1k train split. To have a held-out control set, we only use 500 of the 1000 classes (choosing every other class) for training; we select 280 images per class (140k training examples in total). We consider three training strategies for the added ImageNet images: **i)** with *synthetic captions*, generated using BLIP (Li et al., 2022) (Synthetic); this setup tests the effect of images with objects but with captions that do not necessarily name them (e.g., for an image of a Keeshond, BLIP-generated caption will likely contain '*dog*' but not '*Keeshond*'). **ii)** with *template captions* (Template) such as *"a picture of a $label."*; such captions are not visually descriptive but explicitly name the object in the image. **iii)** we skip the pretraining phase entirely (No Pretrain) and perform the task mix training on the randomly initialized alignment module; on standard benchmarks, skipping pretraining has been reported not to notably affect performance (Karamcheti et al., 2024).

Changes to Task Mix Phase. We incorporate ImageNet as an open-ended QA Task where the model is prompted to name the image object without candidate answers. We use the open-ended QA formulation in training to avoid model adaptation to the multiple-choice formulation of the task we use at test time on FOCI. We again use 500 (out of the 1000) ImageNet classes and sample 150 examples per class (75k training examples in total). We do not otherwise change the LLaVA task mix data.

Results. We report the results of this ablation in Table 3. Skipping the pretraining step entirely (No Pretrain) reduces the average FOCI performance by over 2 accuracy points: this suggest that pretraining of the alignment module on image-text pairs is important for fine-grained object classification, unlike what was recently reported for other tasks (Karamcheti et al., 2024). Training on images with both Synthetic and Template captions has a very limited effect on FOCI performance and the unseen Test Half of ImageNet. Training on Synthetic brings a $\sim$ 1.5-point gain for the 500 ImageNet object classes seen in training (Train Half in Table 3); in comparison, the Template captions bring a much more significant gain of 5% for seen object classes: this strongly suggests that *explicitly mentioning* the objects in the captions is key for learning the alignment module that allows LVLMs better fine-grained object classification; just having images containing the object does not suffice (or is, at least, less effective). Note that only the feed-forward alignment module is trained in the first phase, so the improvements with Template captions can only be the result of having learned a better alignment and not due to the image encoder or LLM (both frozen) obtaining better representations of objects and their mentions, respectively. Including ImageNet as open-ended QA Task to the second task mix training phase has a larger effect on performance. For 500 of ImageNet-1k seen in training (Train Half), we observe a 6% improvement, but also a 2-point improvement on the images from the held-out Test Half and on FOCI.

6 Conclusion

In this work, we evaluate the capabilities of LVLMs for fine-grained object classification over different domains. We address the ambiguity of open-ended QA-based object classification evaluation and propose to replace it with a multiple-choice formulation, in which we retain the task difficulty by mining difficult (semantically closest classes) choices with a CLIP model. This way, we create FOCI, a novel benchmark consisting of 9 fine-grained multi-choice object classification datasets. We benchmark 12 public LVLMs, demonstrating that their performance on FOCI is largely uncorrelated with that on other image understanding and reasoning benchmarks: this renders fine-grained object classification a skill that is complementary to what the existing benchmarks test the LVLMs for. Our ablations identify the quality of the image encoder and the amount of explicit caption mentions of image objects in LVLM training data as factors that drive the performance. We hope our work stimulates wider research efforts on improving LVLMs for fine-grained object classification, in particular conceptual innovation (e.g., more effective training data and protocols for object classification with LVLMs) that goes well beyond mere scaling of LVLM pretraining to billions of image-text pairs.

Limitations

We identify three main limitations for our work:

First, while the goal of this work is not to evaluate every possible domain, we still likely exhibit a bias towards Anglospheric concepts as multiple datasets were created at British and US universities and use images sourced from the English internet. ImageNet in particular shows such biases (Liu et al., 2021) in image source and for its classes. While we briefly consider performance over geographic distribution shifts in the Appendix, we still likely overestimate performance for diverse cultural objects and concepts from around the globe.

Another limitation stems from the multiple-choice formulation: while it allows for well-defined answers, users 'in the wild' are more likely to use an open-ended formulation. While we expect results between the two formulations to correlate, some objects may be harder to classify in a multiple-choice setup due to the presence of challenging confounder options, and vice versa, some objects may be easier to classify in multiple-choice with the correct name as an option.

Finally, we only evaluate public LVLMs using LLMs of 7B parameters or less. We do not consider larger models (e.g., LLaVA 1.5 with Vicuna-13B) or proprietary LVLMs (e.g., GPT4 (OpenAI, 2023b) or Gemini (Anil et al., 2023)) because the inference time is too high on our compute (or not possible at all VRAM-wise) for the former and too expensive with >100,000 of API calls for the latter.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. _eprint: 2404.14219.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. *CoRR*, abs/2204.14198. ArXiv: 2204.14198.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *CoRR*, abs/2312.11805. ArXiv: 2312.11805.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities. *CoRR*, abs/2308.12966. ArXiv: 2308.12966.

Marco Bellagente, Jonathan Tow, Dakota Mahan, Duy Phung, Maksym Zhuravinskyi, Reshinth Adithyan, James Baicoianu, Ben Brooks, Nathan Cooper, Ashish Datta, Meng Lee, Emad Mostaque, Michael Pieler, Nikhil Pinnaparaju, Paulo Rocha, Harry Saini, Hannah Teufel, Niccoló Zanichelli, and Carlos Riquelme. 2024. Stable LM 2 1.6B Technical Report. *CoRR*, abs/2402.17834. ArXiv: 2402.17834.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. 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*, volume 8694 of *Lecture Notes in Computer Science*, pages 446–461. Springer.

Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2022. PaLI: A Jointly-Scaled Multilingual Language-Image Model. *CoRR*, abs/2209.06794. ArXiv: 2209.06794.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 14948–14968. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, and Chunhua Shen. 2023. MobileVLM : A Fast, Strong and Open Vision Language Assistant for Mobile Devices. *CoRR*, abs/2312.16886. ArXiv: 2312.16886.

Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, and Chunhua Shen. 2024. MobileVLM V2: Faster and Stronger Baseline for Vision Language Model. *CoRR*, abs/2402.03766. ArXiv: 2402.03766.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *CoRR*, abs/2305.06500. ArXiv: 2305.06500.

J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. InternLM-XComposer2: Mastering Free-form Text-Image Composition and Comprehension in Vision-Language Large Model. *CoRR*, abs/2401.16420. ArXiv: 2401.16420.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2022. EVA: Exploring the Limits of Masked Visual Representation Learning at Scale. *CoRR*, abs/2211.07636. ArXiv: 2211.07636.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *CoRR*, abs/2306.13394. ArXiv: 2306.13394.

Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. 2021a. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 8320–8329. IEEE.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021b. Natural Adversarial Examples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 15262–15271. Computer Vision Foundation / IEEE.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. 2023. Open-domain Visual Entity Recognition: Towards Recognizing Millions of Wikipedia Entities. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 12031–12041. IEEE.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE.

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. 2021. OpenCLIP.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. 2024. Prismatic VLMs: Investigating the Design Space of Visually-Conditioned Language Models. *CoRR*, abs/2402.07865. ArXiv: 2402.07865.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3D Object Representations for Fine-Grained Categorization. In *2013 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013*, pages 554–561. IEEE Computer Society.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. OBELISC: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. *CoRR*, abs/2306.16527. ArXiv: 2306.16527.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? _eprint: 2405.02246.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension. *CoRR*, abs/2307.16125. ArXiv: 2307.16125.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023b. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *CoRR*, abs/2301.12597. ArXiv: 2301.12597.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually Grounded Reasoning across Languages and Cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 10467–10485. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved Baselines with Visual Instruction Tuning. *CoRR*, abs/2310.03744. ArXiv: 2310.03744.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual Instruction Tuning. *CoRR*, abs/2304.08485. ArXiv: 2304.08485.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023c. MMBench: Is Your Multi-modal Model an All-around Player? *CoRR*, abs/2307.06281. ArXiv: 2307.06281.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. 2013. Fine-Grained Visual Classification of Aircraft. *CoRR*, abs/1306.5151. ArXiv: 1306.5151.

Thomas Mensink, Jasper R. R. Uijlings, Lluís Castrejón, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araújo, and Vittorio Ferrari. 2023. Encyclopedic VQA: Visual questions about detailed properties of fine-grained categories. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3090–3101. IEEE.

George A. Miller. 1994. WordNet: A Lexical Database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated Flower Classification over a Large Number of Classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing, ICVGIP 2008, Bhubaneswar, India, 16-19 December 2008*, pages 722–729. IEEE Computer Society.

OpenAI. 2023a. GPT-4 Technical Report. *CoRR*, abs/2303.08774. ArXiv: 2303.08774.

OpenAI. 2023b. GPT-4 Technical Report. *CoRR*, abs/2303.08774. ArXiv: 2303.08774.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. 2012. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern*

Recognition, Providence, RI, USA, June 16-21, 2012, pages 3498–3505. IEEE Computer Society.

Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Peter Steiner, Xiaohua Zhai, and Ibrahim Alabdulmohsin. 2024. No filter: Cultural and socioeconomic diversity in contrastive vision-language models.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021a. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint*, abs/2103.00020. _eprint: 2103.00020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021b. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Vikram V. Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. 2023. GeoDE: a Geographically Diverse Evaluation Dataset for Object Recognition. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Mert Bülent Sariyildiz, Yannis Kalantidis, Diane Larlus, and Karteek Alahari. 2021. Concept Generalization in Visual Representation Learning. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9609–9619. IEEE.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal Few-Shot Learning with Frozen Language Models. *arXiv:2106.13884 [cs]*. ArXiv: 2106.13884.

Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. 2019. Learning Robust Global Representations by Penalizing Local Predictive Power. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10506–10518.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023a. LVLM-eHub: A Comprehensive Evaluation Benchmark for Large Vision-Language Models. *CoRR*, abs/2306.09265. ArXiv: 2306.09265.

Model	**Checkpoint**
Idefics-1 (Laurençon et al., 2023)	HuggingFaceM4/idefics-9b-instruct
Idefics-2 (Laurençon et al., 2024)	HuggingFaceM4/idefics2-8b
BLIP2 Flan-T5-XL (Li et al., 2023b)	Salesforce/blip2-flan-t5-xl
InstructBLIP Flan-T5-XL (Dai et al., 2023)	Salesforce/instructblip-flan-t5-xl
InstructBLIP Vicuna (Dai et al., 2023)	Salesforce/instructblip-vicuna-7b
InternLM XComposer 2 (Dong et al., 2024)	internlm/internlm-xcomposer2-vl-7b
LLaVA 1.5 (Liu et al., 2023a)	llava-hf/llava-1.5-7b-hf
LLaVA-Next (Mistral) (Liu et al., 2024)	llava-hf/llava-v1.6-mistral-7b-hf
MobileVLM V2 (Chu et al., 2024)	mtgv/MobileVLM_V2-7B
Pali-Gemma [1]	google/paligemma-3b-mix-224
Phi-3-Vision (Abdin et al., 2024)	microsoft/Phi-3-vision-128k-instruct
Qwen-VL-Chat (Bai et al., 2023)	Qwen/Qwen-VL-Chat

Table 4: The tested public LVLM with the corresponding checkpoint from HuggingFace we used. [1]Model Card, tech report pending at time of writing.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023b. LVLM-eHub: A Comprehensive Evaluation Benchmark for Large Vision-Language Models. *CoRR*, abs/2306.09265. ArXiv: 2306.09265.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. *CoRR*, abs/2311.16502. ArXiv: 2311.16502.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid Loss for Language Image Pre-Training. *CoRR*, abs/2303.15343. ArXiv: 2303.15343.

Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2023a. InternLM-XComposer: A Vision-Language Large Model for Advanced Text-image Comprehension and Composition. *CoRR*, abs/2309.15112. ArXiv: 2309.15112.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023b. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *CoRR*, abs/2309.01219. ArXiv: 2309.01219.

A Evaluation Details

Models & Inference: In Table 4, we specify the exact checkpoint we used for each model. We adapt the respective official code of each model for inference. All models use greedy decoding.

We use the following prompt for all models. Depending on the task, we change the default question at the beginning to prime the model for the dataset domain:

```
Default: Which of these choices is shown
    in the image?
IN-Animal: Which of these animals is shown
    in the image?
IN-Plant: Which of these plants is shown
    in the image?
FGVC-Aircraft: Which of these aircrafts is
    shown in the image?
Flowers102: Which of these flowers is
    shown in the image?
Food101: Which of these dishes is shown
    in the image?
Oxford-Pet: Which of these pets is shown in
    the image?
Stanford-Cars: Which of these cars is shown
    in the image?
Choices:
A. $CHOICE1
B. $CHOICE2
C. $CHOICE2
D. $CHOICE3
Answer with the letter from the given
    choices directly.
```

We expect the model to answer with a letter and count the example as correct if the generated answer begins with the letter corresponding to the correct answer.

Dataset Details: In general, we evaluate on the full test split (or, if no public test split exists like with ImageNet, the validation split) of every dataset.

The datasets that we constructed from ImageNet-21k (Animal, Plant, Food, Artifact) are the exception: due to the large amount of classes, we only use 10 images per class instead of the full 50 to keep computation time manageable. In addition, we use the processed version of ImageNet-21k and not the original (>1TB large) version for disk space reasons; the processed version has all images resized to 224×224px. During creating of the four datasets, we remove all classes that have no unique label (keeping only the first occurrence of a label) to achieve a 1-to-1 mapping between classes and labels.

B Training Details

We closely follow the architecture and training protocol of LLaVA 1.5 (Liu et al., 2023a). As LLM, we use the instruction-trained StableLM 2 1.6B Zephir (Bellagente et al., 2024) (`stabilityai/stablelm-2-zephyr-1_6b`), which is a small but performant LLM. The default image encoder is `OpenAI CLIP ViT-L/14-224`. Training is done on a single NVIDIA RTX 3090 with training one model taking less than 2 days.

We train the models using AdamW optimizer (Loshchilov and Hutter, 2019) with a cosine learning rate decay schedule. For the pre-training phase, we use learning rate 1e-3, weight decay 0, and batch size 256. For the task-mix training phase, we use learning rate 2e-4, weight decay 0, and batch size 128; we do not fine-tune the full LLM but apply LoRA (Hu et al., 2022) to all weights with $r = 64$, $\alpha = 128$.

C LVLM Performance on Popular Benchmarks

We collate public results on select popular benchmarks for evaluating LVLMs (GQA (Hudson and Manning, 2019), MMBench (Liu et al., 2023c), and MMMU (Yue et al., 2023)) for the models of Table 1. Comparing these results against the performance in object classification shows that the latter is an independent skill that does not directly correlate with these benchmarks.

D Additional Evaluation on More Datasets

In this section, we consider general object classification datasets (not covering a specific domain) and consider how LVLMs handle image distribution shifts for the same object using ImageNet and its variants and GeoDE (Ramaswamy et al., 2023).

ImageNet Image Distribution Shifts. There are several datasets that collect new images for the classes of ImageNet-1k (Deng et al., 2009), or at least for a subset of them. Here, we consider **ImageNet-Adversarial** (Hendrycks et al., 2021b), which contains images for 200 classes that are difficult to correctly classify for a model trained on the ImageNet-1k training split; **ImageNet-Rendition** (Hendrycks et al., 2021a), which contains for 200 classes images of the objects where the image is painted, a plushy, origami, or other renditions; and **ImageNet-Sketch** (Wang et al., 2019), which contains black-and-white drawings for all 1000 classes.

CLIP models generally excel at transferring between the different image distributions due to their large-scale training (Radford et al., 2021b). We evaluate in Table 6 if LVLMs see similar results despite training the alignment with the image encoder on magnitudes less data and generally only

Model	GQA	MMBench	MMMU
Idefics-1	—	35.2	28.7
Idefics-2	—	76.8	43.5
BLIP2 Flan-T5-XL	*44.0	—	34.4
InstructBLIP Flan-T5-XL	*48.4	—	32.9
InstructBLIP Vicuna 7B	*49.2	38.3	—
InternLM XComposer 2	—	79.6	43.0
LLaVA 1.5 7B	62.0	64.3	—
LLaVA-Next Mistral 7B	64.8	68.7	—
MobileVLM V2 7B	62.6	69.2	—
Pali-Gemma	**65.6	—	—
Phi-3-Vision	—	80.5	40.4
Qwen-VL-Chat	57.5	60.6	35.9

Table 5: Performance on standard benchmarks for image understanding and reasoning. * unlike other models, has not included GQA in training task mix. ** with model fine-tuned on GQA, not the mix version used for testing.

with natural images. We observe that the ranking between the models is similar to our evaluation on FOCI in Table 2. The changes in accuracy from ImageNet-1k to the variants are qualitatively similar to the underlying CLIP models for the LVLMs. This suggests that other representations of objects (like sketches) are encoded similarly enough by the image encoder that the LVLM can 'recognize' without extra training on different image types.

Geographic Shifts with GeoDE. We now consider geographic distribution shift using GeoDE (Ramaswamy et al., 2023), a dataset with 40 classes for which there are images evenly distributed around the globe for six regions: Europe, Africa, Southeast Asia, West Asia, East Asia, and the Americas (which does not include here the US or Canada). Results of the tested LVLMs are reported in Table 7. While GeoDE is a generally easy dataset with high accuracy throughout, we still observe substantial differences between the regions: European images consistently enjoy the highest accuracy, all non-African regions follow close by with 0-3 points worse than Europe, and finally, the African images noticeably trail behind by 2-4 points lower accuracy compared to the overall average accuracy. This shows that geographic biases in the training data, both for the image encoder and for the LVLM (Pouget et al., 2024), result in disadvantages for large parts of the population.

E More CLIP Results

We present the results for conditional accuracy of LVLMs for all datasets in Figure 7.

F Full Experiment Results

Complementary to the Figures in the main paper, we report the raw results of MobileVLMv2 in Table 8 and for our trained models in Table 9.

models	IN-1k	IN-adversarial	IN-rendition	IN-sketch	∅
Idefics-1	60.09	50.03	72.20	50.13	58.11
Idefics-2	73.39	79.84	93.23	68.21	78.67
BLIP-2 Flan-T5-XL	66.12	67.48	90.48	64.85	72.23
InstructBLIP Flan-T5-XL	66.15	69.69	90.58	64.46	72.72
InstructBLIP Vicuna	56.27	59.75	76.82	54.84	61.92
InternLM XComposer 2	65.65	73.08	83.29	56.99	69.75
LLaVA 1.5	62.44	68.53	79.30	55.88	66.54
LLaVA-Next	60.86	67.20	78.12	53.50	64.92
MobileVLM v2	61.16	64.59	79.63	54.66	65.01
Pali-Gemma	69.56	68.45	92.15	65.55	73.93
Phi-3-Vision	61.71	56.71	79.18	56.01	63.40
Qwen-VL-Chat	71.20	70.99	90.59	67.16	74.98

Table 6: Results for ImageNet-1k and four distribution-shifted versions.

models	Europe	Africa	Southeast Asia	Americas	West Asia	East Asia	All
Idefics-1	85.48	79.85	84.65	83.61	84.06	84.22	83.56
Idefics-2	90.15	86.59	90.00	89.40	90.03	89.65	89.23
BLIP-2 Flan-T5-XL	91.24	87.49	90.91	89.64	90.45	89.32	89.79
InstructBLIP Flan-T5-XL	88.64	84.10	88.46	86.87	87.99	87.60	87.20
InstructBLIP Vicuna	76.36	70.53	75.61	75.34	74.78	76.19	74.70
InternLM XComposer 2	91.54	87.59	90.81	90.48	90.63	89.54	90.04
LLaVA 1.5	86.06	82.81	86.27	84.66	86.35	84.17	84.99
LLaVA-Next	86.75	82.90	86.55	85.35	85.46	84.65	85.24
MobileVLM v2	82.13	75.16	79.27	79.26	81.09	77.99	79.05
Pali-Gemma	90.94	87.12	90.53	90.14	90.68	90.09	89.84
Phi-3-Vision	89.76	86.49	89.44	88.75	88.46	87.61	88.39
Qwen-VL-Chat	90.94	87.31	88.86	89.24	90.79	89.32	89.34

Table 7: Result on the GeoDE dataset for each region and the overall accuracy for all examples together.

Model	IN-food	IN-artifact	IN-animal	IN-plant	Aircraft	Flowers102	Food101	O.-Pet	S.-Cars	∅
1.7B	40.89	38.67	30.36	28.19	29.07	47.72	61.03	41.26	36.08	39.25
3B	44.28	42.30	34.83	31.97	32.85	47.70	68.67	46.69	41.52	43.42
7B	46.50	44.58	37.60	33.75	35.01	54.89	74.38	53.69	46.29	47.41

Table 8: Results for the three sizes of MobileVLM v2.

models	IN-food	IN-artifact	IN-animal	IN-plant	fgvc aircraft	flowers102	food101	oxford pet	stanford cars	∅
Baseline	43.43	40.33	32.18	31.54	30.27	38.33	62.40	50.12	42.08	41.19
CLIP-336	45.01	41.51	32.81	31.20	32.49	44.15	68.73	45.05	42.61	42.62
SigLIP	49.88	47.44	36.70	34.11	33.09	56.82	74.69	54.51	50.08	48.59
No Pretrain	41.55	39.63	31.50	29.80	30.30	40.30	58.44	40.23	36.60	38.71
Synthetic	43.69	40.94	32.85	31.04	32.16	39.68	64.61	47.48	40.90	41.48
Template	44.74	39.93	32.79	31.09	30.09	38.07	62.80	46.31	40.43	40.69
QA Task	44.00	41.45	33.77	31.55	32.34	49.16	67.63	51.62	41.24	43.64

Table 9: Full results for our trained models.

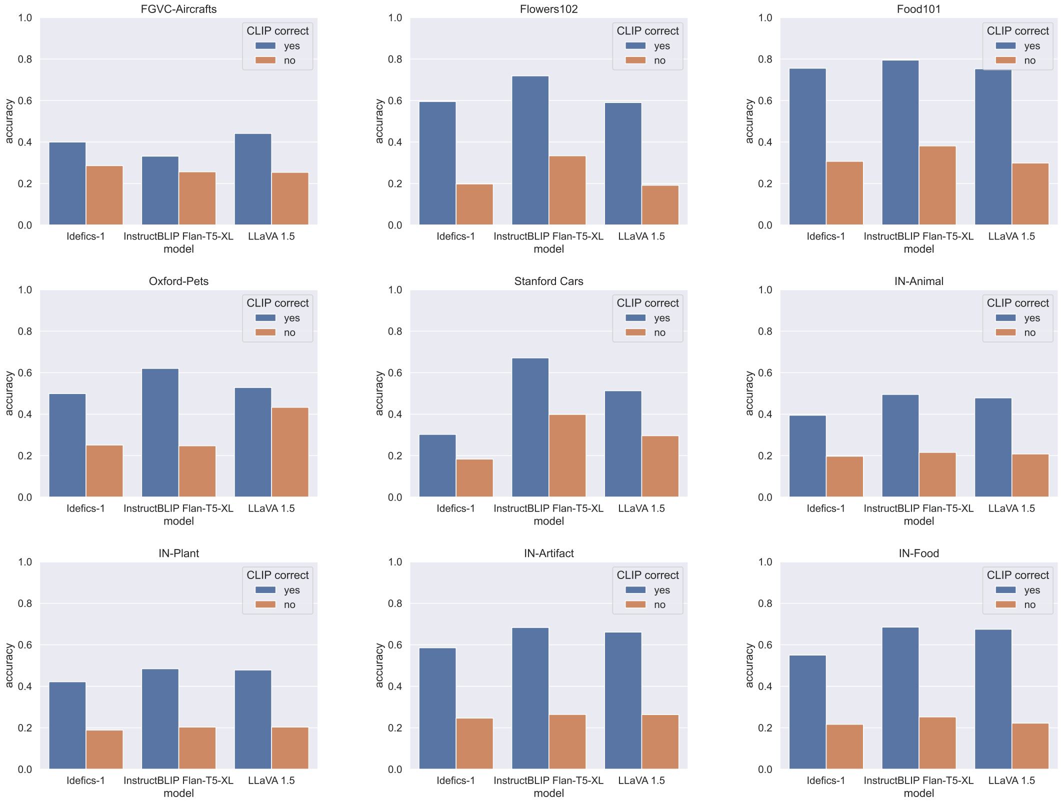


Figure 7: Conditionally accuracy on different datasets of different models if the CLIP image encoder would (in)correctly classify an example in zero-shot.