# Large-Language-Model-Aided Object Detection

Anonymous ACL submission

## Abstract

## 1 Introduction

Object detection is a fundamental task in computer vision, with numerous applications in fields such as robotics, surveillance, and autonomous driving. Recent success in deep learning approaches has enabled steady and impressive progress in the performance of object detection. The progress also allows researchers to tackle more complex tasks along this line: from single object classification to dense object detection and from small, closed label set to large, open and long-tail label set. As we move on to solve denser, more complex image understanding tasks, it is important to take advantage of the visual context. As shown by various psycho-physics studies, human uses context and commonsense to understand visual scenes. We can easily deduce that a tiny object in the sky is not likely to be an elephant. However, most modern object detection pipelines adapt a two-stage design of proposing regions of interest and then classifying the regions into object classes, where neither context nor commonsense is taken into consideration. Without context and commonsense, detection is purely based on visual signals present in the proposed region, which might be sufficient for earlier benchmarks where only one or a few objects are present in each image. As we move on to tackle image with denser semantic information, it is crucial to take advantage of the context such as inter-object spatial relations, scene information, centerness, size, color, lighting, etc., which, combined with commonsense visual knowledge, can provide complementary information for detection.

Language models (Brown et al., 2020; Smith et al., 2022) are trained to predict the next word in a sequence given the context, and have achieved state-of-the-art results on a variety of natural language processing tasks, including commonsense reasoning.(Wei et al., 2022). Scraping web-scale textual data and continually trained with human feedback (Ouyang et al., 2022), large language models (LLMs) can now generate responses by following user instructions. These advances allow researchers to easily extract commonsense knowledge for downstream tasks with prompts.

In this paper, we propose a novel object detection method that leverages the power of large language models. Our approach uses a large language model to generate a description of the objects in an image and uses this description to refine object detections. By using a language model to provide context and refinement, our approach is able to take advantage of the rich linguistic knowledge and understanding of the world that these models have learned during training.

We evaluate our approach on standard object detection benchmarks and demonstrate that it achieves competitive results compared to state-of-the-art object detection methods. In addition, we show that our approach is able to generalize to novel object classes and scenarios, highlighting the potential of large language models for object detection.

## 2 Related Work

**Large Language Model** Large language models (LLMs) such as GPT-3 (Brown et al., 2020) and Turing (Smith et al., 2022) have gained significant attention in the natural language processing (NLP) community due to their ability to generate human-like text and perform a wide range of language tasks. Recently, the development of InstructGPT (Ouyang et al., 2022) has allowed LLMs to better follow human instructions, demonstrating their potential as reasoning agents. An interesting property of LLMs is their emergent ability (Wei et al., 2022), where they can perform tasks that they were not explicitly trained on, simply by leveraging their vast knowledge and understanding of language. This

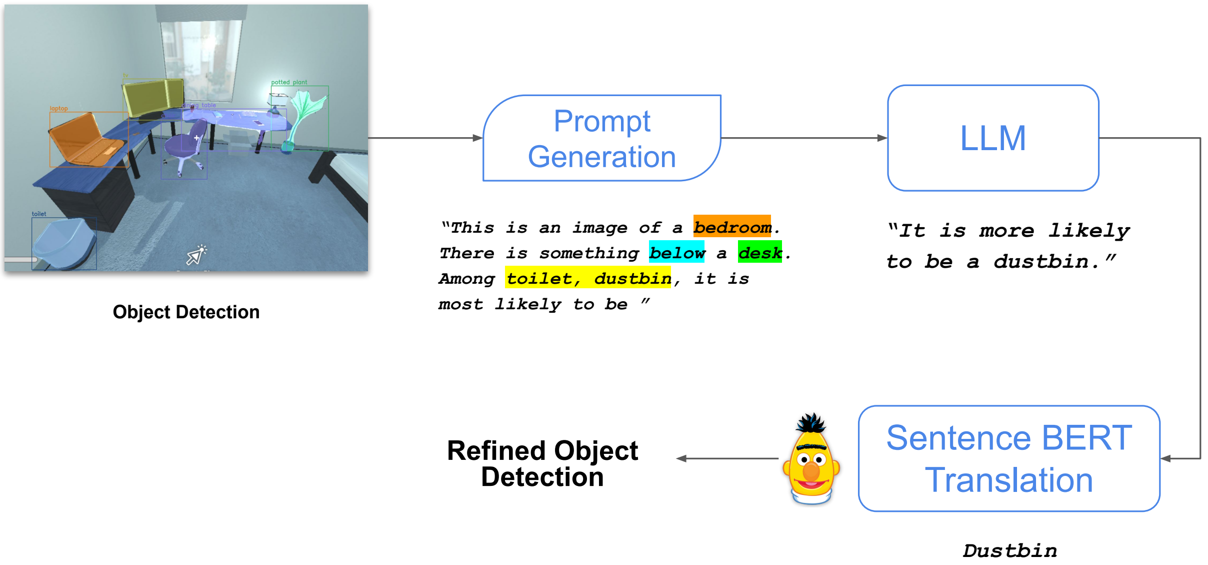
Object Detection
Prompt Generation
"This is an image of a bedroom. There is something below a desk. Among toilet, dustbin, it is most likely to be "
LLM
"It is more likely to be a dustbin."
Sentence BERT Translation
Dustbin
Refined Object Detection

Figure 1: Refinement pipeline overview. In the generated prompt, orange represents the scene; green represents the anchor objects; cyan represents the spatial relations; and yellow represents the candidate list.

makes them promising candidates for use as reasoning agents (Ahn et al., 2022; Huang et al., 2022a,b; Shah et al., 2022), capable of solving complex problems and making decisions based on their understanding of language and context.

**Vision Language Pre-training** Vision language pre-training is a recent trend in NLP that involves training large language models on a combination of visual and textual data. One example of this is Flamingo (Alayrac et al., 2022), which pre-trains a transformer-based language model on a combination of image-text pairs and textual data. Another example is GLIP (Li et al., 2022), which pre-trains a transformer-based language model on a combination of image and text data from the Internet. Both of these approaches have shown promising results in tasks that involve understanding the relationships between language and visual data.

**Injecting Commonsense into Object Detection** There has been recent work on injecting commonsense knowledge into object detection systems in order to improve their performance. DOCK (Singh et al., 2018)is an example of such a system, which utilizes a commonsense knowledge graph to provide context and aid in the detection of objects in images. By using context, DOCK is able to make more informed decisions about which objects are likely to be present in an image and where they are likely to be located. This type of approach (Rabinovich et al., 2007; Divvala et al., 2009; Lee and Grauman, 2010) has the potential to significantly improve the accuracy of object detection systems by leveraging the rich contextual information that is available in commonsense knowledge.

## 3 Method

In this section, we describe the proposed pipeline of using large language model to refine object detection.

### 3.1 Prompt Generation

To leverage the commonsense knowledge existed in LLMs, we need to construct prompts that captures the rich context information in images.

**Scene Description** Scene information provides strong cues to object class. A toilet is unlikely to appear in a kitchen and a coffee machine is unlikely to appear in a bathroom. For objects that are occluded or blurry, human can often leverage scene context to associate objects with correct classes. To help LLM take advantage of scene information, we inject scene description into the prompt. During inference, the CLIP model is used to rank similarities between images and pre-defined scene descriptions.

**Anchor Object Selection** Often times, the class of an object highly correlates with its neighboring objects in the same image. For example, a small object that is in the middle of a person and a coffee machine is likely to be a coffee mug. To help LLM reason about an object with interest by leveraging inter-object relations, we need to select a list of anchor objects to provide to the prompt. During inference, we partition the original detector predictions into trustworthy and untrustworthy sets based on their original prediction scores, sizes as well as distances to the center of the image. Within each prompt, we select the k nearest trustworthy neighbours as the anchors to be included.

**Spatial Relation** Spatial relations matter for determining object class. A dustbin is more likely to be below a desk than above; a coffee mug is more likely to be above a countertop than below. During inference, the spatial relation of objects is determined based on the position of the center of anchor boxes. Based on the angles formed by delta x, y of the centers, for each of the k selected anchor object, we infer about the relative spatial relation between the object and each anchor objects to be one of {"to the left of", "to the right of", "above", "below", "near"}.

**Candidates List** Having a short list of highly probable object class candidate for LLM to select from can help increase accuracy. We experiment with two different methods. The first is to crop out predicted regions in the image and query the CLIP model to re-predicts each individual object. This method re-evaluate each bounding box region independently without considering the context of the whole image. The other method we propose, thus, save the predictions for all class for each region of interest during the forward inference pass of the state-of-the-art object detectors, and retrieve the top predictions with high prediction scores among the classes for each region.

### 3.2 Response Translation

Mapping the response from the LLMs back to the label space is critical for quantitative evaluation. We use the Sentence-BERT model to choose the refined label based on sentence similarity between raw response and label class names.

## 4 Experiments

To evaluate the effectiveness of this approach, we conducted a series of experiments on two object detection datasets: MS-COCO and PASCAL VOC. In the following sections, we describe the details of our experimental setup and present the results of our experiments.

### 4.1 Dataset

**MS-COCO** The MS-COCO (Common Objects in Context) dataset is a widely used dataset for object detection and other computer vision tasks. It consists of over 200,000 images, each annotated with over 80 object categories and captions describing the image content. In our experiments, we used a subset of the MS-COCO dataset consisting of 5,000 images for validation. We chose this dataset due to its large number of annotated images and diverse set of object categories, which provides a challenging yet realistic testbed for evaluating the performance of object detection systems.

**PASCAL VOC** The PASCAL VOC (Visual Object Classes) dataset is another widely used dataset for object detection and other computer vision tasks. It consists of over 20,000 images, each annotated with one or more object categories. In our experiments, we used the PASCAL VOC 2007 and 2012 datasets, which consist of 4952 images for validation. We chose the PASCAL VOC dataset due to its wide variety of object categories and the availability of both detection and segmentation annotations. This dataset provides a more challenging testbed for evaluating the performance of object detection systems due to its smaller size and less diverse set of object categories compared to MS-COCO.

### 4.2 Evaluation

## 5 Results

We see 0.5% improvement in MS-COCO dataset (Table 1) and 0.1% in PASCAL VOC dataset (Table 2) for the Faster R-CNN models. We are still tuning the hyperparameters for better performance. Evaluations on other SOTA object detectors are yet to be conducted.

| | MS-COCO | LM-A-MS-COCO |
|---|---|---|
| Faster R-CNN | 47.3 | 47.8 |
| Mask R-CNN | | |
| YOLOv3 | | |

Table 1: mAP of our LM-Aided refinement of SOTA detectors on MS-COCO.

## 6 Discussion

Our pipeline works better on datasets that has denser objects, which aligns with our assumption in the way that placing object detections in context

|  | PASCAL VOC | LM-A-PASCAL VOC |
|---|---|---|
| Faster R-CNN | 37.2 | 37.3 |
| Mask R-CNN |  |  |
| YOLOv3 |  |  |

Table 2: mAP of our LM-Aided refinement of SOTA detectors on PASCAL VOC.

using language could boost the visual performance. We also notice many failure cases emerging due to wrong predictions of spatial relations. Most often the time the spatial relation of objects in the image is not exactly the same as the spatial relation of bounding boxes due to various vision perspectives. A wrongly predicted spatial relation in the prompt could confuse the LLM and negatively influence the performance.

## 7 Conclusion

In this paper, we propose a novel object detection method that leverages the power of large language models to generate a description of the objects in an image and uses this description to refine object detections. By using a language model to provide context and refinement, our approach is able to take advantage of the rich linguistic knowledge and understanding of the world that these models have learned during training. Our approach achieves competitive results compared to state-of-the-art object detection methods. In addition, we show that our approach is able to generalize to novel object classes and scenarios, highlighting the potential of large language models for object detection.

## References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, and Mengyuan Yan. 2022. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. ArXiv:2204.01691 [cs].

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. ArXiv:2204.14198 [cs].

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. ArXiv:2005.14165 [cs].

Santosh K. Divvala, Derek Hoiem, James H. Hays, Alexei A. Efros, and Martial Hebert. 2009. An empirical study of context in object detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1271–1278.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022a. Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents. ArXiv:2201.07207 [cs].

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. 2022b. Inner Monologue: Embodied Reasoning through Planning with Language Models. ArXiv:2207.05608 [cs].

Yong Jae Lee and Kristen Grauman. 2010. Object-graphs for context-aware category discovery. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8, San Francisco, CA, USA. IEEE.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. 2022. Grounded Language-Image Pre-training. ArXiv:2112.03857 [cs].

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. ArXiv:2203.02155 [cs].

Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge Belongie. 2007.

Objects in Context. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. ISSN: 2380-7504.

Dhruv Shah, Blazej Osinski, Brian Ichter, and Sergey Levine. 2022. LM-Nav: Robotic Navigation with Large Pre-Trained Models of Language, Vision, and Action. ArXiv:2207.04429 [cs].

Krishna Kumar Singh, Santosh Divvala, Ali Farhadi, and Yong Jae Lee. 2018. DOCK: Detecting Objects by Transferring Common-Sense Knowledge. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, volume 11217, pages 506–522. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. 2022. Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model. ArXiv:2201.11990 [cs].

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. ArXiv:2206.07682 [cs].

# A Appendix