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

Anonymous ACL submission

Abstract

Smaller-scale Vision-Langauge 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.

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 the distribution to showcase their knowledge gap.

The contribution of our work is three-fold. (1)

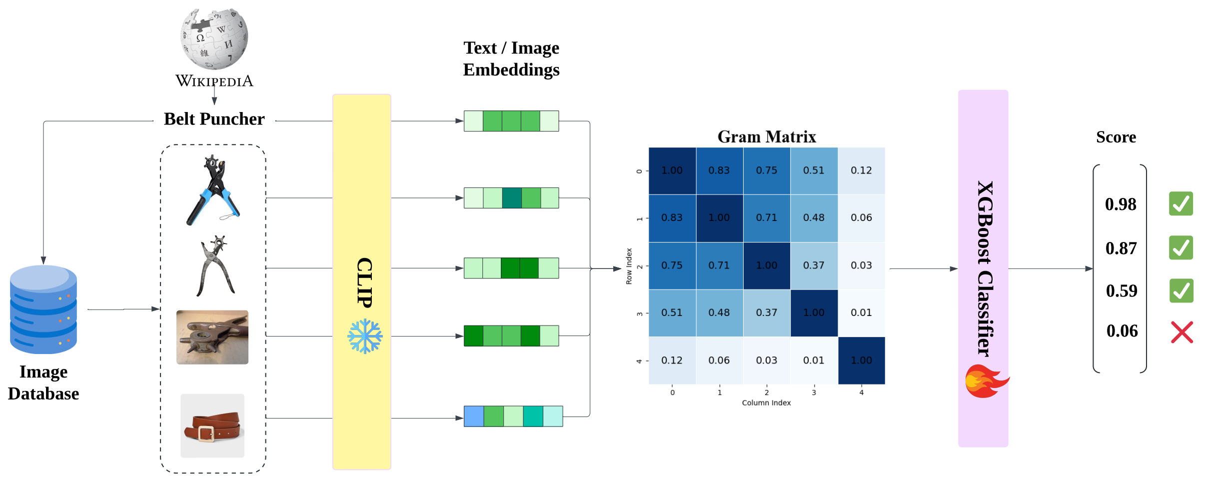

Figure 1: UOUO Data Curation Pipeline. Snowflake means frozen weights, and fire means tune-able weights.

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 et al., 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 A.

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 the collected images for each object category 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.

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 $G_{i,j} = Cosine(E_i^c, E_j^c)$. Additionally, we computed the image-text similarity for each image as $Cosine(E_i^c, T_c)$. Furthermore, we add basic statistical metrics, such as the percentile of the average-similarity with respect to other category members of a given instance, the mean and variance of the average-similarity of the category.

Using these computed features, we applied an XGBoost classifier to label each image instance. This classifier was trained 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

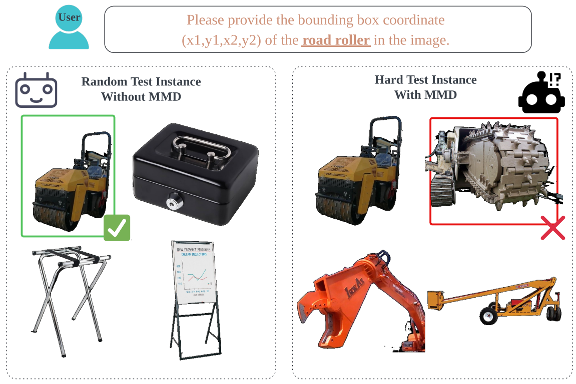

Figure 2: With MMD, we can retrieve harder negative examples and construct higher-quality test instanecs.

grid search to identify the best hyper-parameters. 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

Model	mIoU-mmd	mIoU-rand	acc-mmd	acc-rand
llava-v1.5-7b	0.1755	0.4117	0.4160	0.6954
llava-v1.5-13b	0.2334	0.4711	0.4351	0.7300
llava-v1.6-vicuna-7b	0.2779	0.4783	0.4924	0.7511
llava-v1.6-vicuna-13b	0.2761	0.4945	0.5220	0.7773
llava-v1.6-34b	0.3774	0.5504	0.5745	0.8324
cogvlm-llama3-chat-19b	0.4905	0.6935	0.4278	0.6024
gemini-1.5-pro	0.2654	0.2682	0.6326	0.7986
gpt-4-turbo	0.3396	0.3774	0.6650	0.8970
gpt-4o	0.3286	0.3472	0.6779	0.8814

Table 1: Mosaic Grounding Performance Metrics

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:

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

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.

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 et al., 2024), gpt-4-turbo, gpt-4o (OpenAI et al., 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.

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 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.

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. The UOUO benchmark also 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.

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, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 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, Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry, Lepikhin, Timothy Lillicrap, Jean baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, Luke Vilnis, Oscar Chang, Nobuyuki Morioka, George Tucker, Ce Zheng, Oliver Woodman, Nithya Attaluri, Tomas Kocisky, Evgenii Eltyshev, Xi Chen, Timothy Chung, Vittorio Selo, Siddhartha Brahma, Petko Georgiev, Ambrose Slone, Zhenkai Zhu, James Lottes, Siyuan Qiao, Ben Caine, Sebastian Riedel, Alex Tomala, Martin Chadwick, Juliette Love, Peter Choy, Sid Mittal, Neil Houlsby, Yunhao Tang, Matthew Lamm, Libin Bai, Qiao Zhang, Luheng He, Yong Cheng, Peter Humphreys, Yujia Li, Sergey Brin, Albin Cassirer, Yingjie Miao, Lukas Zilka, Taylor Tobin, Kelvin Xu, Lev Proleev, Daniel Sohn, Alberto Magni, Lisa Anne Hendricks, Isabel Gao, Santiago Ontanon, Oskar Bunyan, Nathan Byrd, Abhanshu Sharma, Biao Zhang, Mario Pinto, Rishika Sinha, Harsh Mehta, Dawei Jia, Sergi Caelles, Albert Webson, Alex Morris, Becca Roelofs, Yifan Ding, Robin Strudel, Xuehan Xiong, Marvin Ritter, Mostafa Dehghani, Rahma Chaabouni, Abhijit Karmarkar, Guangda Lai, Fabian Mentzer, Bibo Xu, YaGuang Li, Yujing Zhang, Tom Le Paine, Alex Goldin, Behnam Neyshabur, Kate Baumli, Anselm Levskaya, Michael Laskin, Wenhao Jia, Jack W. Rae, Kefan Xiao, Antoine He, Skye Giordano, Lakshman Yagati, Jean-Baptiste Lespiau, Paul Natsev, Sanjay Ganapathy, Fangyu Liu, Danilo Martins, Nanxin Chen, Yunhan Xu, Megan Barnes, Rhys May, Arpi Vezer, Junhyuk Oh, Ken Franko, Sophie Bridgers, Ruizhe Zhao, Boxi Wu, Basil Mustafa, Sean Sechrist, Emilio Parisotto, Thanumalayan Sankaranarayana Pillai, Chris Larkin, Chenjie Gu, Christina Sorokin, Maxim Krikun, Alexey Guseynov, Jessica Landon, Romina Datta, Alexander Pritzel, Phoebe Thacker, Fan Yang, Kevin Hui, Anja Hauth, Chih-Kuan Yeh, David Barker, Justin Mao-Jones, Sophia Austin, Hannah Sheahan, Parker Schuh, James Svensson, Rohan Jain, Vinay Ramasesh, Anton Briukhov, Da-Woon Chung, Tamara von Glehn, Christina Butterfield, Priya Jhakra, Matthew Wiethoff, Justin Frye, Jordan Grimstad, Beer Changpinyo, Charline Le Lan, Anna Bortsova, Yonghui Wu, Paul Voigtlaender, Tara Sainath, Shane Gu, Charlotte Smith, Will Hawkins, Kris Cao, James Besley, Srivatsan Srinivasan, Mark Omernick, Colin Gaffney, Gabriela Surita, Ryan Burnell, Bogdan Damoc, Junwhan Ahn, Andrew Brock, Mantas Pajarskas, Anastasia Petrushkina, Seb Noury, Lorenzo Blanco, Kevin Swersky, Arun Ahuja, Thi Avrahami, Vedant Misra, Raoul de Liedekerke, Mariko Iinuma, Alex Polozov, Sarah York, George van den Driessche, Paul Michel, Justin Chiu, Rory Blevins, Zach Gleicher, Adrià Recasens, Alban Rrustemi, Elena Gribovskaya, Aurko Roy, Wiktor Gworek, Sébastien M. R. Arnold, Lisa Lee, James Lee-Thorp, Marcello Maggioni, Enrique Piqueras, Kartikeya Badola, Sharad Vikram, Lucas Gonzalez, Anirudh Baddepudi, Evan Senter, Jacob Devlin, James Qin, Michael Azzam, Maja Trebacz, Martin Polacek, Kashyap Krishnakumar, Shuo yiin Chang, Matthew Tung, Ivo Penchev, Rishabh Joshi, Kate Olszewska, Carrie Muir, Mateo Wirth, Ale Jakse Hartman, Josh Newlan, Sheleem Kashem, Vijay Bolina, Elahe Dabir, Joost van Amersfoort, Zafarali Ahmed, James Cobon-Kerr, Aishwarya Kamath, Arnar Mar Hrafnkelsson, Le Hou, Ian Mackinnon, Alexandre Frechette, Eric Noland, Xiance Si, Emanuel Taropa, Dong Li, Phil Crone, Anmol Gulati, Sébastien Cevey, Jonas Adler, Ada Ma, David Silver, Simon Tokumine, Richard Powell, Stephan Lee, Kiran Vodrahalli, Samer Hassan, Diana Mincu, Antoine Yang, Nir Levine, Jenny Brennan, Mingqiu Wang, Sarah Hodkinson, Jeffrey Zhao, Josh Lipschultz, Aedan Pope, Michael B. Chang, Cheng Li, Laurent El Shafey, Michela Paganini, Sholto Douglas, Bernd Bohnet, Fabio Pardo, Seth Odoom, Mihaela Rosca, Cicero Nogueira dos Santos, Kedar Soparkar, Arthur Guez, Tom Hudson, Steven Hansen, Chulayuth Asawaroengchai, Ravi Addanki, Tianhe Yu, Wojciech Stokowiec, Mina Khan, Justin Gilmer, Jaehoon Lee, Carrie Grimes Bostock, Keran Rong, Jonathan Caton, Pedram Pejman, Filip Pavetic, Geoff Brown, Vivek Sharma, Mario Lučić, Rajkumar Samuel, Josip Djolonga, Amol Mandhane, Lars Lowe Sjösund, Elena Buchatskaya, Elspeth White, Natalie Clay, Jiepu Jiang, Hyeontaek Lim, Ross Hemsley, Zeyncep Cankara, Jane Labanowski, Nicola De Cao, David Steiner, Sayed Hadi Hashemi, Jacob Austin, Anita Gergely, Tim Blyth, Joe Stanton, Kaushik Shivakumar, Aditya Siddhant, Anders Andreassen, Carlos Araya, Nikhil Sethi, Rakesh Shivanna, Steven Hand, Ankur Bapna, Ali Khodaei, Antoine Miech, Garrett Tanzer, Andy Swing, Shantanu Thakoor, Lora Aroyo, Zhufeng Pan, Zachary Nado, Jakub Sygnowski, Stephanie Winkler, Dian Yu, Mohammad Saleh, Loren Maggiore, Yamini Bansal, Xavier Garcia, Mehran Kazemi, Piyush Patil, Ishita Dasgupta, Iain Barr, Minh Giang, Thais Kagohara, Ivo Danihelka, Amit Marathe, Vladimir Feinberg, Mohamed Elhawaty, Nimesh Ghelani, Dan Horgan, Helen Miller, Lexi Walker, Richard Tanburn, Mukarram Tariq, Disha Shrivastava, Fei Xia, Qingze Wang, Chung-Cheng Chiu, Zoe Ashwood, Khuslen Baatarsukh, Sina Samangooei, Raphaël Lopez Kaufman, Fred Alcober, Axel Stjerngren, Paul Komarek, Katerina Tsihlas, Anudhyan Boral, Ramona Comanescu, Jeremy Chen, Ruibo Liu, Chris Welty, Dawn Bloxwich, Charlie Chen, Yanhua Sun, Fangxiaoyu Feng, Matthew Mauger, Xerxes Dotiwalla, Vincent Hellendoorn, Michael Sharman, Ivy Zheng, Krishna Haridasan, Gabe Barth-Maron, Craig Swanson, Dominika Rogozińska, Alek Andreev, Paul Kishan Rubenstein, Ruoxin Sang, Dan Hurt, Gamaleldin Elsayed, Renshen Wang, Dave Lacey, Anastasija Ilić, Yao Zhao, Adam Iwanicki, Alejandro Lince, Alexander Chen, Christina Lyu, Carl Lebsack, Jordan Griffith, Meenu Gaba, Paramjit Sandhu, Phil Chen, Anna Koop, Ravi Rajwar, Soheil Hassas Yeganeh, Solomon Chang, Rui Zhu, Soroush Radpour, Elnaz Davoodi, Ving Ian Lei, Yang Xu, Daniel Toyama, Constant Segal, Martin Wicke, Hanzhao Lin, Anna Bulanova, Adrià Puigdomènech Badia, Nemanja Rakićević, Pablo Sprechmann, Angelos Filos, Shaobo Hou, Víctor Campos, Nora Kassner, Devendra Sachan, Meire Fortunato, Chimezie Iwuanyanwu, Vitaly Nikolaev, Balaji Lakshminarayanan, Sadegh Jazayeri, Mani Varadarajan, Chetan Tekur, Doug Fritz, Misha Khalman, David Reitter, Kingshuk Dasgupta, Shourya Sarcar, Tina Ornduff, Javier Snaider, Fantine Huot, Johnson Jia, Rupert Kemp, Nejc Trdin, Anitha Vijayakumar, Lucy Kim, Christof Angermueller, Li Lao, Tianqi Liu, Haibin Zhang, David Engel, Somer Greene, Anaïs White, Jessica Austin, Lilly Taylor, Shereen Ashraf, Dangyi Liu, Maria Georgaki, Irene Cai, Yana Kulizhskaya, Sonam Goenka, Brennan Saeta, Ying Xu, Christian Frank, Dario de Cesare, Brona Robenek, Harry Richardson, Mahmoud Alnahlawi, Christopher Yew, Priya Ponnapalli, Marco Tagliasacchi, Alex Korchemniy, Yelin Kim, Dinghua Li, Bill Rosgen, Kyle Levin, Jeremy Wiesner, Praseem Banzal, Praveen Srinivasan, Hongkun Yu, Çağlar Ünlü, David Reid, Zora Tung, Daniel Finchelstein, Ravin Kumar, Andre Elisseeff, Jin Huang, Ming Zhang, Ricardo Aguilar, Mai Giménez, Jiawei Xia, Olivier Dousse, Willi Gierke, Damion Yates, Komal Jalan, Lu Li, Eri Latorre-Chimoto, Duc Dung Nguyen, Ken Durden, Praveen Kallakuri, Yaxin Liu, Matthew Johnson, Tomy Tsai, Alice Talbert, Jasmine Liu, Alexander Neitz, Chen Elkind, Marco Selvi, Mimi Jasarevic, Livio Baldini Soares, Albert Cui, Pidong Wang, Alek Wenjiao Wang, Xinyu Ye, Krystal Kallarackal, Lucia Loher, Hoi Lam, Josef Broder, Dan Holtmann-Rice, Nina Martin, Bramandia Ramadhana, Mrinal Shukla, Sujoy Basu, Abhi Mohan, Nick Fernando, Noah Fiedel, Kim Paterson, Hui Li, Ankush Garg, Jane Park, DongHyun Choi, Diane Wu, Sankalp Singh, Zhishuai Zhang, Amir Globerson, Lily Yu, John Carpenter, Félix de Chaumont Quitry, Carey Radebaugh, Chu-Cheng Lin, Alex Tudor, Prakash Shroff, Drew Garmon, Dayou Du, Neera Vats, Han Lu, Shariq Iqbal, Alex Yakubovich, Nilesh Tripuraneni, James Manyika, Haroon Qureshi, Nan Hua, Christel Ngani, Maria Abi Raad, Hannah Forbes, Jeff Stanway, Mukund Sundararajan, Victor Ungureanu, Colton Bishop, Yunjie Li, Balaji Venkatraman, Bo Li, Chloe Thornton, Salvatore Scellato, Nishesh Gupta, Yicheng Wang, Ian Tenney, Xihui Wu, Ashish Shenoy, Gabriel Carvajal, Diana Gage Wright, Ben Bariach, Zhuyun Xiao, Peter Hawkins, Sid Dalmia, Clement Farabet, Pedro Valenzuela, Quan Yuan, Ananth Agarwal, Mia Chen, Wooyeol Kim, Brice Hulse, Nandita Dukkipati, Adam Paszke, Andrew Bolt, Kiam Choo, Jennifer Beattie, Jennifer Prendki, Harsha Vashisht, Rebeca Santamaria-Fernandez, Luis C. Cobo, Jarek Wilkiewicz, David Madras, Ali Elqursh, Grant Uy, Kevin Ramirez, Matt Harvey, Tyler Liechty, Heiga Zen, Jeff Seibert, Clara Huiyi Hu, Andrey Khorlin, Maigo Le, Asaf Aharoni, Megan Li, Lily Wang, Sandeep Kumar, Norman Casagrande, Jay Hoover, Dalia El Badawy, David Soergel, Denis Vnukov, Matt Miecnikowski, Jiri Simsa, Praveen Kumar, Thibault Sellam, Daniel Vlasic, Samira Daruki, Nir Shabat, John Zhang, Guolong Su, Jiageng Zhang, Jeremiah Liu, Yi Sun, Evan Palmer, Alireza Ghaffarkhah, Xi Xiong, Victor Cotruta, Michael Fink, Lucas Dixon, Ashwin Sreevatsa, Adrian Goedeckemeyer, Alek Dimitriev, Mohsen Jafari, Remi Crocker, Nicholas FitzGerald, Aviral Kumar, Sanjay Ghemawat, Ivan Philips, Frederick Liu, Yannie Liang, Rachel Sterneck, Alena Repina, Marcus Wu, Laura Knight, Marin Georgiev, Hyo Lee, Harry Askham, Abhishek Chakladar, Annie Louis, Carl Crous, Hardie Cate, Dessie Petrova, Michael Quinn, Denese Owusu-Afriyie, Achintya Singhal, Nan Wei, Solomon Kim, Damien Vincent, Milad Nasr, Christopher A. Choquette-Choo, Reiko Tojo, Shawn Lu, Diego de Las Casas, Yuchung Cheng, Tolga Bolukbasi, Katherine Lee, Saaber Fatehi, Rajagopal Ananthanarayanan, Miteyan Patel, Charbel Kaed, Jing Li, Shreyas Rammohan Belle, Zhe Chen, Jaclyn Konzelmann, Siim Põder, Roopal Garg, Vinod Koverkathu, Adam Brown, Chris Dyer, Rosanne Liu, Azade Nova, Jun Xu, Alanna Walton, Alicia Parrish, Mark Epstein, Sara McCarthy, Slav Petrov, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 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.

A Appendix

Important statistics of UOUO are listed as follows:

- **Number of categories:**
  - Filtered data directory: 25,864
  - Original data directory: 27,926
  - Percentage of categories kept: 92.6%
- **Total number of images:**
  - Filtered dataset: 678,535
  - Original dataset: 956,167
  - Percentage of images kept: 71.0%
- **Images per category stats:**
  - **Filtered dataset:**
    - Average: 26.235
    - Minimum: 5
    - Maximum: 48
  - **Original dataset:**
    - Average: 26.235
    - Minimum: 5
    - Maximum: 48
- **Average percentage of images kept in each category:** 76.0%

Randomly sampled 100 categories:

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)
Biotic	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)
Biotic	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)
Manufacturing	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)
Manufacturing	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)
Manufacturing	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
Manufacturing	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 3: 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