Multi-modal Zero-Shot Object Detection in Egocentric Videos

Harsha Koneru
Center for Data Science
New York University
hk3820@nyu.edu

Sargun Nagpal
Center for Data Science
New York University
sn3250@nyu.edu

Sharad Dargan
Center for Data Science
New York University
sd5251@nyu.edu

Abstract

This study explores Zero-Shot Object Detection using Natural Language queries within egocentric video frames. Employing diverse object detection methodologies, including traditional and open vocabulary models, a comparative analysis of their efficacy is presented. Reproducing a baseline and conducting comprehensive error analysis, the research delineates optimal data augmentation strategies for enhancing closed vocabulary object detection performance. Additionally, adaptations of the OWL-ViT detector for referring expression tasks and a proposed modified architecture are discussed. Deliberating on the performance gap between closed and open vocabulary detection, the study outlines future steps to bridge this divide, offering a promising foundation for further advancements in open vocabulary object detection within egocentric video contexts.

1 Introduction

Traditional object detection approaches involve recognizing a predefined set of semantic categories in an image through supervised learning, which proves effective in specific domains but falls short in its applicability to open-world scenarios. This limitation arises due to the inability of these techniques to generalize beyond the fixed set of classes, their inability to differentiate between objects of the same category based on their parts and attributes, and the necessity to detect objects in the real world in zero or few shot settings.

One of the main reasons for restricting ourselves to a fixed set of categories is that obtaining training data with open labels is expensive and time-consuming. Furthermore, there is a dearth of datasets with exhaustive annotation of objects with their parts and attributes, and most of the existing ones restricted themselves to niche domains like pedestrians[1], or fashion[2].

The development of strong language encoders and the advent of contrastive training over images and text has enabled models to learn shared representations from text-image pairs present on the internet and has resulted in improvements to zero-shot detection performance[3][4] The development of large-scale parts and attribute-level annotated image datasets[5] presents an opportunity to learn richer representations of objects that integrate knowledge of their characteristics. This has applications in numerous domains, including AR/VR technologies and assistive systems for the visually impaired.

Our work aims to perform Zero-Shot Object Detection based on Natural Language queries on egocentric video frames[6] from the challenging PACO-Ego4D dataset. We leverage traditional object detection models[7], end-to-end object detectors[8], and open vocabulary object detectors[9][10] to solve this task and present a comparison between the approaches. We reproduce the baseline result from previous works[5], perform comprehensive error analysis to determine data augmentation strategies to improve upon the baseline in the closed vocabulary setting and adapt the OWL-ViT object detector for our referring expression detection task. We also describe a potential new pipeline

that could be used for this task. Finally, we discuss the gap in performance between closed and open vocabulary object detection and enumerate steps that could be taken in the future to close that gap. We believe this is a great starting point for future efforts in open vocabulary object detection on egocentric videos.

2 Related Work

Datasets with parts and attributes. While there are exhaustive parts and attributes datasets related to specific domains[2][1], datasets such as PASCAL-Part[11], PartImageNet[12], and CityScapes-Panoptic-Parts[13] provide parts and attributes annotations for common objects. However, unlike the 75 object categories in PACO[5], these datasets provide annotations for very few objects, the maximum being part annotations for 9 object categories provided by CityScapes.
Unimodal Object Detection. Traditional object detection models, primarily unimodal, focused on image-based detection within fixed semantic categories. These models include one-stage detectors like SSD[14] and YOLO [15], two-stage detectors such as Faster-RCNN[16] and Mask-RCNN[7], and transformer-based detectors like DETR[17].
Vision Language Learning. This involves the embedding of text and images into a shared space, and helps us achieve zero-shot generalization on several vision tasks. Most recent work in this domain leverages contrastive learning on text-image pairs found on the internet[3][18][19].
Multimodal Object Detection. This has evolved from adapting unimodal methods to multimodal tasks, to recent innovations like MDETR[10], OWL-ViT[9] and GLIP[20] which emphasize end-to-end models for object detection guided by language queries. MDETR and GLIP formulate their pretraining problem as a phrase grounding task, working with an image and a single text query. While this approach yields rich vison-language representations, it is unsuitable for Open Vocabulary Detection and Multi Modal Instance Search, as it limits the number of classes which can be processed in the forward pass. Therefore, using OWL-ViT was more computationally feasible for our task.

3 Problem Definition and Algorithm

3.1 Task

Zero-shot instance detection necessitates an algorithm capable of determining the bounding box for a specific object instance based on a provided "query" describing that instance. In this context, a level-k (Lk) query is defined as a description of an object instance using k attributes and/or parts. For instance, queries like "blue mug" or "mug with a blue handle" fall under L1, while "blue striped mug with a white handle" is categorized as an L3 query. Each query is linked to one positive image accompanied by a bounding box and a distractor set, consisting of up to 100 hard negative images. We use the Average Recall at k (AR@k), to evaluate the performance of a model on this task.

3.2 Algorithm

We explore two different classes of models, i.e. Closed-Vocabulary and Open-Vocabulary models.

3.2.1 Closed-Vocabulary Models

Closed-Vocabulary Models are trained to recognize a fixed set of predefined object classes, lacking the ability to identify or classify new, unseen object types. The baseline closed vocabulary model is a Mask R-CNN with ViT-L Backbone, which has a custom architecture with separate heads for predicting the Bounding Box, Objects and Parts classification, Semantic Segmentation, and Attribute Prediction. Mask R-CNN is a 2-stage model that extends Faster R-CNN by adding a parallel branch for pixel-level instance segmentation, efficiently identifying and segmenting objects in images with precise bounding boxes and masks. Based on Error Analysis, the baseline model struggled with detecting white/grey objects and those blending into the background. To address this, we retrained it using various data augmentation techniques.

3.2.2 Open-Vocabulary Models

We use Open-World Models for referring expression object-detection using Natural Language Queries. We investigated late fusion (OWL ViT[9]), early fusion (MDETR[10], GLIP[20]), and custom pipeline

(DETR+CLIP) for open vocabulary detection. We selected Google's OWL ViT[9] for our experiments with Open-Vocabulary models. OWL ViT uses a standard Vision Transformer architecture (with minimal modifications) and with detection heads. It is pre-trained contrastively with image and text encoders on large-scale image-text data and finetuned on detection data.

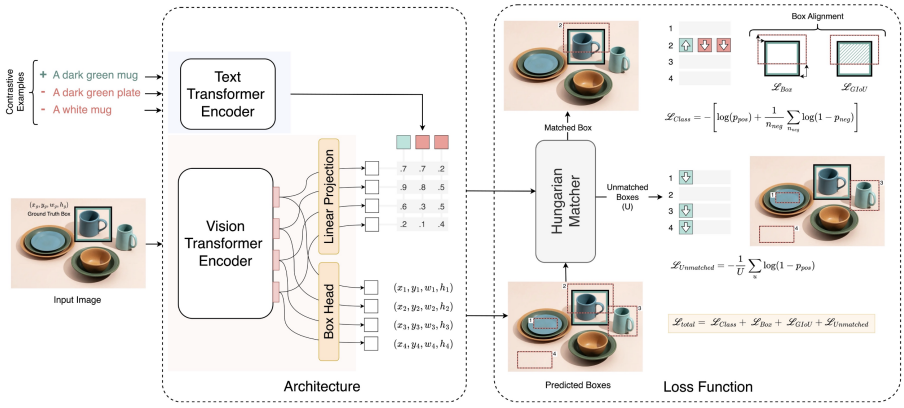

Figure 1: OWL-ViT Adaptation: Architecture and Loss Function

We finetuned the OWL ViT model using contrastive classification loss $\mathcal{L}_{class}$ on positive and negative queries for each predicted bounding box for the image:

$$\mathcal{L}_{class} = -\left[\log(p_{pos}) + \frac{1}{n_{neg}} \sum_{n_{neg}} \log(1 - p_{neg})\right]$$

where p_{pos}, p_{neg} are the probabilities of the bounding box belonging to the positive and negative queries respectively. We also add a component to suppress the spurious bounding boxes detected with a high probability for positive query, denoted as $\mathcal{L}_{unmatched}$:

$$\mathcal{L}_{Unmatched} = -\frac{1}{U} \sum_{u} \log(1 - p_{pos})$$

where U is the number of predicted bounding boxes that are not matched as the best-aligned prediction with the ground truth box. The total loss for this includes these 2 loss terms along with L1 Loss on Bounding box coordinates $\mathcal{L}_{Box}$ and Generalized IoU Loss $\mathcal{L}_{GIoU}$.

$$\mathcal{L}_{total} = \mathcal{L}_{class} + \mathcal{L}_{Unmatched} + \mathcal{L}_{GIoU} + \mathcal{L}_{Box}$$

This loss function was designed to address the shortcomings of the model without these terms based on empirical experiments.

We also formulated a pipeline using a Deformable DETR for box prediction and a CLIP model for vision language alignment. First, the image is passed through a pre-trained Deformable DETR model to get boxes. The class information is discarded as the purpose of the DETR model in our case is just as a box predictor. We then extract the embeddings of the detections through CLIP's vision module. We then align these vision embeddings with the text embeddings of the queries obtained from CLIP's text module. We trained this model on the contrastive dataset, and updated the gradients based on the cross-entropy loss from the alignment, the L1 box loss, and the GIoU loss coming from the DETR model.

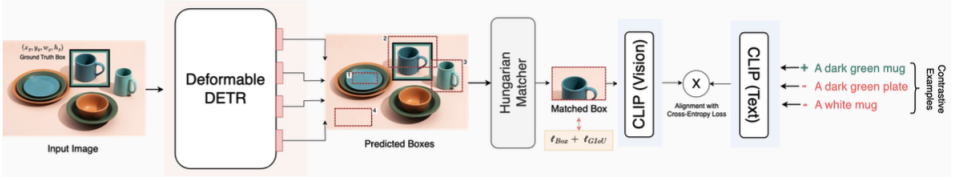

Figure 2: DETR + CLIP Pipeline

4 Experimental Evaluation

4.1 Data

4.1.1 Ego4D Dataset

We use the Ego4D [6] dataset which is one of the largest Egocentric (first person) video datasets that covers hundreds of scenarios (household, outdoor, workplace, leisure, etc.) of daily life activity captured in-the-wild by 926 unique camera wearers from 9 different countries. The annotations for the parts and attributes of objects are from the PACO[5] dataset which contains 75 object categories, 456 object-part categories, and 55 attributes for 15,667 images in the train dataset.

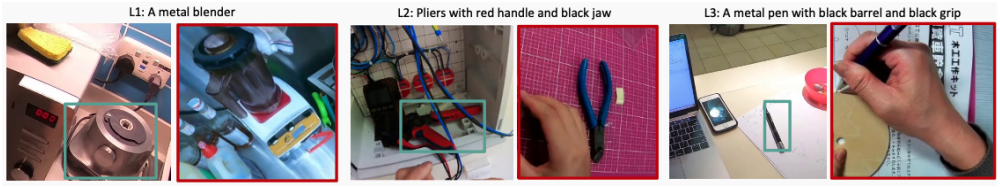

Figure 3: Positive and Distractor Images for sample queries

The Ego4D Test dataset consists of 2,282 queries with a mix of L1, L2, and L3 queries (defined in section 3.1) overall along with their PACO Annotations.

4.1.2 Contrastive Dataset Curation

Our approach addressed the absence of natural language queries in our training data by generating contrastive positive and hard negative queries from the ground truth dataset to train open-vocabulary detection models. The process involved sampling from object parts and attributes to form positive queries. Correspondingly, we sampled from object parts and complementary attributes to create negative queries. These negative queries comprised four categories designed to facilitate learning distinctions among attributes of the same object while safeguarding against catastrophic forgetting, thereby preserving the model's capacity to discern between distinct objects.

4.2 Methodology

To improve upon the Mask-RCNN baseline, we used contrast, brightness and random rotation augmentations to retrain the model. All experiments were done using Meta AI's detectron2 library.

For open-vocabulary object detection, we first established the baseline performance of the OWL ViT Base model on our test dataset to establish it's out-of-box performance. Further, we went on to finetune these models on our contrastive training dataset using the approach described in Section 3.2.2, and then compared the performance with the out-of-box performance of the model.

We use Average Recall (AR@k) metric, for all comparisons between model performances. All experiments were run using NYU High-Performance Computing (HPC) environment, using multiple NVIDIA RTX-8000 GPUs. The experiment's intermediate results and progress were tracked using the Weights and Biases tracking platform.

4.3 Results

4.3.1 Quantitative Results

Table 1 shows the Average Recall at k (AR@k) metrics obtained by closed and open vocabulary models across queries of different difficulty levels. Closed vocabulary models exhibited a superior performance compared to their open vocabulary counterparts. Within the closed vocabulary framework, incorporating augmentations such as contrast, brightness, and random crop enhanced AR@1 metrics. In the open vocabulary domain, the out-of-the-box performance of MDETR was poor, coupled with its computational expensiveness due to its early fusion architecture. Conversely, OWL-ViT, functioning as a late fusion model, demonstrated comparatively better performance. However, when subjected to intricate L3 queries, the OWL-ViT faced challenges, primarily stemming from its fine-tuning

exclusively with object names as queries. Nevertheless, fine-tuning the OWL-ViT model, integrating spurious box suppression, yielded noticeable improvements in AR@k metrics across all queries, particularly accentuated in the more complex L2 and L3 query categories.

Query Level	All		L1		L2		L3	
Metric (AR@k)	@1	@5	@1	@5	@1	@5	@1	@5
Closed Vocabulary Models								
Mask RCNN + ViT (Baseline)	24.52	**36.63**	21.74	**38.21**	21.98	**34.92**	27.25	**37.21**
Mask RCNN + ViT + Augmentation	**24.86**	35.92	**22.13**	37.86	**22.51**	34.07	**27.45**	36.45
Open Vocabulary Models								
MDETR (OOB)	0	0.24	0	0.31	0	0.13	0	0.22
OWL ViT B/32 (Baseline)	0.51	0.73	**2.35**	3.19	0.28	0.49	0	0
OWL ViT - FT	0.33	0.68	1.13	1.96	0.28	0.85	0.07	0.11
OWL ViT - FT + Spurious Box Supression	**0.67**	**1.57**	1.62	**3.26**	**0.56**	**1.84**	**0.41**	**0.77**

Table 1: Results

4.3.2 Qualitative Analysis

This Error analysis demonstrates the improvement in the quality of bounding boxes after adding the box-suppression loss. Further, this also shows that model fails in some cases, where it just considers a part of an object in the image and detects it as an object.

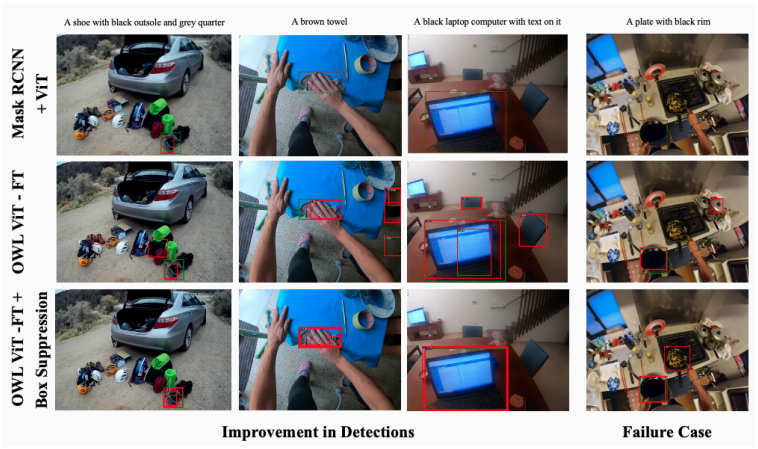

Figure 4: Qualitative Error Analysis

4.4 Discussion

Closed vs Open Vocabulary models: Closed vocabulary models offer higher accuracy within their set categories but lack adaptability to unrecognized classes, limiting their real-world use. Conversely, open vocabulary models show potential but require substantial development to surpass closed models, requiring enhancements in both data handling and architectural design.

Early vs Late Fusion in Open Vocabulary Models: Early fusion integrates multimodal data at an initial stage, offering improved vision-text understanding; however, it tends to face scalability challenges, since the same image must be passed through the model multiple times for each query.

Sensitivity to the Dataset: Using negative queries with the same object but different attributes helps the model differentiate between attributes while using negative queries with different objects prevents catastrophic forgetting. This balance in the dataset is crucial to succeed with contrastive learning.

5 Conclusions

This research compared closed and open-vocabulary models for referring expression tasks in egocentric video frames. Through the meticulous construction of contrastive queries and fine-tuning of models like OWL-ViT, and Deformable DETR + CLIP-based architectures, significant advancements were achieved. Notably, fine-tuning contributed substantially to improvements in the AR@k metric, particularly evident in addressing the complexities associated with higher-level language queries (L2 and L3). Additionally, the integration of a loss term aimed at suppressing spurious boxes resulted in enhanced detection capabilities. While closed vocabulary models demonstrated superior performance, their limitation of applying only to a fixed set of classes hampers real-world applicability. Conversely, open vocabulary models, with their early-fusion approach, provided richer contextual representations but suffered from computational inefficiencies when compared to late-fusion models. This study underscores the need for a balanced consideration of performance, adaptability, and computational efficiency in designing models for real-world applications in the domain of egocentric videos and beyond.

Shortcomings and Future Work

Contrastive Images: We currently only use contrastive queries to train our models. Using negative images, in addition to negative queries for contrastive training is expected to improve model performance.

Scaling-up: Expectedly, the OWL-ViT Large model should outperform the Base, warranting its finetuning. Additionally, scaling the data for more comprehensive training should be considered.

Improved architectures: We experimented with architecture using Deformable DETR and OpenAI CLIP embeddings, and it shows promising initial results. Future efforts will refine this architecture, focusing on box refinement, contextual features, and DETR layer expansion.

Poor performance of open-vocabulary models compared to closed-vocabulary models: The performance of open-vocabulary models is expected to improve if we use early-fusion models. However, these are computationally expensive to train and future work can address making vision-language early fusion more efficient using techniques like multi-query batch dimension.

6 Lessons Learned

Challenges: The project exposed us to the challenges of working with High-Performance Computing (HPC), setting up project environments, and running and tracking multiple GPU batch jobs. Reproducing projects, even in line with models devised by tech giants such as Google (OWL-ViT) and Microsoft (GLIP), posed a persistent challenge. However, the most challenging task was to debug problems with training our deep-learning models. Unlike conventional software development, the complexities within neural networks obscure the origins of errors. For example, a problem as simple as forgetting a negative sign in our IoU loss term led to degenerate boxes with zero width or height, with no clear indication of the source of error. Effective visualizations, error analysis, and logging of our experiments helped us debug these problems.

Takeaways: Our project served as an invaluable lesson in the significance of implementing good software engineering principles within Deep Learning projects. Using tools like Weights & Biases for experiment tracking, logging configurations and metrics, and saving model checkpoints, emerged as essential practices. We also learned about the iterative nature of model enhancement through rigorous error analysis, visualizing model predictions, and studying learning curves. Finally, our project equipped us with a wealth of knowledge on computer vision and multi-modal models and research, which sparked a curiosity to continue exploring further advancements in this field.

7 Student Contributions & Acknowledgements

HK implemented closed vocabulary models, and DETR + CLIP experiments. SD, SN implemented OWL-ViT and MDETR. All team members contributed valuable insights and feedback to each other's work and contributed equally to the report. We thank our mentors, Prof. Mengye Ren and Ying Wang for their continued support throughout this project. We also thank Prof. Brian McFee for his support and guidance.

References

[1] Dangwei Li, Zhang Zhang, Xiaotang Chen, Haibin Ling, and Kaiqi Huang. A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054*, 2016.

[2] Menglin Jia, Mengyun Shi, Mikhail Sirotenko, Yin Cui, Claire Cardie, Bharath Hariharan, Hartwig Adam, and Serge Belongie. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 316–332. Springer, 2020.

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[4] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[5] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7141–7151, 2023.

[6] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[8] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[9] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pages 728–755. Springer, 2022.

[10] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021.

[11] Yifan Zhao, Jia Li, Yu Zhang, and Yonghong Tian. Multi-class part parsing with joint boundary-semantic awareness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9177–9186, 2019.

[12] Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts. In *European Conference on Computer Vision*, pages 128–145. Springer, 2022.

[13] Daan de Geus, Panagiotis Meletis, Chenyang Lu, Xiaoxiao Wen, and Gijs Dubbelman. Part-aware panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5485–5494, 2021.

[14] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[15] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[17] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 4904–4916, 2021.

[19] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for zero-shot transfer learning. *Neurocomputing*, 555:126658, 2023.

[20] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.