Title

Open Vocabulary Part Grounding in Multimodal Large Language Models

Author

Sinha, Raunak

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Open Vocabulary Part Grounding in Multimodal Large Language Models

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Computer Science

by

Raunak Sinha

2024

ABSTRACT OF THE THESIS

Open Vocabulary Part Grounding in Multimodal Large Language Models

by

Raunak Sinha
Master of Science in Computer Science
University of California, Los Angeles, 2024
Professor Nanyun Peng, Chair

We investigate the complexities involved in open-vocabulary part segmentation, a task significantly more challenging than bounding box detection due to the precision and granularity required at the pixel level. Bounding box prediction allows for broad object localization, whereas segmentation demands exact delineation of object boundaries, making it a more difficult and nuanced task. Moreover, part grounding introduces an additional layer of complexity compared to object grounding. While object grounding relies on recognizing the whole object, part grounding requires understanding the intricate relationships between the object's components, each of which may have distinct shapes and functions. Through an extensive evaluation of multiple models, including DesCo, LISA, and VLPart, we assess their performance across both tasks on the PACO dataset, revealing their limitations in addressing the intricacies of part segmentation.

The results indicate that DesCo PACO (+ve) achieves the highest performance in bounding box detection, with an Average AP of 23.37, despite being trained without descriptive input. This highlights the relative simplicity of object grounding, where coarse localization is sufficient to achieve high accuracy. However, part grounding and segmentation models, such as LISA Fine-tuned, exhibit significantly lower performance, with an Average AP of 13.4, underscoring the additional complexity of localizing and identifying object parts. Part segmentation demands a deeper understanding of the object's internal structure and functional distinctions, which current models struggle to

consistently capture.

Furthermore, this study explores the impact of descriptive input on model performance. While descriptive augmentation has minimal effect on bounding box detection, it significantly improves segmentation accuracy. Descriptions provide critical context that enables models such as LISA Description to differentiate between visually similar parts, resulting in an improved Average AP of 16.3. However, despite these improvements, models like VLPart continue to struggle with generating accurate part masks, often producing irrelevant or misaligned predictions. These challenges are compounded by the need to resolve ambiguities between part categories and to capture parts that vary greatly in scale, shape, and context, making part grounding inherently more difficult than object grounding.

The thesis of Raunak Sinha is approved.

Bolei Zhou

Kai-Wei Chang

Nanyun Peng, Committee Chair

University of California, Los Angeles

2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

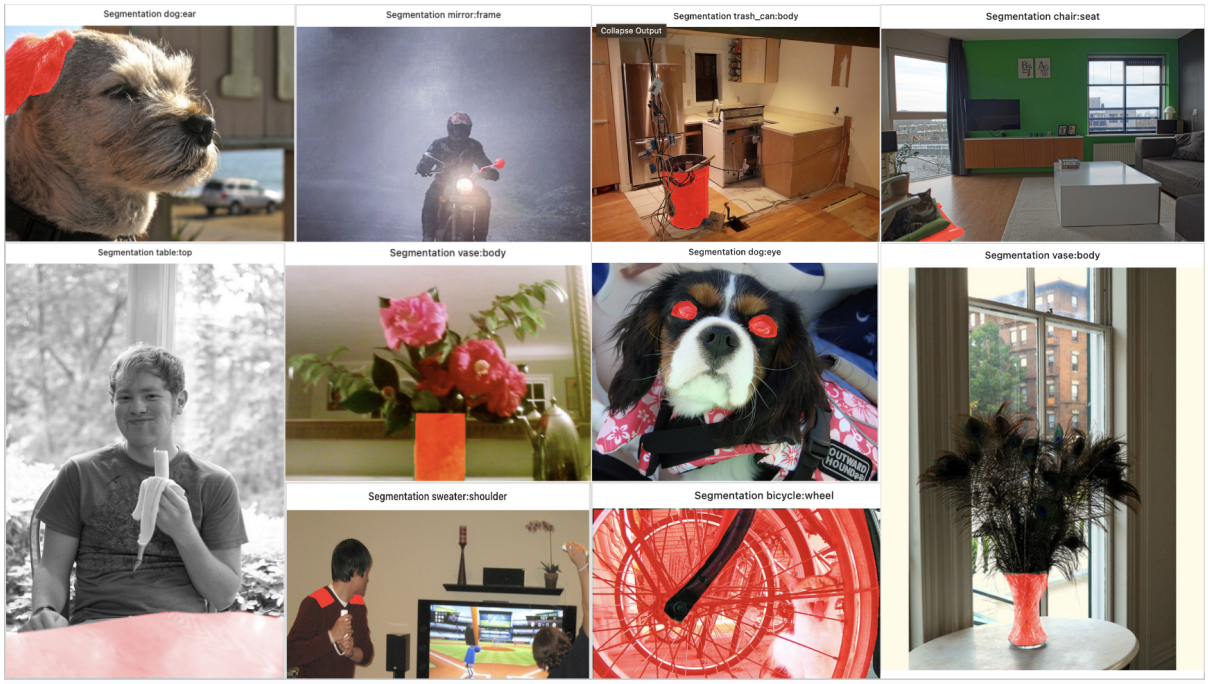


Figure 1.1: Images with part-segments annotated from the PACO dataset. The different scenes show the complexity of the task in terms of diversity of parts in multi-object scenes.

(a) "Blade of Grass" vs "Blade of Sword"

(b) "Face of Clock" vs "Face of Horse"

Figure 1.2: Showcasing difference between similarly named parts across two different objects

Real-world tasks increasingly demand a fine-grained understanding of objects, extend-

ing beyond mere object attributes and descriptions to part-level comprehension. Current grounding models, such as DesCo, demonstrate high proficiency in recognizing and grounding objects within images in an open-vocabulary setting. However, this capability does not naturally extend to part-level understanding, which presents its own set of challenges (Figure 1.1).

Part-level understanding is complicated by the variability in the meaning of part names across different object types. For instance, the "face" of a clock is vastly different from the "face" of a horse, and the "blade" of grass is visually distinct from a "blade" of a knife. This issue is less pronounced in object grounding, where each concept (e.g., clock vs. horse) is distinctly different (Figure 1.2). Existing approaches, such as those proposed by VLPart and LISA, treat each object part as a new class. This significantly increases the number of classes to learn, with limited annotations per class, and overlooks shared concepts across different objects. For example, while the "leg" of a dog differs from the "leg" of a chair, the concept of a "leg" used for standing is common to both. A global understanding of parts across objects would facilitate extending this notion to novel and unseen objects. These difficulties are compounded by the scarcity of labels for object parts both in terms of the number of annotations per part and the total number of objects with uniquely identified parts.

Object detection and segmentation have undergone substantial evolution, advancing from recognizing a limited set of categories to handling an open vocabulary of objects. Open-vocabulary systems are capable of identifying any object in the world, extending beyond fixed categories in existing datasets. These systems have greater real-world impact as they can recognize unseen categories expanding to detection cases that go beyond the limited corpus of labeled data. Despite these advancements, there remains a critical need for intelligent vision systems to understand and segment objects into their constituent parts in a detailed and flexible manner. Traditional models typically focus on object-level detection, but the ability to discern fine-grained details, such as the parts of an object, is essential for numerous applications, including robotics manipulation, behavior analysis, and detailed image editing.

It's tempting to apply existing open-vocabulary object detection methods to part-level recognition since parts constitute an object. This is popularly achieved by replacing weights from the final classifier layer with the text-embedding from the part-category name. However, these methods do not generalize well to part-level recognition. Existing approaches to open-vocabulary object detection struggle with part segmentation due to their reliance on object-level data, which often lacks the granularity required for part-level recognition. The primary challenges are twofold: (1) The recall issue, where models trained on object-level data fail to generalize to the finer granularity of parts, and (2) The precision issue, where insufficient part-level annotations hinder effective part segmentation.

The model for open-vocabulary part segmentation is designed to segment objects not only by their general categories but also by finer granularity. For example, a dog can be segmented into parts like the head, torso, legs, and tail, and further into more detailed parts such as the ear, eye, and nose. Annotating these detailed object parts is highly expensive, and the available datasets for part segmentation are less comprehensive and diverse compared to those for image classification and object detection. Despite gathering data from three sources—Pascal Part, PartImageNet, and PACO—only a limited number of object parts are accessible. Each dataset defines its own set of parts for the same object, making the problem more complex as what visibility of part should be the most acceptable?

While existing object detection methods usually use image captioning data for model training, they rely on novel methodologies to construct alignment between objects in the image and their captions, as these datasets only provide images and captions without dense alignment between text and objects. For datasets with dense caption alignment between objects and text, the frequency of parts in contrast to objects is skewed towards objects, making part-level segmentation more challenging.

To address these challenges, it is essential to develop models that can learn part-grounding with limited data and expand this capability to open-vocabulary settings for both objects and parts. Despite the nascent state of parts/attributes grounding, state-of-the-art (SOTA) segmentation models have shown proficiency in understanding

parts across a range of types. Most of these models explore open-vocabulary in terms of new object types, but rarely are new part types studied. For instance, based on human prior, dogs and cats can be thought to have similar parts of head, body, tail, nose, ears, neck etc. But existing datasets do not consider the possibility of a new part. This might be a problem if a concept of a part like "wing" for example was never seen before, which is an integral part of objects like birds. To enhance the part-grounding capability of models like DesCo, it is necessary to train on comprehensive datasets, experiment with various training strategies, and employ methods like SAM (Segment Anything Model) to generate captions for segments and perform DesCo-style training.

To enhance the part-grounding capability of DesCo, we train the model on the PACO dataset, which encompasses bounding boxes for 75 different object categories along with 456 object part-categories and 55 attributes. This dataset provides a rich and diverse set of annotations crucial for developing a nuanced understanding of object parts.

In our experimentation, we are employing various training strategies for fine-tuning DesCo:

- **Object-Agnostic Part Classes:** Instead of treating each "part of object" as a unique class, we adopt a strategy where "part" is considered a unique class, independent of the object it belongs to. For instance, the "leg of a bird" and the "leg of a chair," while visually distinct, are conceptually similar. By training a model solely on parts without object-class mentions, we aim to generalize the concept of parts across different objects, enhancing the model's ability to recognize parts in novel objects.

- **Descriptive Training for Generalization:** Traditional captioning methods do not facilitate generalization to an open-vocabulary setting. Therefore, similar to DesCo's training style, we employ contrastive descriptions of object parts. This approach involves using detailed descriptions to learn about parts, thereby improving the model's ability to generalize from descriptive text to visual parts.

By employing these diverse and innovative training strategies, we aim to significantly

improve the part-grounding capabilities of DesCo, making it more adept at handling the complexities of part-level segmentation in an open-vocabulary context.

CHAPTER 2

Analysis of current models

2.1 Related work

2.1.1 Image segmentation

Image segmentation, a fundamental task in computer vision, assigns a class label to every pixel in an image. Early techniques relied on manual feature extraction and traditional algorithms like thresholding and clustering. The advent of deep learning revolutionized segmentation methods, with Convolutional Neural Networks (CNNs) (Lecun et al., 1998) becoming pivotal. Notable architectures include Fully Convolutional Networks (FCNs) (Long et al., 2015) and U-Net (Ronneberger et al., 2015), which utilize encoder-decoder structures for dense pixel predictions and precise localization. Research has further advanced with methods like dilated convolutions (Yu and Koltun, 2016), pyramid pooling modules (Zhao et al., 2017), and non-local operators (Wang et al., 2018) enhancing semantic information encoding. Instance and panoptic segmentation have introduced innovations like DETR-based structures (Carion et al., 2020), mask attention (Endo, 2023), and dynamic convolution for instance-level segmentation. Recently, models like SAM (Kirillov et al., 2023) and X-Decoder (Zou et al., 2023) have demonstrated exceptional segmentation quality and multi-task compatibility, while SEEM (Zou et al., 2024) supports diverse human interactions. These advancements continue to push the boundaries of detailed and accurate image analysis, making segmentation a continually evolving area of research.

2.1.2 Semantic Correspondence

The aim of semantic correspondence is to establish spatial visual correspondences between different instances of the same object category, enhancing the ability to recognize and relate visual features across varied contexts. This concept is essential in tasks such as object detection, segmentation, and image captioning, where understanding the relationship between visual and textual data is crucial. Early works focused on leveraging pre-trained models to compute feature map similarities, with methods using a pre-trained CNN to establish these correspondences. Performance was further improved by employing Vision Transformers (ViT) (Dosovitskiy et al., 2020) to enhance the representation capabilities, demonstrating superior results in cross-domain correspondence tasks. Recent advancements have seen the application of self-supervised models like DINO (Caron et al., 2021), which leverage self-supervised learning to generate robust feature representations. DINO facilitates the alignment of novel objects with base objects without requiring extensive labeled data, significantly improving the model's ability to generalize across different object categories by establishing more accurate and detailed correspondences (Sun et al., 2023). Transformer-based models, such as CLIP (Radford et al., 2021), have further advanced the field by aligning images and text in a shared embedding space using a contrastive learning approach. Further, models like OSCAR (Li et al., 2020) integrate object tags into the learning process, enhancing the model's ability to establish semantic correspondences between objects in images and their textual descriptions. UNITER (Chen et al., 2020) and VinVL (Zhang et al., 2021) have introduced sophisticated methods for integrating visual and textual information by employing large-scale pre-training on diverse datasets, enabling them to capture richer and more nuanced semantic correspondences.

2.1.3 Vision-and-language representation learning

Vision-language representation learning has advanced significantly, aiming to unify visual and textual understanding for tasks like visual question answering (Goyal et al., 2017; Gurari et al., 2018; Agrawal et al., 2016; Srivastava et al., 2020; Malinowski et al.,

2018)(VQA), image captioning (Xu et al., 2015; Vinyals et al., 2015; Rennie et al., 2017; Anderson et al., 2018; Sharma et al., 2018), and cross-modal retrieval (Frome et al., 2013; Faghri et al., 2018; Lee et al., 2018). Central to this progress are contrastive learning techniques, as seen in models like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), which align image-text pairs, enabling open-vocabulary generalization beyond fixed categories. Models such as UNITER (Chen et al., 2020) and VinVL (Zhang et al., 2021) have pushed this further by leveraging pre-training on large-scale datasets, aligning fine-grained object and language correspondences, improving performance in reasoning tasks involving object attributes and spatial relationships. With the integration of large language models (LLMs) in models like Flamingo (Alayrac et al., 2022) and LISA (Lai et al., 2024), vision-language systems have evolved to handle more complex, context-aware tasks, demonstrating the potential for deeper reasoning and broader real-world applications.

2.1.4 Multimodal large language model

Multi-modal large language models (LLMs) integrate textual and visual data to enhance AI comprehension and generation capabilities. Models like CLIP (Radford et al., 2021) use contrastive learning to align images with text, excelling in zero-shot learning tasks without extensive fine-tuning. Inspired by LLMs' reasoning abilities, researchers have developed models such as Flamingo (Alayrac et al., 2022), which uses cross-attention (Vaswani et al., 2017) for visual in-context learning, and BLIP-2 (Li et al., 2023b) and mPLUG-OWL (Ye et al., 2023), which encode image features with a visual encoder before feeding them into the LLM. Otter (Li et al., 2023a) incorporates few-shot capabilities through in-context instruction tuning, while LLaVA (Liu et al., 2024) and MiniGPT-4 (Zhu et al., 2023) align image-text features followed by instruction tuning. VisionLLM (Wang et al., 2024) and Kosmos-2 (Peng et al., 2023) enhance interaction and grounding capabilities, respectively, while DetGPT (Pi et al., 2023) bridges multi-modal LLMs and open-vocabulary detectors for detection based on user instructions. These advancements, supported by vision-language pre-training, transform tasks like image captioning (Xu

et al., 2015; Vinyals et al., 2015; Rennie et al., 2017; Anderson et al., 2018)and visual question answering (Goyal et al., 2017; Gurari et al., 2018), making multi-modal LLMs versatile tools for complex content generation and understanding.

2.1.5 Open-vocabulary object detection

Open-vocabulary object detection (OVOD) extends the capabilities of object detection models beyond fixed categories to recognize and localize novel objects described by arbitrary text inputs. This addresses the limitations of traditional models confined to a limited number of seen categories during training. Models like ViLD (**?**), RegionCLIP (Zhong et al., 2021), and PB-OVD (Gao et al., 2022) leverage pseudo region annotations from pre-trained vision-language models to enhance detection capabilities, with ViLD using knowledge distillation to align region embeddings with text and image embeddings. RegionCLIP refines regional features with contrastive learning, improving generalization to unseen categories. DetPro (Du et al., 2022) enhances category embeddings through automatic prompt learning, and GLIP integrates detection and grounding data to improve the alignment of textual descriptions with visual regions. Detic (Zhou et al., 2022b) expands novel class recognition using image classification data, and VLDet (Lin et al., 2022) continuously extracts region-word pairs from image-text pairs, this is seen as a set-matching problem between image-regions and word-embeddings. OVOD faces challenges in data scalability and integration of high-quality language models, but these advancements pave the way for more adaptable vision systems capable of detailed object and part recognition in complex scenarios.

2.1.6 Part Segmentation

Part-segmentation has emerged as an essential task in computer vision, focusing on segmenting objects into their constituent parts, beyond traditional object-level detection. Early work such as Pascal-Part pioneered part-annotated datasets, enabling initial advancements in this domain but remained limited in category diversity and granularity.

PartImageNet further expanded the scope by offering a broader dataset, but fine-grained hierarchies of parts remain a challenge. Recent efforts like VLPart (Sun et al., 2023), LISA (Lai et al., 2024), and DesCo (Li et al., 2023c) introduced vision-language models that improve generalization to unseen objects by leveraging semantic correspondence and complex language reasoning, though their specific mechanisms are discussed later. Despite progress, the lack of extensive, detailed part-level annotations across diverse categories continues to limit part-segmentation, driving the need for more robust cross-domain generalization methods to bridge the gap between object and part-level understanding.

2.2 Model breakdown

For making targeted improvements we first need to understand the best models for part-segmentation extension. We discuss three models DesCo, VLPart, and LISA. The discussion is divided into four sections: Method Breakdown, Data Details, Metrics, Result for each model.

2.2.1 DesCo

2.2.1.1 Model Breakdown

Traditional models trained to recognize a fixed set of categories often fail to generalize to novel concepts or domains, limiting their effectiveness in real-world applications. Recent advancements using contrastive objectives on expansive image-text datasets, such as CLIP and GLIP, have produced foundation models adept at various downstream tasks (Shen et al., 2021; Zhou et al., 2022a; Li et al., 2022a; Gao et al., 2021; Cho et al., 2021). However, these models typically rely on language queries that primarily use object names, missing out on the rich, descriptive information necessary for precise object identification. Descriptive queries, which include attributes, shapes, textures, and relationships, can significantly enhance model performance, allowing them to recognize novel classes through detailed descriptions. For instance, an “axe" can be described as a “long handle, with

a sharp blade, used for cutting wood," providing a comprehensive context that simple object names lack. Using object descriptions during training time enables the model to extend to unseen novel objects during test time, this extension is possible because the model learns to rely on object description for identification which can be compositional.

The primary bottleneck in current approaches is the lack of fine-grained descriptions in image-captioning data, which often results from reporting bias, as humans tend to mention entities rather than providing detailed descriptions. Moreover, models are not incentivized to understand these descriptions fully. In a contrastive setting, models tend to align positive phrases with relevant regions while suppressing negatives, but even with such an alignment the model focuses directly on the entity name and bypasses the descriptions. DesCo's focus is on paying higher attention to descriptions rather than entities (Li et al., 2023c).

Furthermore, prior models frequently treat queries as a"bag-of-words", leading to hallucinations due to inconsistencies in training formulations. They struggle with complex queries and fail to interpret descriptive information accurately. DesCo addresses these limitations by leveraging detailed, context-rich language descriptions, enabling the model to handle complex queries and novel concepts more effectively. This approach ensures that models are better equipped to understand and utilize detailed descriptions, significantly enhancing their generalization capabilities and performance in real-world scenarios.

The ablation study for DesCo explores several key modifications to enhance the model's performance. Directly appending descriptions to queries without context-sensitive construction showed no improvement in performance for rare categories and minimal change in context sensitivity, indicating the model's difficulty in effectively utilizing non-contextual descriptions. In contrast, removing the entity name significantly enhanced both contextual sensitivity and object detection, emphasizing the importance of focusing on contextual information. Incorporating hard negative captions further boosted detection accuracy by helping the model grasp subtle nuances in language descriptions. Additionally, using higher-quality language models like those from the GPT family markedly improved detection performance, highlighting the value of robust pre-trained language models for

embedding rich visual information. These findings collectively demonstrate the critical role of context, strategic input modifications, and high-quality language models in optimizing DesCo's performance.

DesCo extends the training of GLIP (Li et al., 2022b) by using expansive description for objects in any given scene. As GLIP is the foundation of DesCo it is imperative to understand the core principle of GLIP (Grounded Language-Image Pretraining), which is to transform any task-specific, fixed-vocabulary classification problem into a task-agnostic, open-vocabulary vision-language matching problem. This approach unifies training data into a grounding format, where each setup includes an image, a text query, bounding boxes, and ground-truth alignment labels. For detection data, the text query is a concatenation of object classes, encompassing both positive and negative examples. For grounding data, which involves image captions with region annotations, GLIP creates complex queries by combining multiple captions from a single image. For image-caption pairs lacking bounding boxes, pseudo labels are generated using a grounding model.

Detection is viewed as language-context-free grounding, while grounding is considered language-context-dependent detection. The model computes an alignment score between image regions and words.

During training, GLIP optimizes for region-word matching loss and localization loss. This unified approach enables the model to perform well across various vision-language tasks, leveraging both visual and textual information to improve accuracy and generalization. By aligning regions with descriptive language, GLIP effectively bridges the gap between visual content and its textual description, enhancing its capability to understand and process complex queries.

2.2.1.2 Dataset Creation

The aim of this work is to enrich object detection models with complex, description-rich language. This is achieved through a novel description-conditioned contrastive training paradigm, where large language models, imbued with extensive world knowledge, generate

detailed descriptions. The models are prompted with questions like “What features should object detection models focus on for an entity in the caption?” to produce nuanced descriptions.

Additionally, the focus is on “context-sensitive queries” where negative and positive phrases can only be distinguished through detailed descriptions. To identify any particular object a perfect match to the description needs to be established. Negative descriptions are created by altering the descriptions slightly. To create challenging scenarios, “Winograd-like" queries are constructed to produce confounding descriptions. The original grounding task is generalized to allow for fully negative queries to enhance robustness.

This approach addresses key observations:

- Entity Shortcuts: Models often learn to rely solely on entity names, ignoring the surrounding context. In such cases, the mutual information of the context and ground-truth, given the entity and image, is nearly zero. Thus, simply augmenting the context with more detailed descriptions does not add value.

- Hallucinations: Traditional phrase grounding methods focus on finding entities in captions that are always present, which can lead to hallucinations. By introducing context-sensitive queries and negative examples, the model is encouraged to pay attention to the full description, reducing the reliance on entity shortcuts and minimizing hallucinations.

Through this meticulous dataset construction process, the models are trained to better understand and utilize detailed, context-rich descriptions, leading to more accurate and robust object detection capabilities.

2.2.2 Metrics

The authors measure model performance using two key metrics: the first being Average Precision (AP), which assesses accuracy, and the second set comprising ΔBox and ΔConf. ΔBox quantifies the variation in bounding box coordinates, while ΔConf evaluates changes

in the alignment score for these boxes.

2.2.2.1 Result

The authors conduct a comprehensive evaluation of GLIP to assess its capacity for understanding descriptive input. They observe that incorporating descriptions during inference adversely impacts model performance. Additionally, the model appears to disregard the provided contextual information. This is evidenced by experiments where negative descriptions are added to entity names, revealing that as long as the entity is explicitly mentioned, the model's predictions remain largely unaffected. Contrary to expectations, the model shows neither stronger alignment with positive descriptions nor significant variation in predictions when presented with negative versus positive descriptions.

They also study the zero-shot transferability of the model across LVIS and OmniLabel, DesCo-GLIP shows an improvement of 8.6 AP over GLIP and for rare categories (APr) there is a 10.0 increase.

The authors conduct multiple ablation studies across different components in the model. Dropping the entity name improved the focus of the model on contextual information. Including hard negatives improves model detection across multiple datasets while ensuring robust contextual comprehension.

The novel learning paradigm introduced in this work employs language supervision to construct context-sensitive queries. While DesCo demonstrates the ability to comprehend complex queries within a contextual framework, extending this capability to parts of objects presents significant challenges. Utilizing language models for description generation can introduce noise, as the generated descriptions may lack accuracy or relevance. Moreover, successful deployment of DesCo requires careful prompt engineering to ensure precise interpretation. The application of the DesCo paradigm to object detection, which predicts bounding box coordinates around objects, is not directly translatable to part-level understanding. Bounding boxes are an inherently coarse method of localization,

and using them to delineate object parts can lead to significant overlap between distinct regions, resulting in ambiguous part definitions. Therefore, while the DesCo paradigm shows promise, its application may be more appropriate for image segmentation tasks, which offer finer granularity and are more aligned with the study of object parts.

2.2.3 VLPart

2.2.3.1 Method Breakdown

VLPart (Sun et al., 2023) is trained using multiple levels of information, integrating part-level, object-level, and image-level data. This comprehensive training strategy allows the model to achieve multi-level granular alignment between language and images. By leveraging datasets like Pascal Part, PartImageNet, and PACO, VLPart grows the existing set of classes with part-level annotations. It parses novel objects into their respective parts through dense semantic correspondence with known base objects. This dual-step approach enables VLPart to generalize across diverse data sources and foundational models.

VLPart addresses two levels of openness; open category and open granularity. Open category, a concept familiar in object detection, involves recognizing novel objects. Open granularity, however, deals with the hierarchical structure of object parts. For example, a human figure can be segmented into parts such as head, hands, and legs, with each part further divided into finer components like eyes, nose, and mouth. This hierarchical nature makes part-segmentation inherently challenging, as the appropriate level of detail varies with different objects and scenarios. To expand part categories, VLPart utilizes large vocabularies from object-level and image-level data sources, which lack part-level annotations. Training on different granularity helps the model align vision and language features at multiple levels. VLPart calculates alignment scores between class names and regions, a function that is applicable to both object-level and part-level tasks. The objective is to transfer this alignment capability from large-scale object detection to detailed part segmentation.

For generalization and data discovery, VLPart proposes a new data-annotation pipeline

that employs existing vision foundation models to parse unseen objects into their parts, creating correspondences between known objects and parts and novel, unseen objects as a form of data discovery. This is done due to the absence of part-level supervised data.

In summary, VLPart leverages a combination of convolutional and transformer-based image encoders, a multi-scale feature generation network, a sophisticated detection decoder that integrates text embeddings for flexible classification, and a mask decoder designed for open-vocabulary part segmentation. This comprehensive approach enables detailed and accurate segmentation of objects and their parts across a wide range of categories and granularities. The detection decoder comprises two main components: a region proposal network (RPN) and an R-CNN recognition head. The RPN generates box proposals outlining potential objects and parts within the image. These proposals are refined by the R-CNN recognition head, which adjusts the box locations and assigns classification scores.

Notably, instead of traditional classifier weights, the R-CNN recognition head uses text embeddings of object and part names. These embeddings, derived from the text encoder in CLIP, replace fixed classifier weights, enabling a more flexible and comprehensive classification process. The classification score is calculated through a dot-product operation between the region features, extracted from the feature maps generated by the image encoder, and the text embeddings. This mask decoder includes a class-agnostic head, allowing it to segment objects from novel categories not seen during training. This class-agnostic approach is crucial for open-vocabulary segmentation tasks, where the model must generalize to new object categories.

The training loss for VLPart includes location loss, classification loss, and mask loss for both part-level and object-level data, whereas image-level training only updates the classification loss. The assumption in parsing novel objects into parts is that novel objects will exhibit high similarity with the part taxonomies of base objects. For example, a knife and a sword share similar parts. By establishing dense semantic correspondence between regions in novel objects and parts of base objects, new regions can be pseudo-labeled, facilitating effective part segmentation even for unseen objects.

2.2.3.2 Data Details

The VLPart model is trained on various datasets, each providing different levels of annotations essential for detailed part segmentation. The detailed part annotations help refine part distinctions, the object detection data enhances localization and recognition capabilities, and the image-level classification data broadens the model's understanding of different categories.

- Part-Level Annotated Datasets: All part-level annotated datasets used in VLPart contain segmentation masks for parts and their respective categories. Typically, each part is defined as an object-part pair, allowing the model to discern between similar parts of different objects. For example,"tail of dog" and "tail of cat" are treated as separate categories.

- Object Detection Datasets: Object detection datasets used in VLPart contain objects along with their bounding boxes and, in some cases, their masks. Each object is also assigned a category label, which helps the model learn to identify and localize different objects within an image.

- Image-Level Classification Datasets: Image-level classification datasets provide labels and images of objects but lack bounding boxes. Despite this, these labeled samples are valuable for improving the classification loss by introducing a broader range of object categories. For instance, by considering the maximum-size proposals for an image, the model can incorporate more diverse categories, enhancing its ability to generalize and recognize various objects.

  VLPart uses a pretrained DINO for finding the nearest base object for each novel object and build the dense correspondence. Such correspondence show good alignment in terms of color, texture and pose but there is a often are at the cost of high semantic correspondence between such objects.

2.2.3.3 Metrics

To evaluate their methodology, the VLPart paper employs two levels of generalization: cross-category and cross-dataset generalization. Cross-category generalization involves training the model on seen categories (base parts) and testing it on unseen categories (novel parts), ensuring no overlap between them. This method assesses the model's ability to generalize to new, unseen categories, demonstrating robustness in recognizing and segmenting unknown parts. Cross-dataset generalization, on the other hand, trains the model on one dataset and evaluates it on a different one. Unlike cross-category generalization, class exclusivity between training and testing datasets is not guaranteed, closely simulating real-world applications and evaluating the model's practical utility and adaptability to diverse data.

The paper also uses mean average precision (mAP) to measure performance. mAP assesses the average precision across all classes and IoU thresholds, providing a comprehensive evaluation of detection and segmentation accuracy. Specifically, $mAP_{mask}@[0.5, 0.95]$ calculates the mean average precision for mask predictions across IoU thresholds from 0.5 to 0.95. This metric ensures a robust assessment of both coarse and fine segmentation tasks, validating VLPart's accuracy in various precision levels.

2.2.3.4 Result

They show that adding more image-caption data might marginally help the performance but is not sufficient for part-level performance. Directly introducing part information for training had a strong effect on the model performance, as can be seen with the improvement of 3.5 17.6 mAP across the quadrupeds class for a model trained on Pascal Part dataset. The also observe an average improvement of 3.3 mAP over the baseline across the 40 classes in the PartImageNet validation sent. There is a positive correlation to the increase in the number of part annotations dataset and performance.

2.2.4 LISA

2.2.4.1 Method Breakdown

The work (Lai et al., 2024) introduces an innovative task in computer vision known as reasoning segmentation. Unlike traditional segmentation tasks, which rely on explicit instructions or predefined categories to identify objects within an image, LISA (Large Language Instructed Segmentation Assistant) addresses more complex scenarios where instructions are implicit. This necessitates the model to reason and infer the user's intent based on context or world knowledge. For example, instead of directly identifying "the trash can", the query might be "something that the garbage should be put into", requiring a deeper understanding and reasoning capability.

Inspired by the reasoning capabilities of large language models (LLMs) and their adeptness at understanding user prompts, LISA utilizes this potential to advance visual segmentation tasks. While existing multimodal LLMs can process visual inputs, their primary focus has been on text generation, often neglecting the direct production of meaningful image features. LISA addresses this limitation by integrating the reasoning strengths of LLMs with segmentation, thereby extending the model's ability to handle complex, context-dependent queries with a deeper understanding of both language and visual information.

In discussing LISA, our emphasis is on its emergent capability to understand object parts, rather than its primary focus on reasoning segmentation tasks. However, the following contributions by the paper are noteworthy:

- Complex and Implicit Query Handling: LISA excels in managing intricate queries that require advanced reasoning, understanding nuanced descriptions and contextual clues to identify target objects or regions within an image.

- ReasonSeg Benchmark: The ReasonSeg benchmark was developed to evaluate model performance, consisting of 1218 image-instruction pairs from OpenImages and ScanNetv2. These pairs, annotated with short phrases and long sentences, test

the model's ability to handle varying levels of query complexity.

- Proficiency in Complex Reasoning: LISA can adeptly handle scenarios involving complex reasoning, world knowledge, explanatory answers, and multi-turn conversations, making it suitable for applications demanding contextual understanding and precise responses.

The "embedding-as-mask" paradigm is a novel approach introduced in the LISA model to enable multi-modal Large Language Models (LLMs) to generate fine-grained segmentation masks directly. Multi-modal LLMs, such as LLaVA, Flamingo, BLIP-2, and Otter, support image and text inputs and produce textual outputs but lack the capability to output detailed segmentation masks. VisionLLM offers a partial solution by parsing segmentation masks as sequences of polygons, allowing representation as plain text and enabling end-to-end training within existing multi-modal LLM frameworks. However, this method faces optimization challenges and may compromise generalization ability unless significant computational resources are employed. For example, training a 7B model in VisionLLM requires extensive GPU resources, making it computationally prohibitive.

LISA addresses these limitations by expanding the LLM's vocabulary to include a new token, <SEG>, which signals the need for segmentation output. Given a text instruction x_{txt} and an input image x_{img}, the model processes these inputs and generates a response that includes the <SEG> token. The last-layer embedding corresponding to this <SEG> token is extracted and transformed into a segmentation mask through a Multi-Layer Perceptron (MLP) projection layer. Concurrently, a vision backbone network, such as SAM or Mask2Former, extracts visual embeddings from the input image. These visual embeddings and the <SEG> token embedding are combined in a decoder to produce the final segmentation mask.

By integrating this embedding-as-mask technique, LISA significantly enhances the capability of multi-modal LLMs to perform sophisticated segmentation tasks, effectively bridging the gap between language understanding and visual segmentation. This method allows the model to leverage the strengths of both text and visual data, enabling it to

handle complex queries involving intricate reasoning and contextual understanding while maintaining computational efficiency. For instance, training LISA-7B requires only 10,000 training steps on 8 NVIDIA 24G 3090 GPUs, making it far more feasible compared to VisionLLM.

LISA's optimization is achieved through a weighted sum of the text-generation loss L^{txt} and the segmentation mask loss L^{mask}. The text-generation loss L^{txt} is an auto-regressive cross-entropy loss, which ensures the model's language outputs are coherent and contextually accurate. The segmentation mask loss L^{mask} enhances the quality of the generated segmentation masks. This loss combines per-pixel binary cross-entropy and DICE loss, which together help produce precise and high-quality segmentation results. The binary cross-entropy component focuses on the accuracy of each pixel classification, while the DICE loss addresses the overlap between predicted and ground truth masks, promoting better overall segmentation performance.

2.2.4.2 Data Details

LISA's training process incorporates three distinct types of datasets to ensure comprehensive segmentation capabilities. Semantic Segmentation Datasets such as ADE20K, COCO-Stuff, and LVIS-PACO involve images with multi-class labels. During training, categories are randomly selected for each image, and QA pairs are generated using templates like "USER: <IMAGE> Can you segment the CLASS NAME in this image? ASSISTANT: It is <SEG>." The binary segmentation mask corresponding to CLASS NAME serves as the ground truth, and various templates are used to ensure data diversity. Vanilla Referring Segmentation Datasets including refCOCO, refCOCO+, refCOCOg, and refCLEF provide images paired with explicit descriptions of target objects. These are converted into QA pairs such as, "USER: <IMAGE> Can you segment description in this image? ASSISTANT: Sure, it is <SEG>," where description specifies the target object. Visual Question Answering (VQA) Datasets like LLaVA-Instruct-150k, generated by GPT-4, are included to maintain the model's ability to handle diverse queries. This

dataset directly preserves the VQA capabilities of the multi-modal LLM, ensuring the model's versatility in responding to varied and complex instructions.

2.2.4.3 Metrics

The evaluation of LISA employs two key metrics: gIoU and cIoU. Generalized Intersection over Union (gIoU) calculates the average IoU across all images, providing a balanced measure of segmentation accuracy. In contrast, cumulative Intersection over Union (cIoU) measures the cumulative intersection over the cumulative union, which tends to be biased towards larger regions. This dual-metric approach ensures a comprehensive assessment of LISA's performance, capturing both average accuracy and performance on larger segments.

2.2.4.4 Results

LISA-13B demonstrates substantial improvements over LISA-7B, particularly in handling long queries. This enhancement underscores the potential for further advancements in understanding long dependencies within prompts. In the reasoning segmentation task, LISA achieves a 20% gIoU performance boost for complex reasoning tasks that require an understanding of world knowledge and reasoning abilities.

2.3 Dataset

Our experiments are focused primarily on the PACO dataset. We feel that this dataset has the highest number of part-categories across a diverse set of objects. Additionally, the scene from this dataset are more complex which is more realistic. But we also mention details of other datasets we considered as baseline for our work. These three datasets are the standard dataset for part-understanding. We observed that not all models train/evaluate on the same dataset, so that can be hard for direct comparisons.

2.3.1 PACO

The PACO (Ramanathan et al., 2023) (Parts and Attributes of Common Objects) dataset is a comprehensive resource designed to advance part-level understanding and attribute recognition in computer vision. It includes detailed annotations for 75 object categories, covering 456 object-part categories and 55 attributes. The dataset amalgamates data from LVIS (Gupta et al., 2019) for images and Ego4D (Grauman et al., 2022) for videos, providing a rich vocabulary and temporally aligned narrations to aid in sourcing frames for specific objects. The annotation pipeline encompasses object bounding boxes, part masks, and detailed attribute annotations, ensuring exhaustive coverage of both objects and their parts.

2.3.2 PASCAL-Part

The PASCAL-Part (Chen et al., 2014) dataset extends the PASCAL VOC 2010 dataset by providing detailed segmentation masks for 20 object categories, covering individual body parts such as heads, legs, paws, eyes, and ears. It includes 10,103 images for training and validation and 9,637 images for testing, offering a comprehensive resource for fine-grained part segmentation. For objects without consistent parts, like boats, silhouette annotations are used.

2.3.3 PartImageNet

The PartImageNet (He et al., 2022) dataset is a comprehensive resource designed for part segmentation tasks in computer vision. Comprising approximately 24,000 images across 158 classes from ImageNet, it provides detailed part-level annotations for both non-rigid and rigid objects. Organized into 11 super-categories, PartImageNet offers pixel-level segmentation masks, supporting various vision tasks such as part discovery, semantic segmentation, and few-shot learning.

2.4 Model analysis

We want to compare which model out of DesCo, VLPart and LISA is the most suitable for extending to open-vocabulary part-segmentation. To this extend we conduct comprehensive analysis which involves a detailed examination of each model's functions and inherent limitations. DesCo is inherently different from the other models as this model focuses on bounding boxes whereas VLPart and LISA focus on segmentation, this make the direct comparison of the two very difficult as the metrics for each scenario and not directly comparable.

As each model is originally evaluated using different metrics, we compare the metrics and establish which is the most discerning to evaluate, we also extensively discuss limitations associated with the metrics.

This approach ensures a thorough understanding of each model's capabilities and shortcomings, guiding us to the optimal choice for further development.

2.4.1 Bounding Box vs Segmentation

Segmentation in entity grounding is inherently more challenging than bounding box detection due to the precision and granularity it requires. While bounding boxes offer a coarse localization of objects, segmentation demands pixel-level accuracy, capturing the exact shape and boundaries of each entity. This level of detail requires a more nuanced understanding of the object's structure and its context within the image. In bounding box detection, the task is simplified, as the model only needs to predict the location of four boundary points to enclose the object, allowing for some degree of inaccuracy. However, segmentation must accurately capture complex shapes and contours, making it more sensitive to variations in object appearance, occlusions, and background clutter. This complexity is further compounded when dealing with overlapping objects or fine-grained parts, where even minor inaccuracies can lead to significant errors in scene understanding.

Moreover, segmentation requires models to integrate detailed spatial information and

Figure 2.1: (Top) Segmentation-mask and bbox for part-body of object-handbag (bottom) Segmentation-mask and bbox for part-bottom of object-bottom. While the segmentation mask for body and bottom have clear distinctions, the bbox for these parts are overlapping. Sample and annotations taken from PACO dataset.

context from the entire image, rather than relying solely on high-level cues that may suffice for bounding boxes. The need for fine-grained feature extraction and spatial reasoning emphasizes the difficulty of segmentation tasks, pushing the limits of current model architectures and computational resources. As such, while bounding boxes provide a broad-strokes approach to object localization, segmentation represents a more sophisticated and demanding challenge, essential for applications that require precise delineation of objects and their parts.

This distinction is especially notable in part-grounding. As shown in Figure 2.1, the bounding boxes for different parts of an object, such as the base and body of a handbag, can have significant overlap. Grounding models trained on bounding boxes tend to be more lenient, which has important implications for evaluation. The model has greater leeway to make errors without being heavily penalized; as long as the bounding box is approximately in the region of the part, it can achieve a decent IoU score. In contrast,

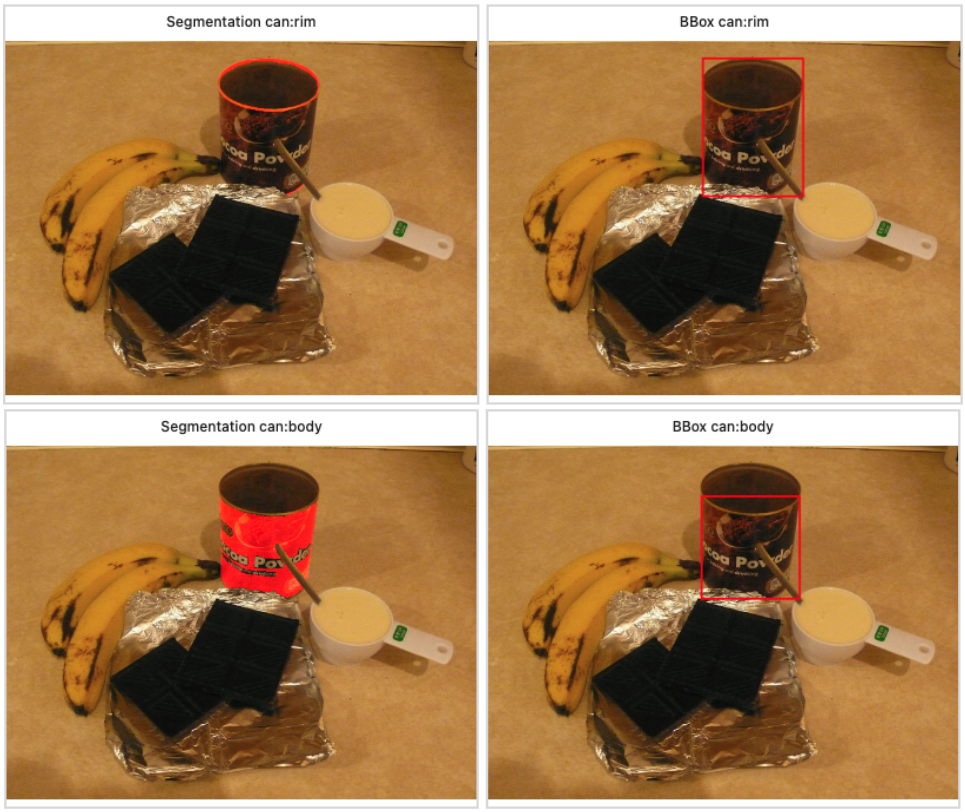


Figure 2.2: (Top) Segmentation-mask and bbox for part-rim of object-can (bottom) Segmentation-mask and bbox for part-body of object-can. Bbox of rim is much bigger than the body of can which is incorrect. Sample and annotations taken from PACO dataset.

segmentation masks need to be far more precise and are generally much smaller than bounding boxes, leaving less margin for error.

Figure 2.2 highlights how bounding boxes can sometimes be misleading. For example, the rim of a can is much smaller than its body, as the segmentation mask correctly indicates. However, the bounding boxes incorrectly suggest the opposite, making the rim appear larger than the body. This underscores the fact that, even in widely accepted annotated datasets for object parts, bounding boxes may not always be the most appropriate form of annotation. The segmentation mask for the two rims of the can does not overlap with any irrelevant parts of the object, while the bounding box for the rim has significant overlaps with other parts of the can, such as the body and label.

The difference between these two types of annotations can lead to varying results

Figure 2.3: Segmentation masks for part-back cover of object-cellular telephone. Each color denotes a new masks, there are five such instances in this image.

in evaluation. For instance, a bounding box that encompasses multiple parts of an object in an image would not be heavily penalized if the predicted bounding box covers a similar region as the original one. With a looser definition of boundaries, there is a higher likelihood of falsely positive points along the edges. On the other hand, the exact shape of the object becomes more critical with segmentation masks. In this case, even small changes in the predicted boundaries can significantly impact the IoU score. These variations in IoU scores can greatly influence the mAP, as mAP is determined by setting thresholds on the IoU scores.

As demonstrated above, predicting bounding boxes for object parts is a much simpler problem, both from a modeling standpoint and in terms of evaluation. Models trained and evaluated solely on bounding boxes, such as DesCo, consistently perform better when comparing metrics like mAP and IoU. However, segmentation models like LISA or VLPart, while achieving higher precision, may suffer from lower recall.

2.4.2 Multiple Masks

Each image-part pair can have multiple masks, for example, in Figure 2.3 there are five instances and segmentation masks for back-cover (part) of cellular telephone (object). Models like VLPart and DesCo support generating multiple masks for the same entry but LISA only outputs one mask per input prompt.

When calculating IoU and mAP, each ground-truth map or bounding box is compared with all available maps or bounding boxes of a particular label. For models like LISA, which predict only one label per query, the evaluation may result in lower scores in the presence of multiple ground-truth maps. This happens because only one ground-truth map will have a strong overlap with the predicted output. For instance, in Figure 2.3, the output mask can, at best, align with one of the five cellphone masks. As a result, evaluating with metrics like AP or IoU penalizes the model for all the other instances where there is no match. To address this, we use the max-IoU pair between each predicted and ground-truth pair for a given label to provide a more accurate evaluation.

A global average for mAP does not provide an accurate reflection of model performance, as some images may contain over 200 masks for the same part category or numerous small parts (in the order of hundreds), which can be too complex for most models to handle effectively. A global average unfairly penalizes models in these cases. A more accurate measure of performance is to compute AP on a per-image basis and then average the results. Therefore, we follow this protocol for evaluating LISA.

2.4.3 VLPart Analysis

To investigate the limitations of the VLPart model, we conduct experiments on both in-domain and out-of-domain objects from the training data. We query the model using images with varying attributes such as texture and color, as well as all possible known labels. Ideally, the model should not predict parts or objects that do not exist. However, as shown in Figure 2.4, we observe that for out-of domain objects such as shovels VLPart not only predicts incorrect parts but also misidentifies parts of shovel as entirely different

objects based solely on shape.

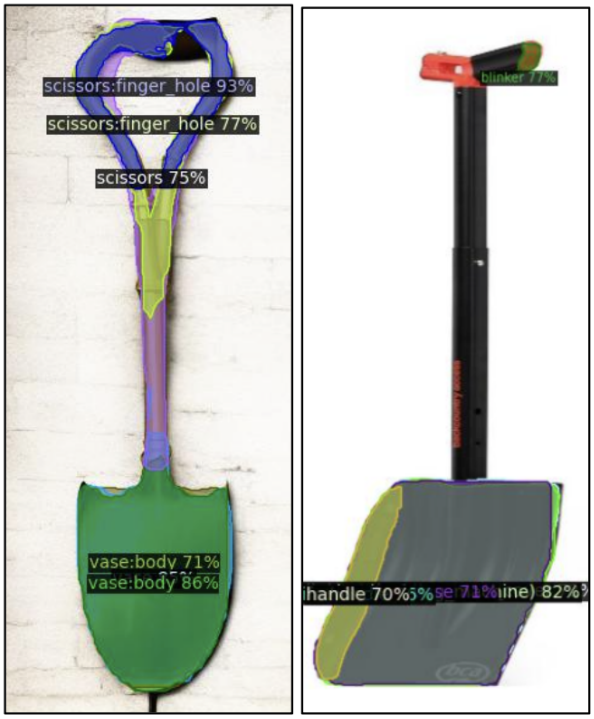


Figure 2.4: Querying VLPart with out-of-domain objects with all training classes as prompts.

Figure 2.5: Querying VLPart with in-domain objects with all training classes as prompts.

This trend is also evident in in-domain objects, such as when the bowl of a spoon is mistakenly classified as a mouse (Figure 2.5). VLPart appears to rely heavily on shape-based recognition for part discovery, as indicated by the high confidence score it assigns to parts being misidentified as objects.

CHAPTER 3

Proposed solution

3.1 Methodology

We experiment with a few fine-tuning strategies for DesCo and LISA. DesCo is originally only trained on object level information so we wanted to naturally extend DesCo to part level descriptions and perform training. We rely on the PACO dataset for supervised part level annotation.

We query Llama 3 (Grauman et al., 2022) to automatically generate nuanced descriptions for each part of a given object. Our focus lies specifically on the visual characteristics that make each part unique and distinctive. Below, we provide a sample query demonstrating this approach:

```
{
    {
    "role": "system",
    "content": "You are an expert who can describe visual features of parts
    of objects for understanding distinctions visually. Output each main
    feature as a list"
    },
    {
    "role": "user", "content": <PART> of <OBJECT>
    }
}
```

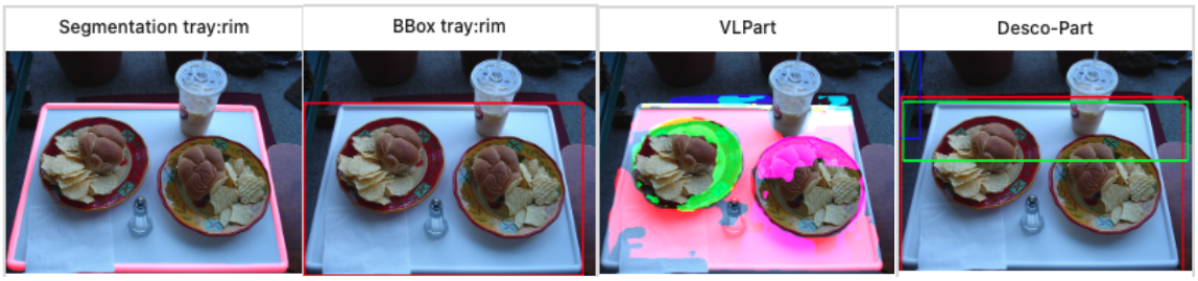


Figure 3.1: Ground-truth segmentation and bounding in comparison to the VLPart and DesCo. VLPart and DesCo output multiple masks/bounding-box for a single prompt.

Once all the descriptions are generated, we clean and process the text for training. During fine-tuning of DesCo and LISA, we append the format "<PART> of a <OBJECT>" with the corresponding "<DESCRIPTION>".

We also experiment with training using only positive descriptions, as well as a combination of both positive and negative descriptions. Our results indicate that using only positive descriptions yields the best performance. Including both positive and negative descriptions seems to confuse the model during training, reducing its overall effectiveness. In our discussion for any given model X, X fine tuned means fine tuning on existing part dataset whereas X Description means fine tuning done with descriptions.

3.2 Result

3.2.1 Qualitative analysis

Instead of predicting one mask, VLPart and DesCo give multiple output mask for a single query which often span a large option of scales (Figure 3.1. This leads to more chances for correct overlap. As long as there is a decent overlap the model will not be penalized for false predictions (this is the bias in LVIS-AP calculations).

Similar to AP, the LVIS implementation of IoU also doesn't penalize for multiple incorrect predictions if there is atlease one that aligns well. We fix IoU comparison by computing IoU for each predicted mask and averaging this per image-part prompt.

Figure 3.2a presents the results for the query "nose of dog". Both LISA Fine-tuned

Description and VLPart achieve high scores for this query. However, VLPart produces multiple candidate masks, only one of which accurately aligns with the ground truth. DesCo-Part, on the other hand, performs significantly worse, with all predicted bounding boxes encompassing broader regions (such as the face) rather than precisely focusing on the specific part, in this case, the nose of the dog.

Figure 3.2b illustrates the results for the query "neck of dog". Here, we observe that LISA Fine-tuned Description outperforms VLPart, delivering a higher degree of segmentation accuracy. While VLPart proposes several regions, none exhibit significant overlap with the ground truth. In contrast, DesCo-Part, with its multiple proposed bounding boxes, increases the probability of finding a more accurate match, as evidenced in this particular instance.

In Figure 3.2c and Figure 3.2e, we observe that none of the models effectively segment the blade of the knife. Figure 3.2c shows that VLPart fails to generate any output, as none of the predicted masks reach a sufficient confidence threshold for the region.

Another clear example of how bounding box detection is a simpler task compared to segmentation can be seen in Figure 3.2d. Although LISA performs well, VLPart is only able to identify part of the inner side. DesCo, despite predicting multiple bounding boxes with high confidence, produces several incorrect predictions. This further highlights that even in an easier task like bounding box detection, inaccuracies can still arise.

In Figure 3.2f, VLPart successfully places the segment roughly in the general region where the spoon should be, but it fails to precisely identify its exact location. In contrast, LISA, fine-tuned on descriptive data, is capable of identifying a nearly accurate segment.

In Figure 3.3, we focus on a direct comparison between LISA fine-tuned on descriptions and VLPart. It is evident that VLPart's predicted masks are not always relevant. Even when VLPart correctly predicts the part, it often includes additional masks for completely unrelated objects or parts. For instance, in Figure 3.3f, when querying for the "body" of the vase, VLPart also outputs a mask for a teddy bear. Similarly, in Figure 3.3b, the mask for the coffee table appears, despite having no connection to the query "top of television

set". Although models like VLPart and DesCo may output completely irrelevant masks that are unrelated to the object in question, the mAP and IoU scores do not penalize them as long as one of the predicted masks is correct. This characteristic can lead to a false sense of confidence in the model's performance, as the metrics do not reflect the inaccuracy of the other predicted masks.

3.2.2 Quantitative analysis

Table 3.1: Comparison of different models on the PACO dataset with average AP scores.

Model	Training Dataset	Average AP
Bounding Box		
DesCo PACO	PACO	17.56
DesCo PACO (+ve)	PACO	23.37
DesCo PACO (+ve) description	PACO + description	20.55
Segmentation		
VLPart	PACO + Pascal Part + PartImageNet	9.6
LISA	PACO + Pascal Part + PartImageNet	9.9
LISA Fine-tuned	PACO	13.4
LISA Description (+ve)	PACO description	16.3

The results in Table 3.1 provide a detailed comparison of the models across both bounding box and segmentation tasks, highlighting key performance differences. Interestingly, for the bounding box task, the DesCo PACO (+ve) model, which was not trained with descriptive input, achieves the highest Average AP of 23.37. This demonstrates that, despite the absence of description-based training, the model is still able to localize objects effectively. The high performance of this model further emphasizes that bounding box prediction allows for a greater margin of error, where broad object localization is sufficient for achieving high scores.

In contrast, the DesCo PACO (+ve) with description model, which incorporates descriptive training, shows a slightly lower score of 20.55. This suggests that the descriptive training, while generally helpful in refining object understanding, may introduce complexity that doesn't directly improve performance in the coarser task of bounding box prediction. This model's lower score, despite the added contextual information, aligns with the notion

that descriptions may not always be advantageous in tasks where precise part identification is not as critical.

Shifting to the segmentation task, the results continue to underscore the difficulty of this task compared to bounding box prediction. LISA Fine-tuned performs the best among the segmentation models, with an Average AP of 13.4, showing that its fine-tuning on part-level data allows it to perform more accurately in segmenting specific regions. Meanwhile, LISA Description performs relatively well with a score of 16.3, highlighting how descriptive input can benefit segmentation tasks that require finer granularity.

VLPart, however, struggles with segmentation, scoring only 9.6, which reflects its difficulty in accurately identifying and segmenting object parts. Despite its ability to generate multiple proposed regions, the model's part discovery remains limited, as seen through lower segmentation performance, where pixel-level precision is necessary.

3.3 Conclusion

In conclusion, this research has explored the complexities involved in open-vocabulary part segmentation, contrasting it with the comparatively simpler task of bounding box detection. Our analysis of models such as DesCo, LISA, and VLPart demonstrates the critical challenges inherent in part segmentation, where precision at the pixel level is required. The performance differences observed between these models underscore how segmentation demands more granularity than the coarse approximations of bounding box detection.

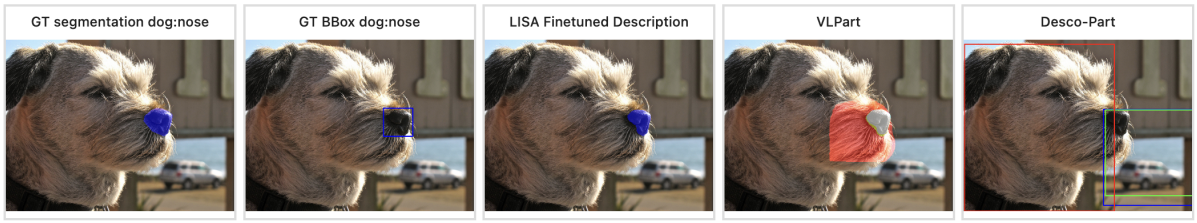
Bounding box prediction, exemplified by DesCo PACO (+ve), consistently performed better, with an Average AP of 23.37, despite not using descriptive input. This illustrates the ease with which bounding box tasks can be accomplished, as the model only needs to approximate the general location of an object. Even without descriptions, the task remains forgiving of small inaccuracies, provided the bounding box sufficiently covers the object.

Conversely, segmentation presents a far greater challenge, with LISA Fine-tuned achieving the highest score of 13.4 in segmentation, a clear drop in performance when compared to bounding box detection. VLPart struggles even more in this context, with an AP of 9.6, reinforcing the notion that the task of part segmentation requires a much deeper level of spatial reasoning and part-specific localization.

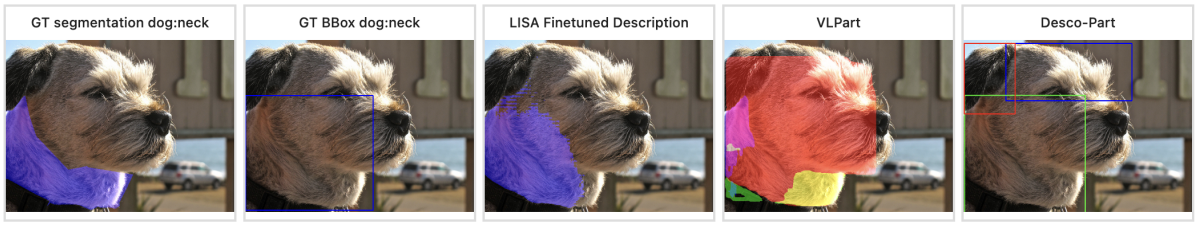
An important takeaway from this research is the role that descriptions of parts play in segmentation tasks. Descriptions provide essential context that enables the model to differentiate between visually similar parts, such as distinguishing between the "blade" of a knife and the "handle." Descriptions, like "curved handle of a spoon" or "sharp edge of a blade," enable the model to make more informed decisions about part segmentation, going beyond purely visual input. This additional context is what allows models like LISA Description to improve performance, as evidenced by its comparatively higher score of 16.3. The results clearly show that while descriptions do not necessarily improve bounding box detection, they are particularly effective in refining part segmentation, where finer granularity and part-specific knowledge are required.

However, despite the advantages that part descriptions bring, models like VLPart still show limitations in generating accurate masks, often outputting irrelevant regions in part segmentation. This highlights the broader challenge in developing models capable of both general object detection and detailed part-level segmentation with the same precision.

Ultimately, the findings from this research reinforce the ongoing need to refine segmentation models to handle the complexity of part-grounding. While descriptions prove to be a helpful addition, future work must focus on further enhancing the ability of these models to deal with the subtleties and intricacies inherent in part segmentation, particularly in open-vocabulary settings where unseen parts and objects need to be effectively localized.



(a) Nose (part) of Dog (object). LISA-finetuned and VLPart notably perform well.



(b) Neck (part) of Dog (object). LISA-finetuned and DesCo-Part perform well.

(c) Blade (part) of Knife (object). All models fail to identify the correct part.

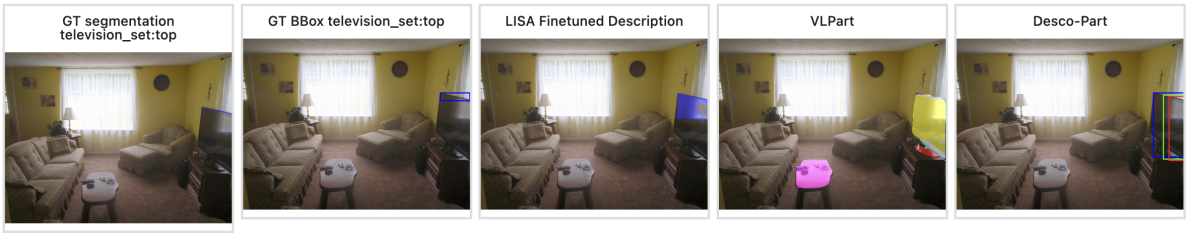
(d) Inner side (part) of Pan (object). LISA performs the best.

(e) Shoulder (part) of Sweater (object).

(f) Bowl (part) of Spoon (object). VLPart tired to localize the part spoon.

Figure 3.2: Comparing differnce cases across all three models

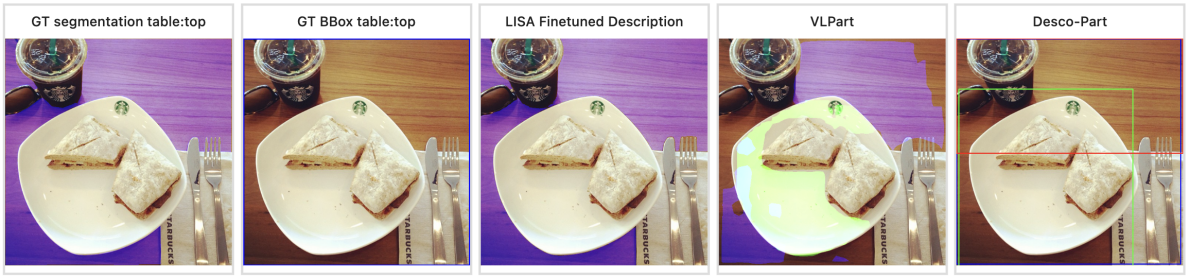
(a) Seat (part) of Bench (object).



(b) Top (part) of Television set (object).

(c) Cover (part) of Bucket (object).

(d) Bottom (part) of Plate (object).



(e) Top (part) of Table (object).

(f) Body (part) of Vase (object).

Figure 3.3: Comparing LISA fine tuned and DesCo-Part.

Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Batra, D., and Parikh, D. (2016). Vqa: Visual question answering. 7

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022). Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736. 8

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086. 8, 9

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. 6

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660. 7

Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., and Yuille, A. (2014). Detect what you can: Detecting and representing objects using holistic models and body parts. 23

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. (2020). Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer. 7, 8

Cho, J., Lei, J., Tan, H., and Bansal, M. (2021). Unifying vision-and-language tasks via text generation. 10

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. 7

Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y., and Li, G. (2022). Learning to prompt for open-vocabulary object detection with vision-language model. 9

Endo, Y. (2023). Masked-attention diffusion guidance for spatially controlling text-to-image generation. 6

Faghri, F., Fleet, D. J., Kiros, J. R., and Fidler, S. (2018). Vse++: Improving visual-semantic embeddings with hard negatives. 8

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M. A., and Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc. 8

Gao, M., Xing, C., Niebles, J. C., Li, J., Xu, R., Liu, W., and Xiong, C. (2022). Open vocabulary object detection with pseudo bounding-box labels. 9

Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., and Qiao, Y. (2021). Clip-adapter: Better vision-language models with feature adapters. 10

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. (2017). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913. 7, 9

Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., Martin, M., Nagarajan, T., Radosavovic, I., Ramakrishnan, S. K., Ryan, F., Sharma, J., Wray, M., Xu, M., Xu, E. Z., Zhao, C., Bansal, S., Batra, D., Cartillier, V., Crane, S., Do, T., Doulaty, M., Erapalli, A., Feichtenhofer, C., Fragomeni, A., Fu, Q., Gebreselasie, A., Gonzalez, C., Hillis, J., Huang, X., Huang, Y., Jia, W., Khoo, W., Kolar, J., Kottur, S., Kumar, A., Landini, F., Li, C., Li, Y., Li, Z., Mangalam, K., Modhugu, R., Munro, J., Murrell, T., Nishiyasu, T., Price, W., Puentes, P. R., Ramazanova, M., Sari, L., Somasundaram, K., Southerland, A., Sugano, Y., Tao, R., Vo, M., Wang, Y., Wu, X., Yagi, T., Zhao, Z., Zhu, Y., Arbelaez, P., Crandall, D., Damen, D., Farinella, G. M., Fuegen, C., Ghanem, B., Ithapu, V. K., Jawahar, C. V., Joo, H., Kitani, K., Li, H., Newcombe, R., Oliva, A., Park, H. S., Rehg, J. M., Sato, Y., Shi, J., Shou, M. Z., Torralba, A., Torresani, L., Yan, M., and Malik, J. (2022). Ego4d: Around the world in 3,000 hours of egocentric video. 23, 30

Gupta, A., Dollár, P., and Girshick, R. (2019). Lvis: A dataset for large vocabulary instance segmentation. 23

Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., and Bigham, J. P. (2018). Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617. 7, 9

He, J., Yang, S., Yang, S., Kortylewski, A., Yuan, X., Chen, J.-N., Liu, S., Yang, C., Yu, Q., and Yuille, A. (2022). Partimagenet: A large, high-quality dataset of parts. 23

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., and Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. 8

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026. 6

Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., and Jia, J. (2024). Lisa: Reasoning segmentation via large language model. 8, 10, 19

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324. 6

Lee, K.-H., Chen, X., Hua, G., Hu, H., and He, X. (2018). Stacked cross attention for image-text matching. 8

Li, B., Weinberger, K. Q., Belongie, S., Koltun, V., and Ranftl, R. (2022a). Language-driven semantic segmentation. 10

Li, B., Zhang, Y., Chen, L., Wang, J., Pu, F., Yang, J., Li, C., and Liu, Z. (2023a). Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*. 8

Li, J., Li, D., Savarese, S., and Hoi, S. (2023b). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR. 8

Li, L. H., Dou, Z.-Y., Peng, N., and Chang, K.-W. (2023c). Desco: Learning object recognition with rich language descriptions. 10, 11

Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., Chang, K.-W., and Gao, J. (2022b). Grounded language-image pre-training. 12

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer. 7

Lin, C., Sun, P., Jiang, Y., Luo, P., Qu, L., Haffari, G., Yuan, Z., and Cai, J. (2022). Learning object-language alignments for open-vocabulary object detection. 9

Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2024). Visual instruction tuning. *Advances in neural information processing systems*, 36. 8

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440. 6

Malinowski, M., Doersch, C., Santoro, A., and Battaglia, P. (2018). Learning visual question answering by bootstrapping hard attention. 7

Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., and Wei, F. (2023). Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*. 8

Pi, R., Gao, J., Diao, S., Pan, R., Dong, H., Zhang, J., Yao, L., Han, J., Xu, H., Kong, L., et al. (2023). Detgpt: Detect what you need via reasoning. *arXiv preprint arXiv:2305.14167*. 8

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR. 7, 8

Ramanathan, V., Kalia, A., Petrovic, V., Wen, Y., Zheng, B., Guo, B., Wang, R., Marquez, A., Kovvuri, R., Kadian, A., Mousavi, A., Song, Y., Dubey, A., and Mahajan, D. (2023). Paco: Parts and attributes of common objects. 23

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V. (2017). Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024. 8, 9

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer. 6

Sharma, P., Ding, N., Goodman, S., and Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics. 8

Shen, S., Li, L. H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.-W., Yao, Z., and Keutzer, K. (2021). How much can clip benefit vision-and-language tasks? 10

Srivastava, Y., Murali, V., Dubey, S. R., and Mukherjee, S. (2020). Visual question answering using deep learning: A survey and performance analysis. 7

Sun, P., Chen, S., Zhu, C., Xiao, F., Luo, P., Xie, S., and Yan, Z. (2023). Going denser with open-vocabulary part segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15453–15465. 7, 10, 15

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30. 8

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164. 8, 9

Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al. (2024). Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36. 8

Wang, X., Girshick, R., Gupta, A., and He, K. (2018). Non-local neural networks. 6

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR. 8

Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al. (2023). mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*. 8

Yu, F. and Koltun, V. (2016). Multi-scale context aggregation by dilated convolutions. 6

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. (2021). Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588. 7, 8

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). Pyramid scene parsing network. 6

Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L. H., Zhou, L., Dai, X., Yuan, L., Li, Y., and Gao, J. (2021). Regionclip: Region-based language-image pretraining. 9

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. (2022a). Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348. 10

Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., and Misra, I. (2022b). Detecting twenty-thousand classes using image-level supervision. 9

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. (2023). Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*. 8

Zou, X., Dou, Z.-Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., et al. (2023). Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127. 6

Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., and Lee, Y. J. (2024). Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36. 6