Evaluating Vision-Language Models for Zero-Shot Detection, Classification, and Association of Motorcycles, Passengers, and Helmets

Lucas Choi
Archbishop Mitty
lucasleechoi@gmail.com

Ross Greer
University of California San Diego
regreer@ucsd.edu

***Abstract*—Motorcycle accidents pose significant risks, particularly when riders and passengers do not wear helmets. This study evaluates the efficacy of an advanced vision-language foundation model, OWLv2, in detecting and classifying various helmet-wearing statuses of motorcycle occupants using video data. We extend the dataset provided by the CVPR AI City Challenge and employ a cascaded model approach for detection and classification tasks, integrating OWLv2 and CNN models. The results highlight the potential of zero-shot learning to address challenges arising from incomplete and biased training datasets, demonstrating the usage of such models in detecting motorcycles, helmet usage, and occupant positions under varied conditions. We have achieved an average precision of 0.5324 for helmet detection and provided precision-recall curves detailing the detection and classification performance. Despite limitations such as low-resolution data and poor visibility, our research shows promising advancements in automated vehicle safety and traffic safety enforcement systems.**

***Index Terms*—Vehicle safety, Zero-shot learning, Vision-language models, Helmet detection, Traffic enforcement systems**

I. Introduction

Motorcycle accidents are frequent causes of injury and death worldwide, especially for occupants not wearing helmets [1]–[4]. Specifically, in India, in 2022, two-wheeler deaths accounted for 44% of total road fatalities with 74,897 deaths, the highest out of all modes of transport[1]. Helmets are 35% effective in reducing the risk of Abbreviated Injury Scale 3+ head injuries [5]. Additionally, 4 people die every hour in India because they do not wear a helmet[2], causing 44,666 deaths in 2019 [6].

Section 129 of the Motor Vehicles Act in India states that "Every person ... on a motorcycle of any class or description shall, while in a public place, wear protective headgear conforming to such standards as may be prescribed by the Central Government." [7] Despite regulations mandating helmet use, compliance is inconsistent, leading to preventable injuries.

Iterations of the CVPR AI City Challenge [8] have prompted researchers to address this challenge, stating "Motorcycles are one of the most popular modes of transportation, particularly in developing countries such as India. Due to lesser protection compared to cars and other standard vehicles, motorcycle riders are exposed to a greater risk of crashes. Therefore, wearing helmets for motorcycle riders is mandatory as per traffic rules, and automatic detection of motorcyclists without helmets is one of the critical tasks in enforcing strict regulatory traffic safety measures." We suggest that, besides the enforcement of traffic safety measures, there is also an even greater benefit in the ability of IoT-style communication between infrastructure or egocentric perception devices. Such systems could detect the presence of motorcyclists and passengers (with or without helmets) and alert the surrounding vehicles whose drivers (autonomous or human) may be otherwise unaware of the vulnerable road users in their proximity [9].

Accordingly, to perceive holistic information about motorcycles and their occupants in a scene, the goal task we evaluate in this paper is the detection and classification of the following objects in every frame of a large video dataset:

1) Motorcycle,
2) Drivers wearing helmets,
3) Drivers not wearing helmets,
4) Passengers wearing helmets,
5) Passengers not wearing helmets,
6) 2nd Passengers wearing helmets,
7) 2nd Passengers not wearing helmets,
8) Children sitting in front of the driver wearing helmets,
9) Children sitting in front of the driver not wearing helmets.

In this research dataset, these scenes are captured by infrastructure-mounted cameras, though the same models can also be applied to egocentric views. This is especially the case given the zero-shot learning approaches we take, which do not require specific-view training data to be applied. We show sample data of these classes in Figure 1.

With many data-driven applications, a common challenge is the ability of a training set to adequately represent the diversity of instances that appear in the real world [10], [11]. For this reason, data-driven methods excel when given the most data possible, as this increases the likelihood of learning similar patterns to a real-world instance. To this end, we create a method that extends beyond the dataset presented by Shuo et al. [8] by employing a pre-trained vision-language foundation

[1]https://opencity.in/analysing-the-morth-road-accidents-report-for-2022/
[2]https://www.indiatoday.in/diu/story/two-wheeler-death-road-accidents-helmets-states-india-1602794-2019-09-24

Fig. 1: Example instances of classes to detect, cropped from the AI City Challenge dataset. From left to right: Motorcycle, Driver with Helmet, Driver with No Helmet, Child Passenger with No Helmet, Passenger 1 with Helmet, Passenger 1 with No Helmet, Passenger 2 with No Helmet.

model for this detection task, specifically, the OWLv2 [12]. Further, in our research, we present a strategy for cascading models to modularly isolate and improve task performance for these important safety systems.

This foundation model strategy is important especially in consideration of challenges presented by dataset shortcomings. The given dataset has no instances of a child passenger with a helmet or a second passenger with a helmet. This is a huge hindrance in accurately detecting the seat position and helmet status in these specific classes using traditional machine-learning approaches due to the lack of data for training. Therefore, the use of zero-shot learning may provide a means to identify these instances in 'real world' test data even without specific training.

The question we explore in this research is to what degree such foundation model approaches, namely OWLv2, are ready for use with real-world data in this motorcycle safety road scene perception domain and where their strengths and weaknesses may lie.

II. Related Research

Conventional machine learning object detection algorithms rely on manual annotations and specialized algorithms, which can be time-consuming and resource-intensive to label, especially as the models are limited to learning from provided datasets. Moreover, these methods often lack the flexibility to adapt to new environments or variations in helmet designs [13].

Foundation models, with billions of parameters trained on enormous collections of information, have recently led to effective zero-shot techniques for a variety of tasks [14], where a learned model can provide strong performance on datasets unseen during training [15]. One such foundation model is OWL-ViT [16]; OWL stands for "open-world localization", referring to this model's ability to function in an "open" world (i.e., non-rigidly specified set of expected classes). The ViT portion of OWL-ViT refers to the Vision Transformer, an architecture that applies the attention mechanism to images instead of the prior standard of convolution. The OWL family of models uses contrastive learning between batches of image patch encodings and text embeddings, with image patch encodings producing proposed classes and proposed bounding boxes, and treating detection as a bipartite matching problem between these decoded image classes and bounding boxes, as introduced in the Detection Transformer (DETR) technique [17]–[19]. Together, these methods were shown to be effective in zero-shot object detection (identifying a bounding box around desired classes of interest within an image). This method was refined and scaled up using self-training as OWLv2 [12], whereby pseudo-box annotations are provided from an existing detector, and it is this further-trained model that we use in the method shared in this research.

For the same application of detecting and classifying the given objects detailed in Section I, many different approaches have been tried in the previous AI City Challenges; in the 2023 AI City Challenge [20], Tran et al. [21] used YOLOv8 for a score of 0.7754 for the mean average precision (mAP). Cui et al. [22] used DETA [23] ensemble and Detectron2 for a mAP of 0.8340. In the 2024 AI City Challenge [8], mainly transformer models combined with ensemble techniques were used. Vo et al. [24] used Co-DETR [25] with a Minority Optimizer for class imbalance and a Virtual Expander for a mAP of 0.4860. Chen et al. [26] used a DETA and DETR fusion model for a score of 0.4824 mAP.

III. Algorithms for Image Processing with Vision-Language Detection

To address the challenges of accurately detecting and classifying motorcycles, their passengers, and helmet usage, we developed a cascading detection algorithm for OWLv2. Furthermore, due to OWLv2's shortcomings, we employed an AlexNet for the seat classification task. This section outlines our cascading detection algorithm using OWLv2 and discusses our approach for the seat classification task.

We first note that there are abstract classes that relate the target classes to one another; for example, "motorcycle" and "person" are the abstract classes represented in the data scheme, where "person" can be further classified based on the attributes of helmet-wearing and seating position. Due to this, our first goal is to detect these high-level classes. Further, we know that there is no driver or passenger without a motorcycle, so we only detect "person" in association with a particular motorcycle instance.

Our detection algorithm, illustrated in Figure 2, begins with a detection stage. We provide a scene image (resized to 960 by

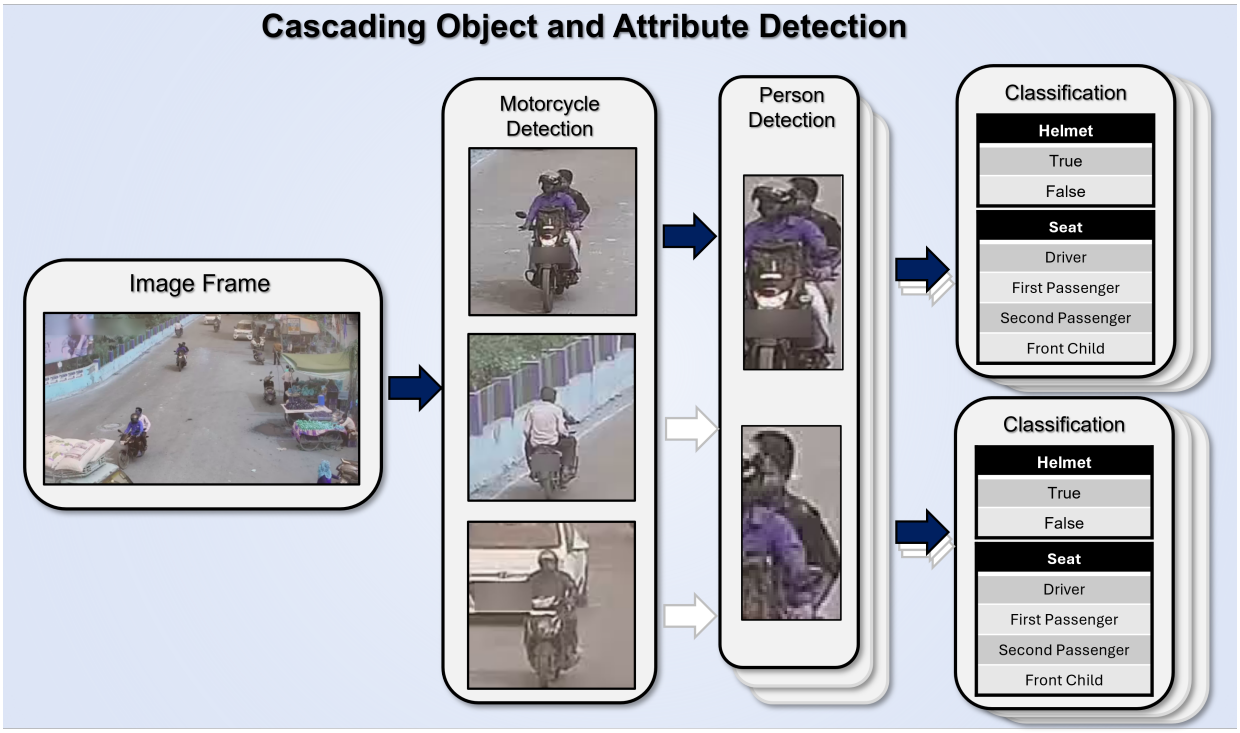

Fig. 2: Our algorithm for detecting the relevant objects for helmet safety, as well as the appropriate attributes, acts in a cascaded style. First, from the original image, we detect all motorcycles. Then, within each motorcycle, we detect all human occupants (drivers and passengers). Then, for each detected human, we perform helmet detection and seat position classification. All detections, including helmet detection for the purpose of classification, are done using OWLViT2, while seat position classification is done using AlexNet.

960 pixels and values normalized in [0, 1], relative to the size of each individual image in the batch) as input to OWLv2 along with the text "motorcycle". The CLIPTokenizer, from [15], encodes the text to be wrapped by the processor with the normalized image.

To detect the person instances on the motorcycles, we expand the re-scaled bounding box by 50 pixels on the left, right, and top sides to encapsulate any person instances surrounding the motorcycle. Using the expanded box, we crop the original image and run the OWLV2 model over this cropped image with the prompt "person" to detect person instances.

The algorithm's subsequent step is to perform the next level of detection, focusing on helmets, by cropping each person instance and running the OWLv2 over the cropped image with a text input of "helmet". Because our task is to classify each person based on whether they are wearing a helmet and not necessarily to detect the helmet itself, we store the boolean result of this detection as an attribute of the person (rather than noting the bounding box).

We note that there is a general difficulty of OWLv2 in differentiating a person's semantic position on the motorcycle (such as driver, passenger, second passenger, etc.), as noted in Section IV. For this particular portion of the task, we take a supervised learning approach. We seek to provide each person detected on the motorcycle with an attribute of location between the positions enumerated in the introduction.

Therefore, we use a neural network (a variant of AlexNet [27], with a final layer output of four) to classify the seating position on a motorcycle of the person instances detected with OWLv2.

Due to the use of the AlexNet in the seat position classification task, we recognize that the whole process is not completely zero-shot. It is rather a hybrid of zero-shot learning and supervised learning, with zero-shot for the association and detection of motorcycles, their passengers, and their helmet status, and supervised learning for the seat classification of the passengers. In this way, the methods in this paper actually address four tasks (motorcycle detection, person detection, helmet detection, and seat classification); three of these are solved in a zero-shot manner, and we include a learned approach to seat classification as this is a relevant safety task that should also be considered in conjunction.

In total, this algorithmic sequence of tasks can provide detections of motorcycles, associated people, their positions, and their helmet status for each image in a video.

IV. Experimental Method and Evaluation

Using the cascaded object and attribution detection algorithm detailed in the previous section, we performed detection on the dataset of 100 videos provided by [8], with further implementation details described in this section.

We first conducted the motorcycle and person detection of our cascaded detection process as described in Section III.

The threshold for OWLv2 is a confidence threshold, meaning it is the minimum confidence score that a predicted bounding box must have to be considered a valid detection. The OWLv2 will discard any detection with a confidence score below the given threshold. The confidence score is calculated as logits on a per-detection basis.

We performed the cascaded detection process with thresholds of 0.1 to 0.7 on the OWLv2 to examine the sensitivity of precision and recall to thresholds. 0.7 was chosen as the last threshold, as OWLv2 made no detections with a threshold higher than 0.7. Using the output of our detections, we calculated the precision and recall at each confidence threshold.

To evaluate the ability of OWLv2 to classify a passenger's helmet status, regardless of error in upstream person detection, we detected helmets within the ground truth bounding boxes of passengers to classify the passenger's helmet status. As in the previous detections, we experimented with a threshold of 0.05 to 0.7.

When performing the seat classification based on the person detection, we attempted to determine a passenger's seat with OWLV2, first using the text prompts provided by the labels in the dataset, such as "passenger 1" and "child passenger." However, with these prompts, OWLv2 tended to miss some passengers and mislabel the people. Assuming this was due to the inputs, we attempted more specific prompts such as "child in front of driver" or "passenger behind driver". Nevertheless, this also yielded similar results. We hypothesize that the prompt inputs were not the determining factor of OWLv2's failure to detect and differentiate the different people on a motorcycle, showing possible shortcomings of model training for this particular type of task. Furthermore, the task of classifying people based on their relative location to other people and the motorcycle may be too specific for the model.

After observing OWLv2's shortcomings with our intersection data, we used a modified AlexNet for the seat classification subtask [27]. We modified the last layer of the AlexNet from 10 outputs to 4 to suit our task.

We used an approximate inverse class frequency to overcome the severe class imbalance in the dataset as shown in Table I. At first, we tested the weighting of 1.147, 7.908, 785.229, and 2093.944, calculated by inverse class weighting. However, this was insufficient, as the model did not appear to learn the child class and appeared to over-favor the driver class. Therefore, we incrementally increased the weighting of the classes of passenger1, passenger2, and child passenger relative to the driver, updating the previously mentioned weights to 1, 10, 800, and 3000, respectively.

We split the data 70/15/15 for the training, testing, and validation. We used a cross-entropy loss. Finally, we trained the model using a learning rate of 0.0001 for 100 epochs. Then, we used the model with the lowest loss on the validation set to make inferences on the test set.

TABLE I: Ground Truth Data

Class	Frequency
Driver	32,889
Passenger 1	4,796
Passenger 2	78
Child Passenger	48
Total	37,811

Fig. 3: Sample images of the dataset of different angles with different environments. From top to bottom: night, foggy, crowded

A. Data

The dataset provided by [8] contains 100 videos taken by infrastructure-mounted cameras in India. They are annotated with bounding boxes of motorcycles and up to four passengers who may or may not be wearing helmets. Each video is 20 seconds long, sampled at 10 Hz, and has a resolution of 1920×1080. Example images from the dataset are shown in Figure 3. The ground truth data is comprised of class frequencies, as shown in Table I, and has 26349 helmet-wearers and 11462 unhelmeted people, meaning 69.7% are helmeted.

TABLE II: Precision and Recall scores of OWLv2 Motorcycle detections. No detections were made above the threshold of 0.7.

Threshold	Precision (IoU 0.5)	Recall (IoU 0.5)
0.7	0.7124	0.002095
0.6	0.7357	0.1232
0.5	0.6249	0.3384
0.4	0.5548	0.4453
0.3	0.4849	0.5258
0.2	0.3951	0.6108
0.1	0.2460	0.7226

B. Results

Our OWLv2 detected motorcycles with accuracies shown in Table II and detected persons with accuracies as shown in Table III. For motorcycle detection, the average precision is 0.4122, calculated by the area under the curve of Figure 4, and for person detection, the average precision is 0.3561, obtained from Figure 5.

Our helmet-status classification was done through helmet detection with OWLV2 on the provided ground truth bounding boxes of passengers, with a representative classification based on the helmet's presence or absence. This resulted in the precisions and recalls in Table IV, tested over multiple thresholds, resulting in an average precision of 0.5324, as further illustrated in Figure 6. A naive classifier, which always predicts the rider to be wearing a helmet, would have a precision of 69.7% and a trivial recall of 100% based on the ground truth data described in Section IV A; at all thresholds, our precision is higher than the naive classifier, showing a reduction in false positives and negatives.

The IoU in Tables II, III, and IV stands for intersection over union, which is the metric for evaluating the accuracy of a predicted bounding box. The IoU is calculated as follows: $IoU = \frac{A \cap B}{A \cup B}$ or $IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union}$, where A stands for the predicted bounding box and B stands for the ground truth bounding box. In our evaluation, for a given detected bounding box, if the IoU with the ground truth is greater than or equal to 0.5, then the detection is considered a "true positive".

Finally, our neural network's seat classification achieved an accuracy of 95.17% on the validation set, with the classification results on the test set displayed in Figure 3. We note that the severe class imbalance does leave the child passenger class unsuccessfully classified, though this does not have much impact on the accuracy metric. This reveals an insufficiency in the model learning and cautions us of evaluating performance for such an imbalanced dataset without examining class performance in the confusion matrix.

C. Sensitivity of Helmet Detection to OWLViT Detection Threshold

Due to the nature of the dataset, it is important to find an optimal threshold in our detections. As many videos within the data are often unclear, too high of a threshold may omit the detections within the unclear regions of the data. On the contrary, too low of a threshold may yield unrelated detections, such as detecting a bike as a motorcycle. Therefore, an optimal threshold between these two extremes is necessary to achieve the highest accuracy. We show the results of exploring multiple thresholds in our research and note that continual tuning will be important when applying these methods to additional datasets or tasks.

TABLE III: Precision and Recall scores of OWLv2 Passenger Detection. No detections were made above the threshold of 0.5.

Threshold	Precision (IoU 0.5)	Recall (IoU 0.5)
0.5	1.0	6.6136e-5
0.4	0.9437	0.02183
0.3	0.8861	0.1066
0.2	0.6992	0.2568
0.1	0.2672	0.5432

TABLE IV: Precision and Recall scores of OWLv2 Helmet Classification. No detections were made above the threshold of 0.7.

Threshold	Precision (IoU 0.5)	Recall (IoU 0.5)
0.7	0.9565	0.005845
0.6	0.9557	0.1046
0.5	0.9221	0.1980
0.4	0.8775	0.2698
0.3	0.8280	0.3370
0.2	0.7852	0.4143
0.1	0.7398	0.5298
0.05	0.7146	0.6370

V. Concluding Remarks and Future Research

Zero-shot learning demonstrates the potential of this application as it can overcome some limitations of incomplete and biased training datasets. As noted, the provided dataset lacks instances of child passengers with helmets and second passengers with helmets, making training traditional supervised-learning models difficult. Zero-shot learning leverages pre-training on diverse data, classifying unseen instances more

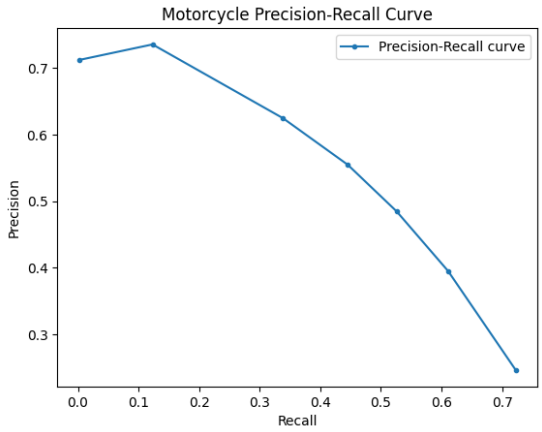

Fig. 4: Precision-Recall Curve of Motorcycle Detection. Initially, a slight increase in precision indicates improved confidence in early predictions. However, precision declines steeply as recall rises, highlighting the model's challenge in maintaining accuracy while capturing more true positives.

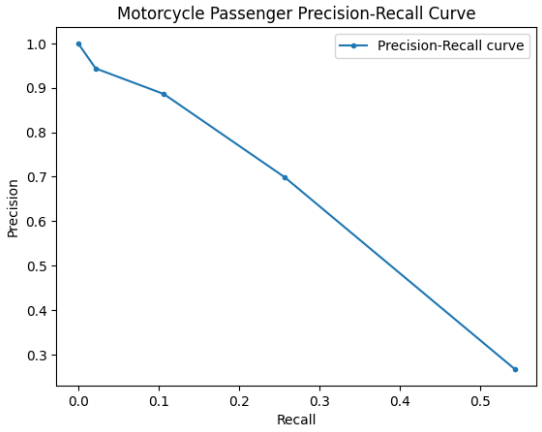

Fig. 5: Precision-Recall Curve of Passenger Detection. The curve demonstrates a high precision at low recall values. Despite the trade-off of precision and recall, the shape suggests a robust model performance in balancing the two.

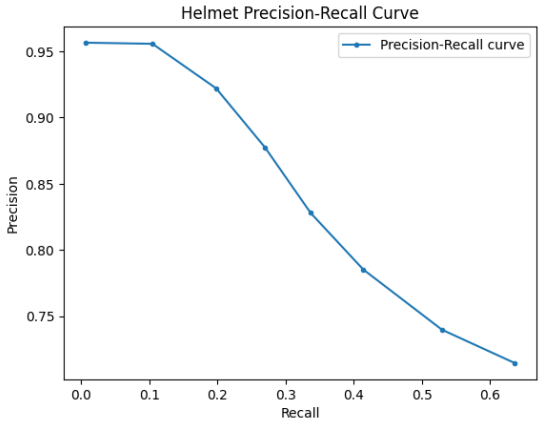

Fig. 6: Precision-Recall Curve of Helmet Classification
The curve has a high initial precision, progressively decreasing as recall increases. Efforts to capture more true positives resulted in a higher incidence of false positives.

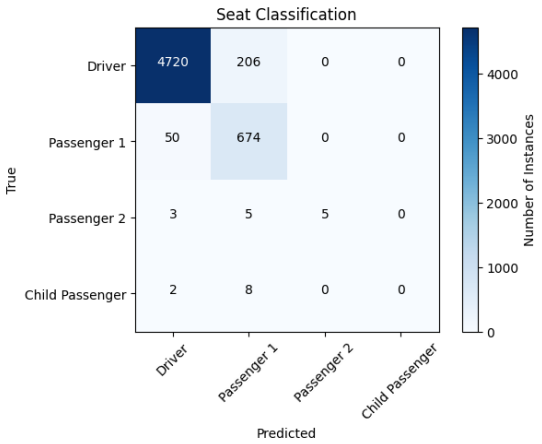

Fig. 7: Confusion Matrix of Seating Position Classification

accurately. With further fine-tuning and training, zero-shot learning has a strong potential for accurately handling real-world data.

Several sources of error are demonstrated in the helmet classification. The ground truth bounding boxes do not always encompass the whole person. Many boxes were taken from the lower half of their body as they entered the frame of the video. Additionally, overlapping bounding boxes with passengers and drivers, where drivers have helmets on, but the passengers do not, often confuses the OWLV2, claiming that it had detected the helmet in both cases. This also impacts person detection, as the OWLv2 cannot detect the passenger due to the driver obstructing most of the passenger's body.

Furthermore, AlexNet and the OWLv2 foundation model were challenged when faced with 'real-world' noise-filled scenarios. Many of the videos provided in the dataset had very low resolutions, with blurred-out time stamps at the top left and bottom right obstructing the view of motorcycles. Data collected during the night further reduced visibility, as the headlights of motorcycles and cars create a blurry effect throughout the video. The regular poor conditions of fog or heavy air pollution compounded these factors, as shown in Figure 3. All of these various aspects made image detection and classification challenging and sometimes near impossible.

Future investigations are necessary to apply zero-shot learning in the real world. In this application, accurately detecting helmets will help to raise awareness as the detections will provide a more accurate measure of the frequency at which people do not wear helmets, as well as assist in enforcing the wearing of helmets. The ability to respond to unanticipated data is crucial for safety systems, as real-world scenarios often surpass the scope of any pre-existing dataset. Ongoing development and refinement of the model will be imperative to fully harness their potential in practical safety systems. Our future research will focus on enhancing the accuracy, robustness, and consistency of zero-shot learning models in our detections.

To handle noisy data, pre-training the OWLv2 on further diverse datasets will allow it to better handle uncertain detections. Furthermore, preprocessing the data will mitigate some of these issues. Moreover, a possible improvement is the further integration of AlexNet and OWLv2 for seat classification. A hybrid approach using these two models will involve ensemble methods to balance their strengths for a more accurate result [28].

Finally, we will address task-specific shortcomings. For example, at times, the OWLv2 model fails to get the bounding box over the whole person, specifically the head, which is especially crucial for this task. A primary focus will be improving the model's ability to localize and classify these critical areas accurately.

Despite current limitations and imbalances in data, this research shows the potential of foundation models and language-based prompting toward the zero-shot handling of important safety challenges. We address all components of the AI City Challenge Helmet Detection and Occupancy tasks, showing

possibilities for the OWL model to address the sub-tasks of detection and association of vehicles, their occupants, and safety state information. This application has the potential to extend upon I2V communication. The detections from the infrastructure point of view can be sent to the vehicle's egocentric perception in order to alert drivers of the presence of motorcycles for safer intersection driving.

REFERENCES

[1] N. Abdi, T. Robertson, P. Petrucka, and A. M. Crizzle, "Do motorcycle helmets reduce road traffic injuries, hospitalizations and mortalities in low and lower-middle income countries in africa? a systematic review and meta-analysis," *BMC public health*, vol. 22, no. 1, p. 824, 2022.

[2] C. D. F. d. Souza, J. P. S. d. Paiva, T. C. Leal, L. F. d. Silva, M. F. Machado, and M. D. P. d. Araújo, "Mortality in motorcycle accidents in alagoas (2001-2015): temporal and spatial modeling before and after the "lei seca"," *Revista da Associação Médica Brasileira*, vol. 65, pp. 1482–1488, 2020.

[3] L. Abedi and H. Sadeghi-Bazargani, "Epidemiological patterns and risk factors of motorcycle injuries in iran and eastern mediterranean region countries: a systematic review," *International journal of injury control and safety promotion*, vol. 24, no. 2, pp. 263–270, 2017.

[4] A. L. Cavalcanti, B. Lucena, I. S. Rodrigues, A. L. Silva, T. T. Lima, and A. F. C. Xavier, "Motorcycle accidents: morbidity and associated factors in a city of northeast of brazil," *Tanzania journal of health research*, vol. 15, no. 4, 2013.

[5] A. Jayaraman, J. Padmanaban, J. Kakadiya, R. Rajaraman, M. Patel, and A. M. Hassan, "Helmet effectiveness in reducing serious/fatal injuries to motorised two-wheeler riders in india,"

[6] S. VASAN and G. GURURAJ, "Unhelmeted two-wheeler riders in india," *The National Medical Journal of India*, vol. 34.

[7] "The motor vehicles act, 1988."

[8] S. Wang, D. C. Anastasiu, Z. Tang, M.-C. Chang, Y. Yao, L. Zheng, M. S. Rahman, M. S. Arya, A. Sharma, P. Chakraborty, S. Prajapati, Q. Kong, N. Kobori, M. Gochoo, M.-E. Otgonbold, G. Batnasan, F. Alnajjar, P.-Y. Chen, J.-W. Hsieh, X. Wu, S. S. Pusegaonkar, Y. Wang, S. Biswas, and R. Chellappa, "The 8th AI City Challenge," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2024.

[9] R. Greer, S. Desai, L. Rakla, A. Gopalkrishnan, A. Alofi, and M. Trivedi, "Pedestrian behavior maps for safety advisories: Champ framework and real-world data analysis," in *2023 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1–8, IEEE, 2023.

[10] A. Ghita, B. Antoniussen, W. Zimmer, R. Greer, C. Creß, A. Møgelmose, M. Trivedi, and A. C. Knoll, "Activeanno3d-an active learning framework for multi-modal 3d object detection," in *35th IEEE Intelligent Vehicles Symposium (IV) 2024*, 2024.

[11] R. Greer, B. Antoniussen, M. V. Andersen, A. Møgelmose, and M. M. Trivedi, "The why, when, and how to use active learning in large-data-driven 3d object detection for safe autonomous driving: An empirical exploration," *arXiv preprint arXiv:2401.16634*, 2024.

[12] M. Minderer, A. Gritsenko, and N. Houlsby, "Scaling open-vocabulary object detection," 2023.

[13] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, pp. 1–21, 01 2019.

[14] R. Greer and M. Trivedi, "Towards explainable, safe autonomous driving with language embeddings for novelty identification and active learning: Framework and experimental analysis with real-world data sets," *arXiv preprint arXiv:2402.07320*, 2024.

[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[16] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, *et al.*, "Simple open-vocabulary object detection," in *European Conference on Computer Vision*, pp. 728–755, Springer, 2022.

[17] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, pp. 213–229, Springer, 2020.

[18] R. Greer, J. Isa, N. Deo, A. Rangesh, and M. M. Trivedi, "On salience-sensitive sign classification in autonomous vehicle path planning: Experimental explorations with a novel dataset," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 636–644, 2022.

[19] R. Greer, A. Gopalkrishnan, N. Deo, A. Rangesh, and M. Trivedi, "Salient sign detection in safe autonomous driving: Ai which reasons over full visual context," *arXiv preprint arXiv:2301.05804*, 2023.

[20] M. Naphade, S. Wang, D. C. Anastasiu, Z. Tang, M.-C. Chang, Y. Yao, L. Zheng, M. S. Rahman, M. S. Arya, A. Sharma, Q. Feng, V. Ablavsky, S. Sclaroff, P. Chakraborty, S. Prajapati, A. Li, S. Li, K. Kunadharaju, S. Jiang, and R. Chellappa, "The 7th ai city challenge," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2023.

[21] D. N.-N. Tran, L. H. Pham, H.-J. Jeon, H.-H. Nguyen, H.-M. Jeon, T. H.-P. Tran, and J. W. Jeon, "Robust automatic motorcycle helmet violation detection for an intelligent transportation system," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 5341–5349, June 2023.

[22] S. Cui, T. Zhang, H. Sun, X. Zhou, W. Yu, A. Zhen, Q. Wu, and Z. He, "An effective motorcycle helmet object detection framework for intelligent traffic safety," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 5470–5476, June 2023.

[23] J. Ouyang-Zhang, J. H. Cho, X. Zhou, and P. Krähenbühl, "Nms strikes back," *arXiv preprint arXiv:2212.06137*, 2022.

[24] H. Vo, S. Tran, D. M. Nguyen, T. Nguyen, T. Do, D.-D. Le, and T. D. Ngo, "Robust motorcycle helmet detection in real-world scenarios: Using co-detr and minority class enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 7163–7171, June 2024.

[25] Z. Zong, G. Song, and Y. Liu, "Detrs with collaborative hybrid assignments training," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6748–6758, 2023.

[26] Y. Chen, W. Zhou, Z. Zhou, B. Ma, C. Wang, Y. Shang, A. Guo, and T. Chu, "An effective method for detecting violation of helmet rule for motorcyclists," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 7085–7090, June 2024.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[28] R. Greer and M. Trivedi, "Ensemble learning for fusion of multiview vision with occlusion and missing information: Framework and evaluations with real-world data and applications in driver hand activity recognition," *arXiv preprint arXiv:2301.12592*, 2023.