UOUO: Uncontextualized Uncommon Objects for Measuring Knowledge Horizons of Vision Language Models

Xinyu Pi*[1] Mingyuan Wu*[2] Jize Jiang*[2] Haozhen Zheng*[2]
Beitong Tian[2] Chengxiang Zhai[2] Klara Nahrstedt[2] Zhiting Hu[1]
[1]University of California San Diego [2]University of Illinois Urbana-Champaign
xpi@ucsd.edu, {mw34, jizej2, haozhen3}@illinois.edu
* indicates equal contribution

Abstract

Smaller-scale Vision-Language Models (VLMs) often claim to perform on par with larger models in general-domain visual grounding and question-answering benchmarks while offering advantages in computational efficiency and storage. However, their ability to handle rare objects, which fall into the long tail of data distributions, is less understood. To rigorously evaluate this aspect, we introduce the "Uncontextualized Uncommon Objects" (UOUO) benchmark. This benchmark focuses on systematically testing VLMs with both large and small parameter counts on rare and specialized objects. Our comprehensive analysis reveals that while smaller VLMs maintain competitive performance on common datasets, they significantly underperform on tasks involving uncommon objects. We also propose an advanced, scalable pipeline for data collection and cleaning, ensuring the UOUO benchmark provides high-quality, challenging instances. These findings highlight the need to consider long-tail distributions when assessing the true capabilities of VLMs. Code and project details for UOUO can be found at https://zoezheng126.github.io/UOUO-Website/.

1 Introduction

The advent of Vision-Language Models (VLMs) has marked a revolutionary leap in the integration of natural language processing and computer vision, largely due to the capabilities of the self-attention mechanism and the Transformer architecture (Vaswani et al., 2023). These technologies allow VLMs to effectively process and fuse information from both text and images, leading to significant advancements in tasks that require multimodal understanding, such as visual question answering and image captioning (Radford et al., 2021; Li et al., 2023; Alayrac et al., 2022; Xu et al., 2023; Young et al., 2014).

VLMs, trained on large-scale datasets, typically boast high performance on general tasks involving everyday objects and common scenarios (Li et al., 2024; Du et al., 2022; Wang et al., 2023). However, models of smaller scale, defined here as having fewer than 70 billion parameters, often claim to match the capabilities of their larger counterparts on general domain tasks (Lin et al., 2015; Agrawal et al., 2016; Yu et al., 2016; Liu et al., 2024; Goyal et al., 2017; Yu et al., 2023b) while offering advantages in computational efficiency and storage. Despite these claims, the No-Free-Lunch Theorem (Wolpert and Macready, 1997) suggests that these smaller models may compromise on their ability to handle less common or more complex scenarios that lie in the long tail of data distributions.

One natural and intuitive hypothesis is that they are sacrificing their fitness to the elements on the long tail of the distribution. Empirical observations of real-world data frequently align with Zipf's and Power Law (Piantadosi, 2014; Clauset et al., 2009), which indicates that while some objects and concepts are exceedingly common, a vast number of them are rare and fall into the long tail of the distribution . Understanding how well VLMs handle these rare and uncommon instances is crucial for assessing their true robustness and applicability across diverse and nuanced contexts.

Despite the importance of this evaluation, there is currently a lack of dedicated benchmarks that systematically test VLMs on objects and concepts that are significantly outside the everyday norm. To address this gap, we introduce the "Uncontextualized Uncommon Objects" (UOUO) benchmark. The object class distribution of UOUO is systematically out of common image sources such as ImageNet (Russakovsky et al., 2015), COCO (Lin et al., 2015), and Open Image Dataset (Kuznetsova et al., 2020). Our goal is to rigorously test and quantify the performance of both large-scale and small-scale VLMs on elements from the long tail of

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 6432–6441
November 12-16, 2024 ©2024 Association for Computational Linguistics

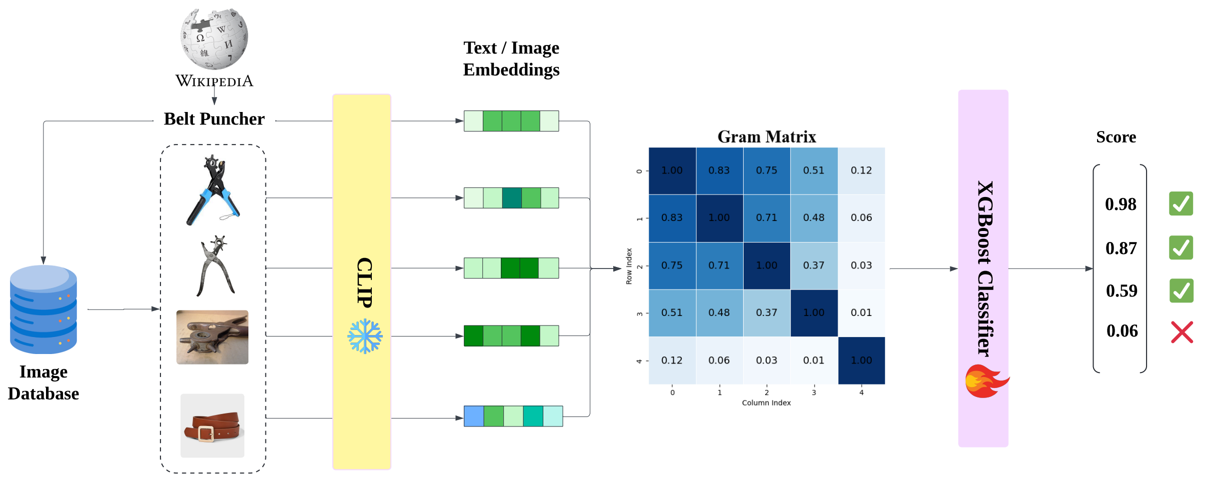

Figure 1: UOUO Data Curation Pipeline. Snowflake means frozen weights, and fire means tune-able weights.

the distribution to showcase their knowledge gap.

The contribution of our work is three-fold. (1) We compile a million-scale dataset specifically designed to include uncommon and uncontextualized objects, which are rarely encountered in everyday contexts but are significant in specialized domains. (2) We evaluate the performance gap between large-scale and small-scale VLMs when dealing with these rare elements, showcasing the significant knowledge and performance gap between large- and small-scale model on the long-tail distributions. (3) We propose a systematic pipeline for automatic and scalable data collection and cleaning, ensuring high-quality and representative testing instances.

2 Related Work

Real-world VQA Benchmarks Based on our survey, the typical real-world visual question answering datasets (excluding mathematics, celebrity, landmark, place, OCR and chart-reading) used in popular open-source VLMs such as LLaVa (Li et al., 2024), CogVLM (Wang et al., 2023) BLIP2 (Li et al., 2023), Qwen VL (Bai et al., 2023) and MiniCPM-V (Yu et al., 2023a) includes the following: COCO (Lin et al., 2015), RefCOCO (Yu et al., 2016), NoCAPs (Agrawal et al., 2019), MMBench (Liu et al., 2024), VQA-v2 (Goyal et al., 2017), OK-VQA (Marino et al., 2019), MME (Fu et al., 2024), GQA (Hudson and Manning, 2019).

Much to our surprise, it turns out that the image sources of GQA, RefCoCo, OK-VQA, MME Coarse-Grained Recognition, VQA-v2, and a significant proportion of MMBench are all direct random samples from COCO. Only NoCAPs features novel object classes (sourced from the 600-categories Open Image Dataset (Kuznetsova et al., 2020) outside COCO's less-than-100 common classes. This showcases the significant limitation of categorical diversity of extant VQA datasets. The knowledge and performance gap between the small- and large- scale VLMs might be concealed in such low coverage and diversity.

Existing Datasets with Uncommon Object Labels In extant datasets, Stanford Cars (Krause et al., 2013), CUB-bird (Wah et al., 2011), Deepfish (Saleh et al., 2020), ROCOv2 (Rückert et al., 2024), FGVC-Aircraft (Maji et al., 2013) also features rare object labels. Some non-academic mine & stone datasets, and chemical objects datasets can also be found on internet. However, the typical emphasis of these datasets is either *fine-grained subtype* or subspecies of common objects, or *domain-specific expert knowledge*. In realistic use cases such as autonomous car or embodied robotics, such knowledge might have limited generalizability.

3 Data Curation and Filtering

3.1 Domain Selection and Scraping

To construct the UOUO (Uncontextualized Uncommon Objects) benchmark, we began by selecting specific domains that are rich in specialized knowledge yet contain objects and tools that are rarely encountered by the general public. Our focus was on the industry sector, given its diversity and the presence of numerous specialized tools and equipment. These artificial tools are significantly out of the distribution of ImageNet, COCO, and Open Image Dataset.

We used Wikipedia as a starting point, targeting the page dedicated to manufacturing (https://en.wikipedia.org/wiki/Manufacturing). For each sub-sector identified within this domain, we employed GPT-4-Turbo (OpenAI, 2024) to generate a list of the top 50 objects or tools pertinent to experts in the field but obscure to the general populace. This list was generated through prompt-based querying, asking the model to identify objects that are crucial within the industry but not commonly known.

Once we had our list of uncommon objects, we performed a Google Image Search for each object name. For each query, we collected the top 50 image results. This approach allowed us to gather a diverse set of images representing each object under different conditions and contexts. For detailed dataset statistics of UOUO, we refer readers to Appendix C.

Mannual Annotation The image instances collected from Google Image Search can be noisy, with perhaps one fifth irrelevant instances for each queried uncommon category. To ensure the quality and relevance of the dataset, we implemented a rigorous annotation and cleaning process, combining manual and automated techniques. Our team manually reviewed and annotated on a subset of the collected categories of images to identify and remove outliers and noisy data. Categories with consistent visual representation across examples were retained, while those filled with ambiguous or irrelevant images were discarded. This initial curation aimed to maintain high fidelity to the object's intended representation. The instruction for manual annotation of UOUO can be found in Appendix B.

Automatic Data Cleaning We utilized the CLIP model to further enhance the dataset. CLIP (Contrastive Language–Image Pre-training) provides embeddings for both images and text, enabling us to compute similarities within and across categories. For each image, we extracted its CLIP image embedding E_i^c and the text embedding T_c of its corresponding category name (Radford et al., 2021; Sun et al., 2023). We calculated the cosine similarity between all pairs of image embeddings within each category to construct a GRAM matrix G, where $G_{i,j} = \text{Cosine}(E_i^c, E_j^c)$. Additionally, we computed the image-text similarity for each image as $\text{Cosine}(E_i^c, T_c)$, alongside statistical metrics such as the percentile, mean, and variance of the average similarity within each category.

The complete feature set includes image embeddings, GRAM percentiles (25th, 50th, and 75th), GRAM mean and variance, the instance's mean similarity with other images, and the percentiles of

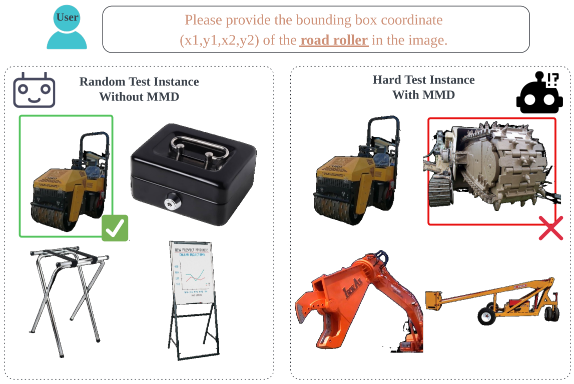

Figure 2: With MMD, we can retrieve harder negative examples and construct higher-quality test instances.

its pairwise similarities. Additionally, it incorporates image-text similarity metrics, corresponding percentiles, z-scores, and the instance label.

Using these computed features, we applied an XGBoost classifier to label each image instance. This classifier was trained on manually cleaned data from 500 categories to distinguish between high-quality and low-quality instances based on their similarity scores.

We optimized our XGBoost classifier (Chen and Guestrin, 2016) through 5-fold cross-validation and grid search to identify the best hyperparameters. The optimal configuration consisted of a maximum tree depth of 6, 200 estimators, a learning rate of 0.15, a subsample ratio of 1.0, gamma value of 0.1, and a colsample-bytree of 1.0. Additionally, the regularization parameters included reg-lambda of 1.5 and reg-alpha of 0.0, with a minimum child weight of 1.0.

The classifier achieved an accuracy of 0.8754 on cross-validation, closely aligning with human judgment, and exhibited Macro-Average Precision, Recall, and F1-Score of 0.8631, 0.8353, and 0.8460, respectively.

4 Test Instances Generation

Background Removal and Decontextualization Connectionist neural networks (including VLMs) are notoriously known for their tendency of overfitting to spurious correlations present in the training data. For instance, in our collected data, bulldozers are often seen in construction scenes laden with materials such as sand, concrete, and bricks. This high co-occurrence can lead models to rely on these contextual cues rather than truly understanding and recognizing the bulldozer itself. To mitigate this issue and ensure that models focus on the objects rather than their typical environments, we implement a robust background removal process to decontextualize all candidate objects in our dataset. To achieve effective background removal, we utilize a state-of-the-art, off-the-shelf background removal model (BRIA-AI, 2024).

Testing Instances Generation To assess the performance of Vision-Language Models on our UOUO benchmark, we generated challenging test instances designed to probe the models' capabilities beyond common knowledge. Specifically, we employ the CLIP embeddings combined with the Maximum Mean Discrepancy (MMD) with a Gaussian RBF kernel (Dziugaite et al., 2015) to identify and retrieve hard negative examples.

Let $\mathbf{x}$ and $\mathbf{y}$ be the sets of CLIP embeddings for two different object categories, each of shape (n, d), where n is the number of embeddings and d is the embedding dimension.

The Maximum Mean Discrepancy (MMD) between sets of embeddings $\mathbf{x}$ and $\mathbf{y}$ is calculated as follows:

$$\mathrm{MMD}(\mathbf{x}, \mathbf{y}) = k(\mathbf{x}, \mathbf{x}) + k(\mathbf{y}, \mathbf{y}) - 2 \cdot k(\mathbf{x}, \mathbf{y})$$

where the Gaussian Radial Basis Function (RBF) kernel value $k(\mathbf{a}, \mathbf{b})$ is defined as:

$$k(\mathbf{a}, \mathbf{b}) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{a}_i - \mathbf{b}_j\|^2\right)$$

For our calculations, we set $\sigma = 10$.

We use the Mosaic Image Augmentation Technique (Ge et al., 2021) to generate testing data in a scalable way. Each testing data point is created from ***four*** images, each background-removed. The four images contain objects of different categories but share some similar visual properties such as structures, colors, or textures. The selection of these images is determined by the Maximum Mean Discrepancy (MMD) distance between the categories they belong to. The closer the MMD distance, the more similar in features they might appear. We create an 800x800 canvas large enough to accommodate all four images. Then, each of the four images is augmented and positioned on the canvas's top-left, top-right, bottom-left, or bottom-right. The ground-truth bounding box for the object grounding is generated from the segmentation mask of background removal and normalized to be dimension-insensitive, accounting for potential differences in the VLM's rescaling process. Figure 2 showcases an exemplar test instance.

Model	mIoU-mmd	mIoU-rand	acc-mmd	acc-rand
llava-v1.5-7b	0.18	0.41	0.42	0.70
llava-v1.5-13b	0.23	0.47	0.44	0.73
llava-v1.6-vicuna-7b	0.28	0.48	0.49	0.75
llava-v1.6-vicuna-13b	0.28	0.49	0.52	0.78
llava-v1.6-34b	0.38	0.55	0.57	0.83
cogvlm-llama3-chat-19b	0.49	0.69	0.43	0.60
gemini-1.5-pro	0.27	0.27	0.63	0.80
gpt-4-turbo	0.34	0.38	0.67	0.90
gpt-4o	0.33	0.35	0.68	0.88

Table 1: Mosaic Grounding Performance Metrics

5 Experiment

Procedures Following the aforementioned test instance generation, we test both open source VLMs that are trained to perform grounding, including: llava-v1.5-7b, llava-v1.5-13b (Liu et al., 2023), llava-v1.6-vicuna-7b, llava-v1.6-vicuna-13b, llava-v1.6-34b (Li et al., 2024), cogvlm-v1.5-vicuna-7b (Wang et al., 2023), and propriety VLMs including: gemini-1.5-pro (Team, 2024), gpt-4-turbo, gpt-4o (OpenAI, 2024).

We test VLMs' performance on both randomly generated test instances and the MMD-augmented hard instances. We employ two metrics to quantitfy the performance: *mIoU* - Mean IoU (Intersection over Union), a standard metric for object segmentation; and *Accuracy* , which we prompt the VLM to output one positions from "top-left, top-right, bottom-left, bottom-right", and directly evaluate whether the answer matches the ground truth. The prompts used in this experiment can be found in Appendix A.

Observations and Analysis We present all experimental results in Table 1. (a) Comparing horizontally across columns, we observe significant performance drops of smaller-scale models in both *mIoU* and *Accuracy* with the application of MMD-based hard instance generation. Notably, the performance drops of many of them are around 30%. This provides solid support for our initial hypothesis that smaller-scale models have some, but insufficient fitness to the long-tail distribution objects. Furthermore, the drastic performance change showcases MMD's effectiveness in generating hard instances and non-robustness of existing grounding models. (b) Comparing vertically within columns, the central tendency is that larger scale models (except Genimi which might not be trained to perform Grounding) perform much better than small-scale models in accuracy. This reveals the concealed gap of knowledge horizon of small- and large- scale models, which is usually unobservable in benchmarks consist of common objects. (c) The observation that GPT-4 series can still handle the task

Figure 3: A glimpse into COCO and UOUO, with demo images of COCO cited from official website.

remarkably well (near 90% and 70% on random and MMD settings, respectively) showcases the task's solvability, revealing the soundness of our automatically constructed test instances.

6 Conclusion

In our work, we introduced the UOUO benchmark to assess VLMs on objects out of everyday distributions. Our findings show that while smaller VLMs perform well on tasks of common objects, they struggle significantly with uncommon objects, unlike larger models which handle these challenges much better. This highlights the need to consider long-tail distributions in evaluations. The systematic data curation, filtering, and hard test instance generation pipeline for UOUO construction has high extensibility, paving the road of future research of long-tail distribution objects. UOUO itself could also be expanded in this way, extending beyond the domain of manufacturing and to other broad category of objects.

7 Limitations

One limitation of our work is the reliance on automated data collection and cleaning processes, though efficient, may introduce biases or fail to capture nuanced representations compared to fully manual curation. We also note that the Mosaic Image Augmentation was applied with the assumption that the model takes single-image inputs. Our preliminary experiment showed most VLMs have limited to none multi-image inference support, thus multi-image inputs results are not included in UOUO benchmark. The UOUO benchmark currently emphasizes static images, potentially overlooking the dynamic and context-dependent nature of object recognition in real-world scenarios. Future extensions should explore a wider range of uncommon objects across various fields and consider the inclusion of video or sequential data to better reflect real-world applications. Addressing these limitations will enhance the comprehensiveness and applicability of the UOUO benchmark.

Acknowledgement

This work was supported by DARPA ECOLE HR00112390063 and by the National Science Foundation grants NSF CNS 19-00875, NSF CNS 21-06592, NSF OAC 18-35834 KN, NSF CCF 22-17144. This research used the Delta advanced computing and data resource which is supported by the National Science Foundation (award OAC 2005572) and the State of Illinois. Any results and opinions are our own and do not represent views of National Science Foundation.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2016. Vqa: Visual question answering. *Preprint*, arXiv:1505.00468.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957.

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. *Preprint*, arXiv:2204.14198.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

BRIA-AI. 2024. Bria background removal v1.4 model card. `https://huggingface.co/briaai/RMBG-1.4`. Accessed: 2024-06-13.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. ACM.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. 2015. Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. 2021. Yolox: Exceeding yolo series in 2021. *Preprint*, arXiv:2107.08430.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *Preprint*, arXiv:1612.00837.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Preprint*, arXiv:1902.09506.

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*, pages 554–561.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning.

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. 2024. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. 2013. Fine-grained visual classification of aircraft. Technical report.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Steven T. Piantadosi. 2014. Zipf's word frequency law in natural language: a critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130.

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. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. Imagenet large scale visual recognition challenge. *Preprint*, arXiv:1409.0575.

Johannes Rückert, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Cynthia S. Schmidt, Sven Koitka, Obioma Pelka, Asma Ben Abacha, Alba G. Seco de Herrera, Henning Müller, Peter A. Horn, Felix Nensa, and Christoph M. Friedrich. 2024. Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. *Preprint*, arXiv:2405.10004.

Alzayat Saleh, Issam H Laradji, Dmitry A Konovalov, Michael Bradley, David Vazquez, and Marcus Sheaves. 2020. A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Scientific Reports*, 10(1):14671.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *Preprint*, arXiv:2303.15389.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

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

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogvlm: Visual expert for pretrained language models. *Preprint*, arXiv:2311.03079.

David H Wolpert and William G Macready. 1997. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82.

Peng Xu, Xiatian Zhu, and David A. Clifton. 2023. Multimodal learning with transformers: A survey. *Preprint*, arXiv:2206.06488.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. *Preprint*, arXiv:1608.00272.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2023a. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023b. Mm-vet: Evaluating large multimodal models for integrated capabilities. *Preprint*, arXiv:2308.02490.

Appendix

A Prompt Details

mIoU Prompt Please provide the bounding box coordinate (x1,y1,x2,y2) of {object name} in the image with the format \n item1:(x1,y1,x2,y2).

Accuracy Prompt Identify the location of the given object in this 2x2 mosaic image. The possible answers are: 'top left', 'top right', 'bottom left', 'bottom right', or 'none'. Only give a deterministic response as one of the possible answers. If the object is not present, the response should be 'none'. Please do not give more than one response. \n object name: {object name}\n Location:

Prefix setting All other settings follow the model's defaults. For instance, in the case of llava, the prompt prefix is: A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human's questions. USER: <image>\n{question} ASSISTANT:

B Instructions for manual annotation

Consider removing individual images, or removing the entire category completely. Any category or image meeting the following criterion should be removed.

- **Category Exclusion Principles**
  1. **Lack of Sufficiently Uniform Images** Categories should be excluded if the collected images do not show enough consistency in appearance.

2. **Ambiguous Collections of Objects**
   Categories representing a collection of multiple objects (e.g. Wax Working Tools, Tools, Kits, etc.) should be excluded.
3. **Insufficient Number of Images Collected**
   Categories should be removed if there are not enough images available.
4. **Not a Tool**
   Items that are not standalone tools (e.g. pastes or liquids that always need to be stored in a container) should be excluded.

- **Image Sample Exclusion Principles**

  1. **Distinctive Image**
     Images that are significantly different from other images within the same category should be excluded.
  2. **Cluttered Composition**
     Images with cluttered backgrounds that make it difficult or impossible to isolate the tool.
  3. **Partial Display**
     Images that only show part of the object should be excluded.
  4. **Not a Real Object**
     Images that depict diagrams, 3D renders, or other non-realistic representations should be excluded.
  5. **Excessive Text**
     Images that contain excessive text, which obscures the main object, should be excluded.

C Important statistics of UOUO

- **Number of categories:**
  - Number of categories originally: 27,926
  - Number of categories kept: 25,864
  - Percentage of categories kept: 92.6%

- **Total number of images:**
  - Number of images originally: 956,167
  - Number of images kept: 678,535
  - Percentage of images kept: 71.0%

- **Images per category stats:**
  - **Original dataset:**
    * Average: 34.382
    * Minimum: 11
    * Maximum: 48
  - **Filtered dataset:**
    * Average: 26.235
    * Minimum: 5
    * Maximum: 48

- **Average percentage of images kept in each category:** 76.0%

D Wikipedia Industry List

See figure 4.

E Randomly sampled 100 categories

See table 2.

V · T · E	Major industries	[hide]
Natural sector		[hide]
Biotic	Agriculture	Arable farming (Cereals · Legumes · Vegetables · Fiber crops · Oilseeds · Sugar · Tobacco) · Permanent crops (Apples et al. · Berries · Citrus · Stone fruits · Tropical fruit · Viticulture · Cocoa · Coffee · Tea · Nuts · Olives · Medicinal plants · Spices) · Horticulture (Flowers · Seeds) · Animal husbandry (Beef cattle · Dairy farming · Fur farming · Horses · Other livestock · Pig · Wool · Poultry · Beekeeping · Cochineal · Shellac · Silk) · Hunting (Fur trapping)
	Forestry	Silviculture (Bamboo) · Logging (Firewood) · Rattan · Tree tapping (Frankincense · Gum arabic · Gutta-percha · Maple syrup · Mastic · Natural rubber · Palm sugar, syrup, & wine · Pine resin) · Wild mushrooms (Fungiculture · Truffles)
	Aquatic	Fishing (Anchovies · Herring · Sardines · Cod · Haddock · Pollock · Mackerel · Shark · Swordfish · Tuna · Crabs · Lobsters · Sea urchins · Squid · Whaling) · Aquaculture (Carp · Catfish · Tilapia · Abalone · Mussels · Oysters · Pearls · Microalgae · Seaweed) · Both (Clams · Sea cucumbers · Scallops · Salmon · Shrimp)
Geological		Fossil fuels (Coal · Peat · Natural gas · Oil shale · Petroleum · Tar sands) · Mining of ores (Aluminum · Copper · Iron · Gold · Silver · Palladium · Platinum · Lithium · Rare-earth metals · Uranium) · Other minerals (Gemstones · Phosphorus · Potash · Salt · Sulfur) · Quarrying (Gravel · Sand · Chalk · Clay · Gypsum · Limestone · Dimension stone · Granite · Marble)
Industrial sector		[hide]
Manufacturing	Light industry	Food (Animal feed · Baking · Canning · Dairy products · Flour · Meat · Prepared · Preserved · Sweets · Vegetable oils) · Beverages (Beer · Bottled water · Liquor · Soft drinks · Wine) · Textiles (Carding · Dyeing · Prints · Spinning · Weaving · Carpets · Lace · Linens · Rope) · Clothing (Accessories · Dressmaking · Furs · Hatmaking · Sewing · Shoemaking · Tailoring) · Printing (Bookbinding · Embossing · Engraving · Secure · Typesetting) · Media reproduction (Cassette tapes · Phonographs · Optical discs) · Metal fabrication (Boilermaking · Builders' & household hardware · Cutlery · Gunsmithing · Locksmithing · Machining · Other smithing · Powder metallurgy · Prefabrication · Surface finishing) · Other fabrication (3D printing · Blow molding · Drawing · Extrusion · Glassblowing · Injection moulding · Pottery · Sintering · Stonemasonry · Woodworking) · Furniture · Other goods (Baggage · Bicycles · Jewellery · Medical supplies · Musical instruments · Office supplies · Outdoors & sports equipment · Personal protective equipment · Toys)
	Electrical & optical	Electronics (Components · Circuit boards · Semiconductors) · Computers (Computer systems · Parts & peripherals · Blank storage media) · Communications equipment (Mobile phones · Network infrastructure) · Consumer electronics (Televisions · Video game consoles) · Instrumentation (Clocks & watches · GPS devices · Scientific instruments) · Medical imaging systems · Optical instruments (Cameras · Gun & spotting scopes · Laser construction · Lens grinding · Microscopes · Telescopes) · Electrical equipment (Batteries · Electrical & fiber optic cables · Electric lighting · Electric motors · Home appliances · Transformers)
	Chemicals	Coal & oil refining (Bitumen · Coke · Diesel fuel · Fuel oil · Gasoline · Jet fuel · Kerosene · Mineral oil · Paraffin wax · Petrochemicals · Petroleum jelly · Propane · Synthetic oil · Tar) · Commodity chemicals (Fertilizers · Industrial gases · Pigments · Pure elements) · Speciality chemicals (Adhesives · Agrochemicals · Aroma compounds · Cleaning products · Cosmetics · Explosives · Fireworks · Paints & inks · Perfumes · Soap · Toiletries) · Fine chemicals · Pharmaceuticals (Antibiotics · Blood products · Chemical & hormonal contraceptives · Generic drugs · Illegal drugs · Supplements · Vaccines)
	Materials	Leather (Liming & deliming · Tanning · Currying & oiling) · Wood (Drying · Sawmilling · Engineered · Lumber · Composite) · Paper (Sizing · Cardboard · Pulp · Tissue) · Rubber (Tires · Vulcanized rubber) · Plastics (Commodity · Engineered · Specialty · Pellets · Synthetic fibers · Thermoplastics & thermosets) · Glass (Borosilicate · Fused quartz · Soda-lime · Float glass · Glass fiber · Glass wool & fiberglass · Safety glass) · Ceramics (Brick · Earthenware · Porcelain · Refractory · Tile) · Cement (Mortar · Plaster · Ready-mix concrete) · Other mineral (Abrasives · Carbon fibers & advanced materials · Mineral wool · Synthetic gems) · Metal refining (Iron · Aluminum · Copper) · Alloys (Steel) · Formed metal (Rolled · Forged) · Cast metal
	Heavy industry	Machinery (Conveyors · Heavy · Hydraulic · Machine tools · Power & wind turbines) · Automobiles · Other heavy vehicles (Aerospace & space · Rail vehicles · Ships & offshore platforms) · Weapons
Utilities		Power (Electric · Gas distribution · Renewable) · Water (Sewage) · Waste management (Collection · Dumping · Hazardous · Recycling) · Remediation · Telecom networks (Cable TV · Internet · Mobile · Satellite · Telephone)
Construction		Buildings (Commercial · Industrial · Residential) · Civil engineering (Bridges · Railways · Roads · Tunnels · Canals · Dams · Dredging · Harbors) · Specialty trades (Cabinetry · Demolition · Electrical wiring · Elevators · HVAC · Painting and decorating · Plumbing · Site preparation)
Service sector		[hide]
Sales		Retail (Car dealership · Consumer goods · General store · Grocery store · Department store · Mail order · Online shopping · Specialty store) · Wholesale (Auction · Brokerage · Distribution)
Transport & Storage		Cargo (Air cargo · Intermodal · Mail · Moving company · Rail · Trucking) · Passenger transport (Airlines · Car rentals · Passenger rail · Ridesharing · Taxis) · Warehousing (Self storage)
Hospitality		Foodservice (Drink service · Cafés · Catering · Fast food · Food delivery · Restaurants · Teahouses) · Hotels
Asset management		Financial services (Banking · Credit · Financial advice · Holding company · Money transfer · Payment cards · Risk management · Securities) · Insurance (Health · Life · Pension funding · Property · Reinsurance) · Real estate (Brokerage · Property management)
Professional		Accounting (Assurance · Audit · Bookkeeping · Tax advice) · Architecture & engineering (Inspection · Surveying · Physical, product, & system testing) · Design (Fashion · Interior · Product) · Legal services · Management (Consulting · Public relations) · Marketing (Advertising)
Healthcare		Medicine (Dentist offices · Hospitals · Nursing) · Residential care · Veterinary medicine
Entertainment & leisure		Gambling (Online) · Sport · Venues (Arcades · Amusement parks · Fairgrounds · Nightclubs)
Other		Administrative (Customer service · Leasing · Renting · Staffing · Private investigation & security) · Maintenance (Janitors · Landscaping) · Repairs · Personal services (Beauty · Dry cleaning · Funeral · Maid service · Pet care · Sex) · Poverty · Travel (Business travel · Cruise lines · Tourism)
Information sector		[hide]
Publishing & Mass media		Written (Books · Periodicals · Software) · Audio-visual (Film · Music · Video games) · Broadcasting (News · Radio · Television) · Internet (Hosting · Social networks · Streaming · Websites)
Education		Primary · Secondary · Tertiary (Vocational school · University) · Testing · Tutoring
Other		Creative · Language · Research and development (Basic research)

Figure 4: Wikipedia Industry List

2D pantograph	AC Recharge Kit	Adhesive scale	Aluminum dross processing machine
Artificial insemination gun	Ballistic clipboard	Ballot Box (for collecting anonymous feedback)	Banjo rim lathe
Bingo balls	Broodstock tanks	Broom	Burnishing Stone
Cable Retention Sleeve	Carding Machine	Cattle Curtain	Cell Model
Climbing rope	Coal centrifuge	Coffee roaster	Cold Storage Backpack
Compressor (hardware)	Cooling Incubator	Copy Stand	Culture trays
Dehooking tool	Deposit Slip Printer	Disc golf basket welder	Disc repair kit
Display Turntables	Distillation column	Electronic rate board	Evaporating Dish
Extrusion laminator	Fiber disc	Fishing rod holders	Flange spreader
Flower press	Foundation crack ruler	Fume Extraction Hood	Goniophotometer
Graduated cylinders	Granule Filler	Inductively Coupled Plasma (ICP) Spectrometer	Irrigation pipelayer
Lacquer polishing brush	Leachate Collection Pipe	Live Feed Incubator	Longlines and ropes
Martingale	Metal scribe	Mobile manufacturing unit (MMU)	Mushroom grow tent
Music on hold player	Network Firewall Hardware	Offshore aquaculture cage	Ore skip
Oscillating shaker	Oxygen concentrators	Packing Gauge	Pellets coating system
Pellicle Formation Tool	Pillory	Pin beater	Pointer stick
Portable battery booster	Pressure vessels	Print Quality Inspection Scope	Pulling post
Purging compound dispenser	Queue stanchion	Quick release hook	Roll Coating Paint Line
Rope pump	Rotary drum bauxite washer	Rotary impeller feeder	Sand filter
Scale Breaker	Schlenk flask	Security drone	Security token device
Shear Line	Shock Absorber	Sign language interpreter gloves	Slab Tongs
Slush ice machines	Soap scum remover	Spin Welder	Spoke cutting machine
Spot meter	Springform pan	Tabbing shears for composite test specimens	Texture sprayer
Tower Climbing Harness	Violin varnish brush	Vixen Plate	Wall Hooks for Art
Waste basket	Water jet cutter for stone	Whalebone Scraper	Wire Mesh Cable Trays

Table 2: List of 100 Randomly Sampled Categories