Multi-modal Zero-Shot Object Detection in Egocentric Videos

Harsha Koneru, Sargun Nagpal, Sharad Dargan

Mentors: Mengye Ren, Ying Wang

NYU Center for Data Science

EGO4D PACO Zero-Shot Challenge

Problem Statement

Humans have an inherent ability to recognize unknown objects in their environment. While current detection models offer utility, the following issues impede their real-world usability.

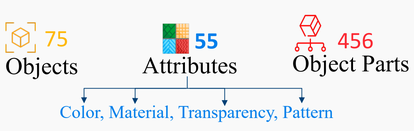
Zero-Shot Open World Detection	Detecting Objects is insufficient
Traditional object detection models are **restricted to a fixed set of object classes**. Open World Detection involves training models to adapt to novel or previously unseen objects.	Models that can identify objects through their **parts and attributes** allow for nuanced object recognition, crucial for advanced applications like robotics and image understanding

Problem: Zero-Shot Referring Expression Detection **using Natural Language Queries** that describe objects through their parts and attributes

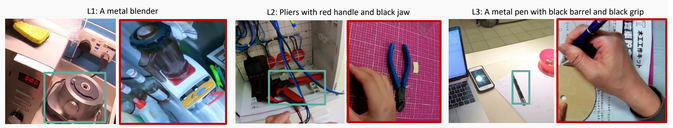
Applications: **AR/VR** technologies - **episodic memory** (such as recalling the location of specific items like "A pair of spectacles with a black rim") and developing **assistive technology** for the visually impaired.

Dataset & Metrics

PACO Ego4D Dataset: One of the largest Egocentric (first person) videos dataset 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.



PACO Ego4D Test Dataset: The Test dataset has queries of varying complexity (Lk) that describe the object with k-attributes



PACO Ego4D is a federated dataset: each object category has a subset of data that is exhaustively labeled for positive and negative annotations.

Evaluation Metric: We compute Average Recall (AR@k), where k=1,5,10 denotes the top-k boxes returned by the model for a given query. For a given query, AR@k is computed at different IoU thresholds (0.50 - 0.95) and then averaged over all thresholds and queries.

Related Work

Datasets with parts and attributes: While datasets such as PartImageNet (He et al., 2022) and PASCAL-Part (Chen et al., 2014) offer part-level annotations, they do so for only a limited set of objects (<10).

Uni-modal object detection: Uni-modal object detectors can be categorized into two-stage detectors like Faster-RCNN (Ren et al., 2015), one stage detectors like YOLO (Redmon et al., 2016) and transformer-based models like DETR (Carion et al., 2020). These methods have achieved great success on common benchmarks like COCO and LVIS.

Vision-Language Learning: Vision-Language Learning includes general representation learning, exemplified by models such as VisualBert (Li et al., 2019) and CLIP (Radford et al., 2021), as well as specialized models for Image Captioning and VQA.

Multi-modal object detection: Multi-modal Object Detection has evolved from adapting unimodal methods to multimodal tasks, to recent innovations like MDETR (Kamath et al., 2021), and OWL-ViT (Minderer et al, 2022), which emphasize end-to-end models for object detection guided by natural language queries.

Methodology

Closed Vocabulary Models

Baseline Model (PACO): Mask-RCNN + ViT-L Backbone with 4 heads for: Box prediction, Object/Part classification, Semantic segmentation and Attribute prediction.

Error Analysis: The baseline model struggled with detecting white/gray objects and those blending into the background. To address this, we retrained it using data augmentation techniques like adjusting brightness and contrast, and applying random rotation.

Open Vocabulary Models

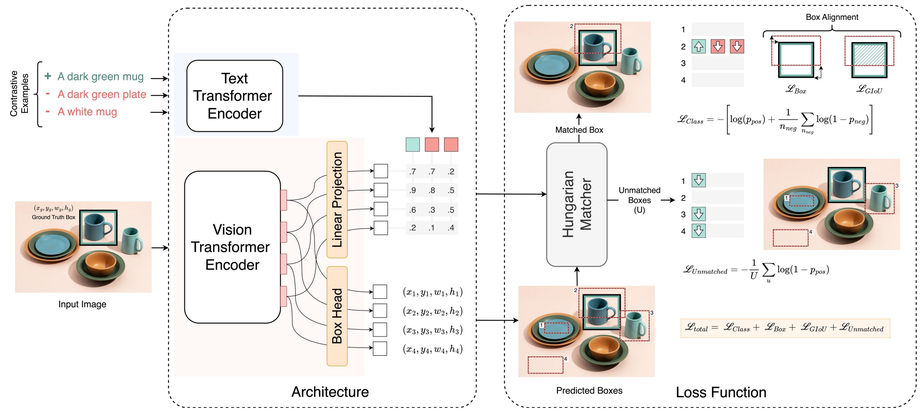
Contrastive Dataset Curation: We built Positive and Hard Negative queries from parts and attributes of objects in the Ego4D dataset.

Object	Part	Attributes	Complement
 Mug	Handle	White	Black, Red, …
		Plastic	Steel, Metal, …
	Base	Brown	Black, White, …
		Wooden	Plastic, Steel, …

Positive Queries:
- + A mug with a **white**, **plastic** handle
- + A mug with a **white** handle and **brown**, **wooden** base

Hard Negative Queries:
- - **Same Object, Different Attributes**: **Mug** with a **white** handle and **blue** base
- - **Different Object, Same Attributes**: **Jar** with a **white**, **plastic** handle
- - **Different Object, Different Attributes**: **Bottle** with a **grey**, **metal** body
- - **Different Object, No Attribute: A Bottle**

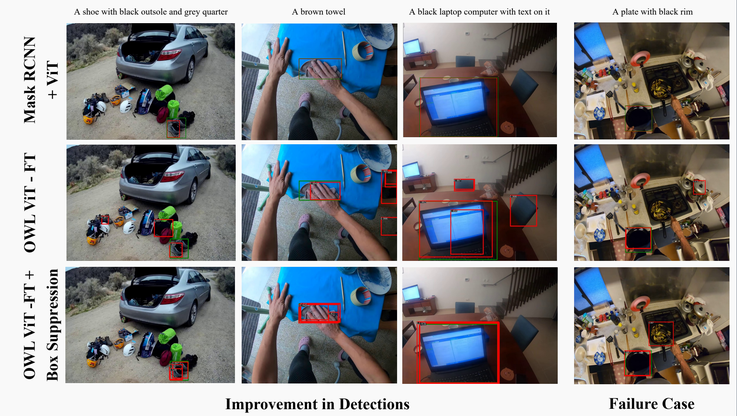
Models: We investigated late fusion (OWL ViT), early fusion (MDETR, GLIP), and custom architectures (DETR+CLIP) for open vocabulary detection. Specifically, we fine-tuned Google's **OWL ViT**, originally pretrained on object names, for the referring expression task. This involved altering the training process and introducing a **new loss term** to suppress spurious bounding boxes.



Results

Query Level	All			L1			L2			L3		
Metric (AR@k)	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10
Closed Vocabulary Models												
Mask RCNN + ViT (Baseline)	24.52	**36.63**	40.81	21.74	**38.21**	42.13	21.98	**34.92**	**38.98**	27.25	**37.21**	**41.57**
Mask RCNN + ViT + Augmentation	**24.86**	35.92	**41.02**	**22.13**	37.86	**43.16**	**22.51**	34.07	38.78	**27.45**	36.45	36.77
Open Vocabulary Models												
MDETR (OOB)	0	0.24	0.52	0	0.31	0.55	0	0.13	0.42	0	0.22	0.37
OWL ViT B/32 (Baseline)	0.51	0.73	0.78	**2.35**	3.19	3.46	0.28	0.49	0.49	0	0	0
OWL ViT - FT	0.33	0.68	0.83	1.13	1.96	2.6	0.28	0.85	0.95	0.07	0.11	0.11
OWL ViT - FT + Spurious Box Supression	**0.67**	**1.57**	**2.03**	1.62	**3.26**	**4.58**	**0.56**	**1.84**	**2.41**	**0.41**	**0.77**	**0.84**

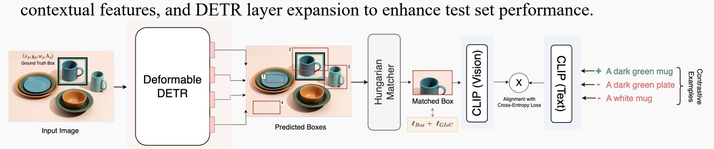
Discussion



- **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 richer contextual information for improved vision-text understanding; however, it tends to face scalability challenges.
- **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 of object level knowledge. This balance in dataset construction is crucial to succeed with contrastive learning.
- **Spurious Box Suppression:** We're modifying the loss function to not only increase the model's accuracy in identifying correct bounding boxes with associated phrases but also to lower its tendency to detect false positives.

Future Work

- **Contrastive Images:** Exploration of using negative images, in addition to negative queries for contrastive training. We expect the AR@k to improve with this additional contrastive element.
- **Scale-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.
- **New architecture:** We suggest combining Deformable DETR with OpenAI CLIP embeddings, and it shows potential. Future efforts will refine this architecture, focusing on box refinement, contextual features, and DETR layer expansion to enhance test set performance.



- **Efficient Vision-Language Early Fusion:** Explore efficient early fusion techniques for vision-language embeddings, like multi-query batch dimension, for richer representations. Current methods are too expensive and unsuitable for referring expression tasks.

References

Ramanathan, Vignesh, et al. "Paco: Parts and attributes of common objects." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

Minderer, Matthias, et al. "Simple open-vocabulary object detection." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022.

He, Kaiming, et al. "Mask r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2017.

Kamath, Aishwarya, et al. "Mdetr-modulated detection for end-to-end multi-modal understanding." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.