See, Say, and Segment: Teaching LMMs to Overcome False Premises

Tsung-Han Wu* Giscard Biamby* David Chan Lisa Dunlap
Ritwik Gupta Xudong Wang Joseph E. Gonzalez Trevor Darrell
University of California, Berkeley

Abstract

Current open-source Large Multimodal Models (LMMs) excel at tasks such as open-vocabulary language grounding and segmentation but can suffer under false premises when queries imply the existence of something that is not actually present in the image. We observe that existing methods that fine-tune an LMM to segment images significantly degrade their ability to reliably determine ("see") if an object is present and to interact naturally with humans ("say"), a form of catastrophic forgetting. In this work, we propose a cascading and joint training approach for LMMs to solve this task, avoiding catastrophic forgetting of previous skills. Our resulting model can "see" by detecting whether objects are present in an image, "say" by telling the user if they are not, proposing alternative queries or correcting semantic errors in the query, and finally "segment" by outputting the mask of the desired objects if they exist. Additionally, we introduce a novel False Premise Correction benchmark dataset, an extension of existing RefCOCO(+/g) referring segmentation datasets (which we call FP-RefCOCO(+/g)). The results show that our method not only detects false premises up to 55% better than existing approaches, but under false premise conditions produces relative cIOU improvements of more than 31% over baselines, and produces natural language feedback judged helpful up to 67% of the time.

1. Introduction

Perception systems engaging with real-world environments often need to understand and respond to complex queries such as "find the keys with the purple heart on them" or "bring me the remote for the television." Solving such complex visual tasks can require active reasoning, world knowledge, and an implicit understanding of the scene which are often unavailable to simple visual perception systems [33]. An extension of referring segmentation [22], "reasoning segmentation," requires that models are capable not only of understanding the query but reasoning on the query as well.

However, what if the "keys with the purple heart" do not exist in the scene? While recent methods for reasoning segmentation have shown remarkable performance on "positive" queries where the query object exists in the scene, most existing approaches for reasoning segmentation fail to account for this "false premise" scenario [33], and happily produce a hallucinated segmentation even when the objects associated with the query do not exist in the image [37, 66] (see Fig. 1). It is desirable for robust reasoning segmentation systems to not only respond in the negative but to also propose corrected expressions when appropriate. Such robust systems should first be able to **"see"**, by detecting *if* an object from the query is present in an image, then **"say"** something *about* the object itself if it's not there, suggesting alternatives to the user's query and optionally providing additional helpful information in the scene, before finally being able to **"segment"** by showing *where* in an image an object is grounded, if the user has not withdrawn their request.

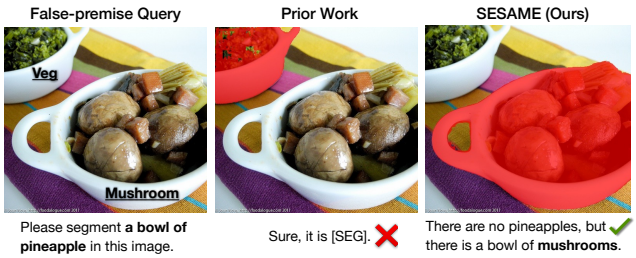

Figure 1. False premise failures with LMMs: contemporary open-source LMMs combined with segmentation decoders are able to generate referring segments effectively but have difficulty on segmentation questions which ask the model to refer to something that is not present in the image. SESAME, our See-Say-Segment LMM, uses model chaining and joint training to overcome this problem.

Until now, "false premise"-aware approaches have focused on the "see" and "segment" components, often using two-stage cascaded approaches, where an auxiliary classifier is used to "see" and a segmentation backbone is used to "segment" [37, 63, 66]. These pipeline-based approaches do not demonstrate reasoning ability when interpreting a reference and can not engage with users in task-directed dialogue; they cannot "say" anything about the query if it is incorrect, unfor-

[1] *Equal contribution.

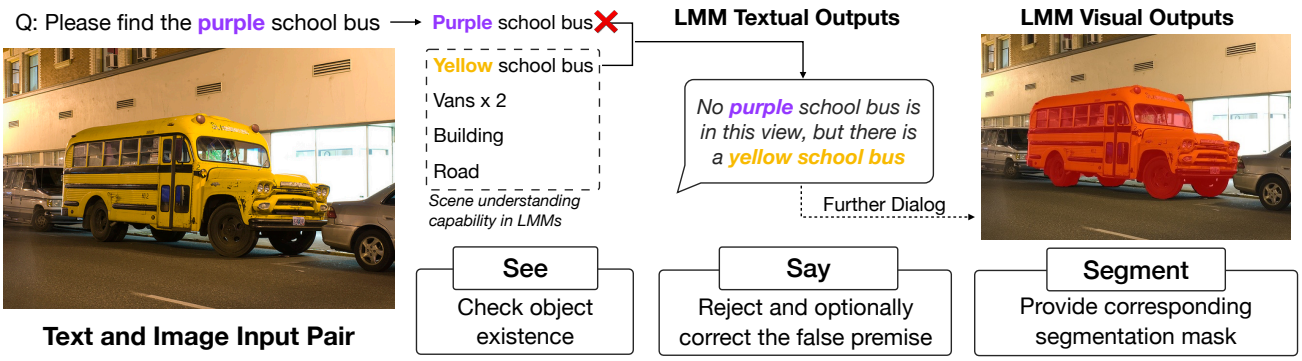

Figure 2. *SESAME* is an LMM that can "see" whether objects are detected in an image and "say" by telling the user if they are there or not. When appropriate, alternative queries can be offered or semantic errors corrected in the query. SEASAME can then "segment" by returning the mask of the desired object.

tunately letting users continue with erroneous assumptions and queries without correction.

Addressing this unexplored area, we introduce a novel dataset and associated benchmarks, dubbed FP-RefCOCO, FP-RefCOCO+, and FP-RefCOCOg. These datasets, an expansion from RefCOCO(+/g) [43, 74], are augmented with context-aware false-premise queries via Large Language Models (LLMs), which are essential to train and evaluate a model's ability to "see", "say" and "segment". In this new task, we find that existing open-source LMM-based referring segmentation approaches often fail to "see" and "say" due to catastrophic forgetting during instruction fine-tuning.

To counter this issue, we develop two reasoning segmentation methods resilient to false premises. Our method includes a cascading models approach and an all-encompassing LMM, ***SESAME*** (**SE**E, **SA**y, seg**ME**nt), which is jointly trained with our novel dataset. By leveraging the reasoning and referencing nature capabilities of contemporary LMMs, these methods can not only "see" and "segment" but also "say" what is necessary to reject or even correct a query. In summary, our contributions include the introduction of a novel benchmark dataset and:

- **An LMM that can "see":** We analyze how existing approaches for reasoning segmentation fail to recognize false premise queries, and show that cascading models and joint data fine-tuning to produce relative accuracy improvements of up to **55.45%** over a baseline's ability to detect false-premise queries.
- **An LMMs that can "say":** We further show our approach is novel in that it can give helpful feedback about the query, and demonstrate that such feedback is judged to be helpful up to **67%** of the time.
- **An LMM that can "segment":** Finally, we demonstrate the importance of false-premise robustness in improvements in segmentation quality, showing that robust false-premise training can result in relative cIoU improvements over baselines of up to **31.65%**.

2. Related Work

Reasoning segmentation, a subset of referring segmentation, introduced by Lai *et al*. [33], focuses on complex reasoning tasks in addition to localized references. Reasoning segmentation exists in contrast to the more global tasks of semantic segmentation, which assigns class labels to every pixel in an image [1, 6, 8, 17, 24, 32, 40, 47, 56, 57, 60, 61, 69, 73, 78–80, 82], instance segmentation, which detects pixels corresponding to instances of objects in a scene [9, 19, 77], and panoptic segmentation, which solve both instance and semantic segmentation problems simultaneously [7, 30, 36, 67]. It is also more fine-grained than approaches for referring object grounding and reasoning [4, 5, 14, 23, 34, 51, 59, 70, 72, 75, 76, 81], as these approaches operate on bounding boxes corresponding to objects, and do not seek to localize the pixels of the objects directly.

The current state-of-the-art for reasoning segmentation, LISA [33], uses a pre-trained large multimodal model fine-tuned to output segmentation tokens for each image. While LISA is capable of complex reasoning, it is trained in a manner that encourages producing segmentation/region outputs, even in the presence of a false-premise query. Similar to LISA, X-Decoder [83] and SEEM [84] can both produce pixel-level segmentation and language tokens, however, focus on multi-task performance, and struggle to perform complex reasoning segmentation tasks [33].

While our proposed method is the first to explore false premises in the field of reasoning segmentation, understanding and detecting false premises has been studied in several other areas in computer vision [12] including visual question answering [54], image/text matching [15, 16, 29, 50, 68], image-grounded conversation [46], tool usage [62] and hallucination detection [55]. Indeed, it has long been known

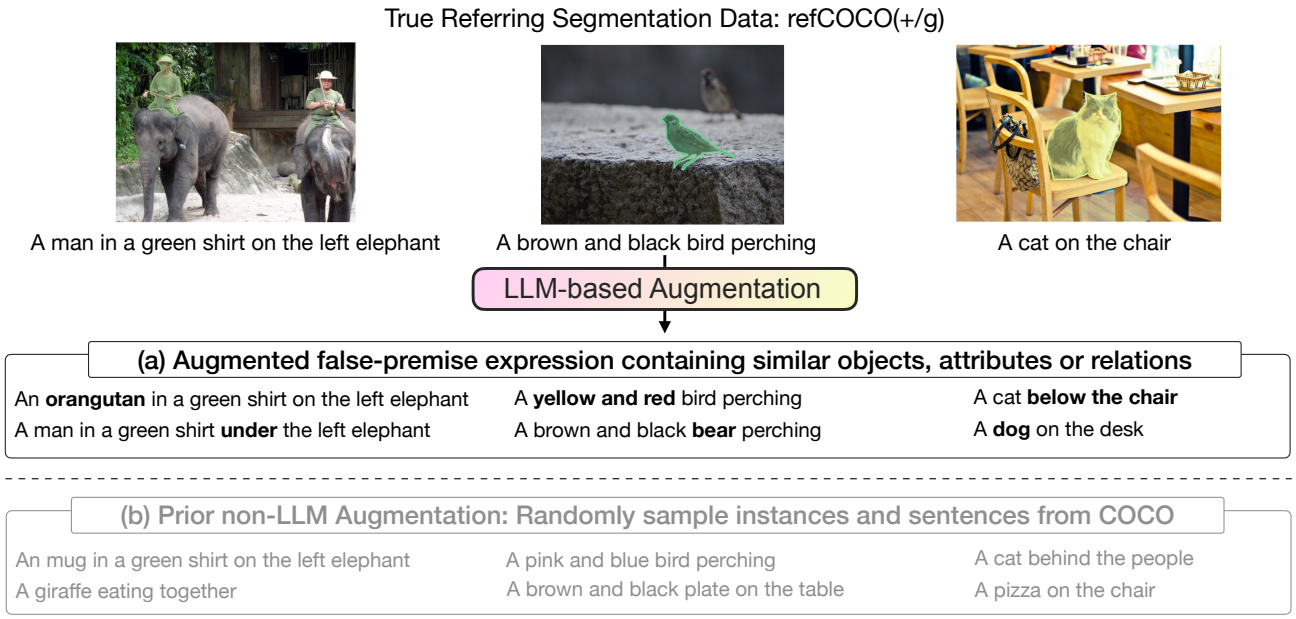

Figure 3. FP-RefCOCO Dataset Creation. Using refCOCO for base images, we employ an LLM to create a false-premise referring segmentation dataset with similar objects, attributes, and relations. Such paired examples enable the the creation of specific correction ground truth that is more specific than baseline methods which simply sample positive and negative examples. This data allows us to train an LMM that has robust reasoning reference capabilities.

in visual question answering that sometimes the image cannot entail any of the possible answers, and both datasets [18, 25, 27, 39, 42, 58] and methods [35, 42, 44, 45, 52, 65] have been developed which can evaluate and correct for false premises in the case of question answering. Generally, methods for detecting false premises fall into a cascaded approach with two components, a "detection" model which is designed to determine if the question is answerable ("see"), and the standard "answering" model, which actually answers the question [35, 44, 45, 52, 63].

Beyond question answering, several explicit measures have been designed which use pre-trained vision and language models to determine how closely text matches with a related image [10, 11, 20, 26, 28, 53, 71], however while such measures can detect image/text relevance, they can neither "segment" nor "say".

Closest to our work, Wang *et al*. [63] introduce a cascaded method for referring segmentation in the presence of false premises composed of an entity detection module, an expression parsing module (which parses objects using a textual scene graph), and a complex entity/relationship matching detection method based on WordNet distances. While this method is capable of understanding false premises and giving feedback in referring expressions, it cannot handle open-domain language commonly found in reasoning segmentation tasks and is restricted to closed-domain tasks with fixed vocabularies. Our work is the first approach that enables false premise detection and language feedback in open-domain reasoning segmentation tasks.

3. A New Dataset and Benchmark for False Premise Correction

Dialog-based models with the ability to segment and reason are traditionally trained on referring expression datasets which tend to only contain positive examples—examples that contain the object pertaining to the query language. Models trained under these conditions will always produce positive results, regardless of the truthfulness of the premise. Prior false-premise strategies to tackle this often integrate a classifier ahead of the segmentation module [37, 66], but this solution can be too restrictive, limiting the LMMs in engaging with diverse, open-domain conversational scenarios.

In response, *we alter both the task and the data* to facilitate the ability for models to provide more human-like responses when presented with a question about a non-existent object. This new task, False Premise Correction, expects models to suggest an alternative referring expression that more closely matches an object in the image if prompted with a query that describes a missing object.

Although existing datasets such as R-RefCOCO [66] include queries referring to non-existent items in images, their method of generating negative expressions through naive random sampling often lacks context awareness. This limitation significantly reduces their effectiveness for false-premise correction tasks. Consider an image with a cat on a chair as in Fig. 3 (b): contextually valid false premises that could

Dataset	Split	Images	Objects	Sentences	Positive Sentences	Negative Sentences
FP-RefCOCO	train	16,992	42,278	234,445	120,191	114,254
	val	1,500	3,805	20,962	10,758	10,204
	testA	750	1,975	11,205	5,726	5,479
	testB	750	1,798	9,514	4,889	4,625
	Total	19,992	49,856	276,126	141,564	134,562
FP-RefCOCO+	train	16,994	42,404	234,892	120,624	114,268
	val	1,500	3,811	21,094	10,834	10,260
	testA	750	1,975	11,061	5,657	5,404
	testB	750	1,810	9,883	5,095	4,788
	Total	19,994	50,000	276,930	142,210	134,720
FP-RefCOCOg	train	21,899	42,246	157,866	80,512	77,354
	val	1,300	2,573	9,554	4,896	4,658
	test	2,600	5,023	18,830	9,602	9,228
	Total	25,799	49,822	186,250	95,010	91,240

Table 1. Details for the FP-RefCOCO datasets, and train/val/test splits. Our dataset splits mirror those of RefCOCO (unc), RefCOCO+ (unc), and RefCOCOg (umd). To each original dataset, we have appended an approximately equal number of negative referring expressions, each coupled with a corresponding corrected sentence, thus creating an augmented dataset specifically for the False Premise Correction task.

be logically corrected to "a cat on the chair" might include phrases like "a cat under the chair" or "a dog on the chair." However, R-RefCOCO typically produces less suitable examples, such as "a pizza on the chair" or "a cat behind the people," which do not align with realistic model correction expectations. Furthermore, these datasets do not provide a direct link between each false premise query and its corresponding correct alternative, a critical aspect for effective training and evaluation in false premise correction.

To address these issues, we present FP-RefCOCO(+/g), a new benchmark dataset building upon the RefCOCO(+/g) referring segmentation datasets [43, 74]. For each image, FP-RefCOCO(+/g) not only incorporates the original positive referring queries but also pairs them with a diverse range of contextually related false premise queries. To generate negative samples, we modify a single element (object, adjective, or relation) in the positive referring expressions by prompting the OpenAI GPT-3.5-turbo model [2]. As depicted in Fig. 3 (a), our LLM-based augmentation strategy yields false premise queries that are more closely aligned with the context. After some basic data cleaning to ensure the responses are parseable, we end up with a nearly 1:1 positive/negative sample ratio and the same train/test/val splits as RefCOCO(+/g). Full statistics are provided in Tab. 1.

The FP-RefCOCO(+/g) benchmark dataset enables the evaluation and training of Language and Multimodal Models (LMMs) in open-domain reasoning and segmentation tasks, focusing on three essential capabilities: "See," "Say," and "Segment." In Tab. 2 and Tab. 3, the statistics showed significant limitations of the state-of-the-art model, LISA [33], particularly in its complete inability to reject non-existent items ("See") with 0% recall on false premise query or to provide any appropriate corrections ("Say"). LISA predicts a segmentation for all false premise sentences, resulting in an approximately 30% reduction in segmentation cIoU compared to the original dataset without any false premise queries. In response, we developed and trained an integrated LMM to achieve notable improvements across all three capabilities, which is detailed in Sec. 4.

4. An LMM that can See, Say, and Segment

To enable intelligent interaction systems that simultaneously possess the abilities to see, say, and segment, we first introduce a novel approach cascading various LMMs with distinct functionalities. We then present *SESAME*, a unified **SE**e, **SA**y, seg**ME**nt model with the aid of our curated dataset described above.

Existing generic LMMs for VQA, such as GPT-4V [49] or LLaVA-v1.5 [38], are adept at identifying objects in images and suggesting alternatives when necessary. However, segmentation-specialized LMMs such as LISA, while capable of generating segmentation masks given diverse language prompts, struggle with queries about non-existent objects. In these cases, LISA often produces segmentation masks but fails to provide relevant feedback, typically offering generic responses like "Sure, it is [SEG]." This behavior deviates from our desired outcome, indicating a need for more context-aware responses in advanced interaction systems. As shown in Tab. 2, there is a significant degradation in the performance of "seeing" and "saying". Intriguingly, the original LLaVA model which LISA is based on, prior to its fine-tuning for segmentation capabilities, did possess

Method	FP-RefCOCO			FP-RefCOCO+			FP-RefCOCOg		
	See	Say	Segment	See	Say	Segment	See	Say	Segment
GRES [37]	65.11	-	47.47	64.00	-	37.61	58.15	-	35.15
SEEM (Focal-L) [84]	51.36	-	37.92	51.32	-	40.01	51.25	-	35.87
LISA [33]	51.36	0.00	44.00	51.32	0.00	39.62	51.25	0.00	39.64
Cascading (Ours)	75.59	0.35	55.18	75.03	0.42	48.64	76.07	0.55	49.98
SESAME (Ours)	**79.84**	**0.63**	**57.93**	**80.00**	**0.61**	**50.81**	**81.78**	**0.67**	**53.79**

Table 2. SESAME (ours) and existing methods on various referring segmentation tasks. The See scores measure the binary classification accuracy. Say is measured via the CLAIR score, which ranks the similarity of the suggested false premise correction against positive referring expressions for the same referent. The Segment scores are (cIoU).

Method	See		Segment (cIoU)			
	Acc (F. Pre)	Acc (T. Pre)	0% FP	25% FP	50% FP	75% FP
LISA	0.00	**100.00**	**67.99**	52.36	39.64	34.15
Cascading (Ours)	58.76	94.64	65.77	58.53	49.98	45.82
SESAME (Ours)	**67.89**	96.64	66.02	**60.64**	**53.79**	**49.79**

Table 3. Ablation on the SEE performance in detail and the amount of false premise data used at test time for the segmentation scores on the FP-RefCOCOg dataset. As *SESAME* (ours) has fairly good SEE capability, it demonstrates superior segmentation performance even when the false premise sampling rate is as high as 75%.

these abilities. This indicates a critical issue in the realm of LMMs – the challenge of catastrophic forgetting during the process of learning new skills.

To address this, we first propose a cascading approach for the False Premise Correction task. The first LMM detects the presence or absence of objects in images and also engages in dialogue with users, providing clarifications or alternative suggestions when necessary. Once an object's presence is confirmed, the query can then be passed to the second LMM specializing in the task of "segmentation." This second-stage model is a segmentation-focused LMM [33] that performs referring segmentation via prompts in the form of "Please help me segment X in the image" and has high performance in both conventional semantic segmentation and complex reasoning segmentation tasks. This method coordinates between the two LMMs via prompt chaining with the first excelling in accurate object detection and contextual language response and the second in detailed image segmentation.

However, a single model with all three capabilities is desired, but as described above existing approaches "forget" how to "see" and "say". We address the catastrophic forgetting problem by utilizing a joint-training strategy. As in [33] we instantiate *SESAME* with LLaVa-v1.5 [38] for the "see" and "say" portions of the pipeline and Segment Anything [31] as the segmentation backbone.

This training utilizes three distinct datasets: the train split ("train" from Tab. 1) of the custom-designed FP-RefCOCO(+/g), the LLaVA VQA instruction finetuning dataset, and the train splits from R-RefCOCO(+/g). We call this the unified training set. The FP-RefCOCO(+/g) dataset comprises both positive and negative queries, with the incorrect ones being amended and always related to their original versions. In contrast, R-RefCOCO(+/g) includes randomly selected nonexistent COCO objects, which lack contextual relevance to the images; this dataset is employed to train the model to simply reject non-existent objects rather than offer corrections.

We distributed FP-RefCOCO(+/g), LLaVA VQA, and R-RefCOCO(+/g) in a 7:2:1 ratio for each training cycle. For FP-RefCOCO(+/g) we specifically maintained a 9:1 ratio of true to false queries to ensure a balanced focus on the model's 'say' and segmentation tasks. Training with R-RefCOCO(+/g) conditions the model to dismiss false premises without proposing alternatives. Relying solely on FP-RefCOCO could lead the model to generate speculative outputs, such as arbitrarily altering 'left' to 'right', without genuine image analysis. This issue is detailed in our ablation study (Fig. 6). Finally, the LLaVA VQA dataset, focused solely on visual question answering, is integrated to retain the model's competence in this field.

This simple yet effective approach enables the fine-tuned model to acquire new segmentation skills while preserving its innate "see" and "say" abilities. Unlike previous methods [37, 66] that used an auxiliary branch for binary responses to detect true/false premise queries, our approach seamlessly integrates the "see" and "say" abilities within the inherent capabilities of the LMM. This integration results in a more streamlined model capable of multitasking. The enhanced LMM demonstrates an improved ability to not only accurately segment objects in images while also engaging in intelligent dialogue, handling both existent and non-existent objects with equal finesse.

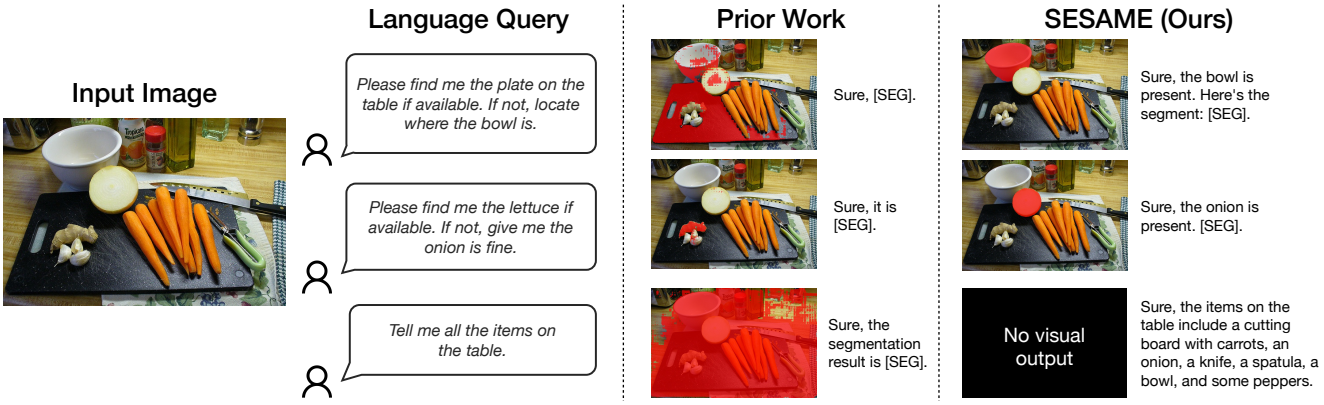

Figure 4. In contrast to prior work (the output of the LISA [33] is shown above), SESAME is able to handle more complicated conditional reasoning and instruction, and is able to not output a segment when it is not requested.

5. Experimental Results

Implementation Details. *SESAME*'s LMM backbone, LLaVA-v1.5-7B, and segmentation decoder, SAM, are fine-tuned for 10 epochs via LoRA [21] using the unified training dataset described in Sec. 4. Our loss function combines dice and binary cross-entropy losses for segmentation, along with cross-entropy loss for sentence prediction tasks. We employed the AdamW optimizer with a learning rate of 3e-4, setting the batch size to 10 and the gradient accumulation steps to 5. The total training time was approximately 24 hours on a single DGX A100 80GB GPU. To ensure fair comparisons, following [33], we carefully excluded images from the training set that were also present in the test or validation sets. This step was crucial to avoid data contamination, especially as we merged FP-RefCOCO, FP-RefCOCO+, and FP-RefCOCOg into a unified training set, each having unique data splits. We will make our code available for future research and applications.

5.1. Results

In our experiments, we assessed the "See", "Say", and "Segment" capabilities of *SESAME*, our cascading method (combining the off-the-shelf LLaVA-v1.5-7B and LISA-7B), and the baseline model LISA-7B [33]. We also compared two non-LMM methods including GRES [37] and [84]. Results are reported on the val sets of FP-RefCOCO(+/g) in Tab. 2.

Detection. Models' ability to "See" is evaluated using binary classification accuracy, a metric that determines whether models can accurately discern the presence or absence of an object referenced in an image. In this evaluation, *SESAME* achieves an accuracy of 79.84%, outperforming both our cascading method (75.59%) and existing baselines. LISA-7B, expectedly, underperforms due to its constant erroneous assumption that the prompted object exists in the scene. The superior detection accuracy of *SESAME* when compared to the cascading method stems from our enhanced fine-tuning approach. This approach integrates a more extensive and balanced collection of both positive and negative referring expression training data, unlike the standard VQA dataset used in the conventional LLaVA-v1.5 fine-tuning process.

Description. We evaluate a model's ability to "Say" through a modified CLAIR metric [3]. CLAIR uses an LLM (OpenAI's GPT-4 [48]) to provide scores for candidate captions compared against a reference caption set, outputting a similarity score in $[0, 1]$. We use the model's corrected referring expression as the candidate and compare it against the set of positive sample referring expressions belonging to the same referent in the image. We modified CLAIR to return the score of the best match to the reference set. When there are multiple objects with referring expression annotations in the same image, we only score the suggested correction against expressions for the same object instance.

The cascading method, employing LLaVA for the "say" function, demonstrates superior performance, surpassing the finetuned LISA which appears to lose all the capability. This finding underscores the problem in LISA's fine-tuning process. In contrast, *SESAME* achieves much higher CLAIR scores (0.63, 0.61, 0.67) than the cascading method (0.35, 0.42, 0.55) on all three datasets. We theorize this is because *SESAME* is fine-tuned on our customized false premise data with ground truth false premise corrections that encourage the model to learn to suggest the same referent rather than other irrelevant objects in the image.

Segmentation. The segmentation score is measured using the cIoU metric, following established protocols in prior basic and false-premise referring segmentation studies [33, 37, 66]. This metric assesses the cumulative intersection over the cumulative union. As expected, as *SESAME* has the best "See" ability, it consistently skips the creation of segmentation masks for false premise queries, thereby achieving the highest cIoU scores. Conversely, existing baseline methods generally produced segmentation masks for most false premise queries, leading to poor cIoU scores.

Figure 5. Not only is SESAME robust to false premises, and does not attempt to incorrectly predict a segmentation mask when an object or concept is not actually present in the image, but it is able to use commonsense reasoning to suggest relevant objects or concepts when a similar instance is present, and output the segmentation mask of that instance.

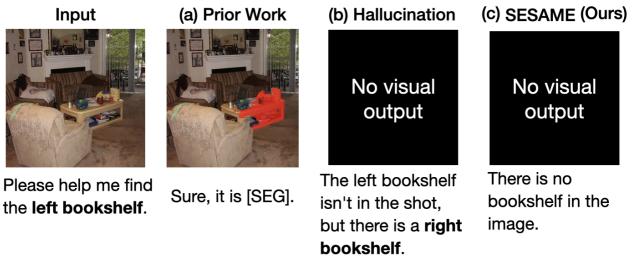

Figure 6. Ablation Studies: (a) Prior work hallucinates a segmentation even though there is no bookcase. (b) If we force the model to correct all the false premise queries, it correctly detects that there is no left bookcase in the image, but still hallucinates a "right bookcase" in the text response, likely because relational expressions are often reversed (e.g., "left" to "right") when modified to form negative samples. (c) *SESAME* addresses the hallucination problem by adding R-RefCOCO(+/g) into the unified train set, and allowing the model to simply respond that the requested object was not found rather than requiring it to provide an alternative expression.

5.2. Discussions

Proportion of False-premise Queries. An essential part of our analysis involved varying the proportion of false premise queries in our test set, as detailed in Tab. 3. A 0% false premise (FP) scenario is equivalent to the RefCOCOg evaluation, while a 50% FP mirrors the FP-RefCOCOg dataset. These findings underscore that despite fine-tuning, models, including *SESAME*, still have considerable potential for improvement, specifically in detecting false premises (FP) with the highest Recall in false premise query being only 67.89%. This capability is particularly crucial as increasing FP proportions directly impacts the performance of downstream segmentation cIoU score.

Handling Complex Instructions. A particularly notable example in Fig. 4 showcases *SESAME*'s ability to process and respond to complex user prompts. This includes segmenting an alternative object based on a conditional query and engaging in VQA only without generating a segmentation mask. In contrast, previous models like LISA lacked these capabilities, severely limiting their human-like interaction potential. This finding also suggests that *SESAME* could be extended to multi-round interactions, where a user might request an intelligent agent to first summarize a scene and then focus on segmenting specific objects of interest.

Significance of False Premise Rejection. We also investigated the impact of integrating the R-RefCOCO(+/g) dataset, specifically designed for false premise rejection, into our training process. Excluding these data often led models to rely on superficial word modifications in their responses instead of genuinely interpreting the image context. This reliance aggravated the issue of hallucination and resulted in lower scores in the "Say" capability. A striking illustration of this phenomenon is presented in Fig. 6.

Method	refCOCO	refCOCO+	refCOCOg
MCN [41]	62.4	50.6	49.2
VLT [13]	67.5	56.3	55.0
CRIS [64]	70.5	62.3	59.9
LAVT [70]	72.7	62.1	61.2
ReLA [37]	73.8	**66.0**	65.0
X-Decoder [83]	-	-	64.6
SEEM [84]	-	-	65.7
LISA-7B [33]	**74.9**	65.1	**67.9**
SESAME (Ours)	74.7	64.9	66.1

Table 4. Even though our method is trained to do both see, say, and segment simultaneously, our model is still on par with prior methods on natural setting.

Performance in Referring Segmentation Benchmarks. Finally, we assessed *SESAME* in traditional referring segmentation benchmarks with only positive queries. The results in Tab. 4 demonstrated the comparison between our method and several existing approaches. While our model is adept at handling false premises and enhancing dialogue interaction, it does not compromise the segmentation abilities. This indicates that our joint training approach, which fine-tunes LMMs with a tailored dataset, successfully achieves great segmentation capabilities while maintaining robust performance in LMM's basic ability to see and say.

6. Conclusion

In this study, we tackle the overlooked issue within the realm of LMMs: false premise segmentation queries. We not only highlight this challenge in existing LMM methodologies but also introduce a pioneering task known as False Premise Correction, necessitating capabilities to "See," "Say," and "Segment." Alongside this new task, we present FP-RefCOCO(+/g), a specially designed dataset for evaluating LMMs on these essential skills. To address this challenge, we employ innovative cascading and joint training techniques. Our integrated LMM, *SESAME*, demonstrates substantial improvement in detecting the presence of objects ("see"), advising users about non-existent objects or modifying queries accordingly ("say"), and precisely segmenting objects that are present in the image ("segment"). This research fills a critical gap in LMM capabilities and sets a strong foundation for future explorations into improving LMM interactions in diverse and real-world applications.

Acknowledgements

This work was supported in part by funding from the United States Department of Deference and the BAIR Commons program.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *TPAMI*, 2017. 2

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 4

[3] David Chan, Suzanne Petryk, Joseph E Gonzalez, Trevor Darrell, and John Canny. Clair: Evaluating image captions with large language models. *arXiv preprint arXiv:2310.12971*, 2023. 6

[4] Chi Chen, Ruoyu Qin, Fuwen Luo, Xiaoyue Mi, Peng Li, Maosong Sun, and Yang Liu. Position-enhanced visual instruction tuning for multimodal large language models. *arXiv preprint arXiv:2308.13437*, 2023. 2

[5] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 2

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 2018. 2

[7] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *CVPR*, 2020. 2

[8] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *NeurIPS*, 2021. 2

[9] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 2

[10] Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for text-to-image generation and evaluation. *arXiv preprint arXiv:2305.15328*, 2023. 3

[11] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge J. Belongie. Learning to evaluate image captioning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5804–5812. IEEE Computer Society, 2018. 3

[12] Ernest Davis. Unanswerable questions about images and texts. *Frontiers in Artificial Intelligence*, 3:51, 2020. 2

[13] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, 2021. 8

[14] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vlt: Vision-language transformer and query generation for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[15] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015. 2

[16] Yansong Feng and Mirella Lapata. Automatic caption generation for news images. *IEEE transactions on pattern analysis and machine intelligence*, 35(4):797–812, 2012. 2

[17] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019. 2

[18] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 3

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2

[20] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528. Association for Computational Linguistics, 2021. 3

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv:2106.09685*, 2021. 6

[22] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 108–124. Springer, 2016. 1

[23] Jianqiang Huang, Yu Qin, Jiaxin Qi, Qianru Sun, and Hanwang Zhang. Deconfounded visual grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 998–1006, 2022. 2

[24] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019. 2

[25] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 3

[26] Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. TIGEr: Text-to-image grounding for image caption evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2141–2152. Association for Computational Linguistics, 2019. 3

[27] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 3

[28] Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh, and Mohamed Coulibali. NUBIA: NeUral based interchangeability assessor for text generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 28–37. Association for Computational Linguistics, 2020. 3

[29] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 2

[30] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, 2019. 2

[31] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv:2304.02643*, 2023. 5

[32] Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. In *CVPR*, 2021. 2

[33] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 1, 2, 4, 5, 6, 8

[34] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Dynamic language binding in relational visual reasoning. *arXiv preprint arXiv:2004.14603*, 2020. 2

[35] Mengdi Li, Cornelius Weber, and Stefan Wermter. Neural networks for detecting irrelevant questions during visual question answering. In *Artificial Neural Networks and Machine Learning–ICANN 2020: 29th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 15–18, 2020, Proceedings, Part II 29*, pages 786–797. Springer, 2020. 3

[36] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation. In *CVPR*, 2021. 2

[37] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23592–23601, 2023. 1, 3, 5, 6, 8

[38] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 4, 5

[39] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4185–4194, 2019. 3

[40] Wei Liu, Andrew Rabinovich, and Alexander C. Berg. Parsenet: Looking wider to see better. *arXiv*, 2015. 2

[41] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *CVPR*, 2020. 8

[42] Aroma Mahendru, Viraj Prabhu, Akrit Mohapatra, Dhruv Batra, and Stefan Lee. The promise of premise: Harnessing question premises in visual question answering. *arXiv preprint arXiv:1705.00601*, 2017. 3

[43] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 2, 4

[44] Akib Mashrur, Wei Luo, Nayyar A Zaidi, and Antonio Robles-Kelly. Semantic multi-modal reprojection for robust visual question answering. In *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–6. IEEE, 2022. 3

[45] Akib Mashrur, Wei Luo, Nayyar A Zaidi, and Antonio Robles-Kelly. Robust visual question answering via semantic cross modal augmentation. *Computer Vision and Image Understanding*, page 103862, 2023. 3

[46] Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P Spithourakis, and Lucy Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. *arXiv preprint arXiv:1701.08251*, 2017. 2

[47] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *ICCV*, 2015. 2

[48] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. 6

[49] OpenAI. Gpt-4v(ision) system card, 2023. 4

[50] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011. 2

[51] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv:2306.14824*, 2023. 2

[52] Prakruthi Prabhakar, Nitish Kulkarni, and Linghao Zhang. Question relevance in visual question answering. *arXiv preprint arXiv:1807.08435*, 2018. 3

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 3

[54] Arijit Ray, Gordon Christie, Mohit Bansal, Dhruv Batra, and Devi Parikh. Question relevance in vqa: Identifying non-visual and false-premise questions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 919–924, 2016. 2

[55] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 2

[56] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2

[57] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *TPAMI*, 2017. 2

[58] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, 2017. 3

[59] Jiajin Tang, Ge Zheng, Cheng Shi, and Sibei Yang. Contrastive grouping with transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23570–23580, 2023. 2

[60] Zhuotao Tian, Pengguang Chen, Xin Lai, Li Jiang, Shu Liu, Hengshuang Zhao, Bei Yu, Ming-Chang Yang, and Jiaya Jia. Adaptive perspective distillation for semantic segmentation. *TPAMI*, 2022. 2

[61] Zhuotao Tian, Jiequan Cui, Li Jiang, Xiaojuan Qi, Xin Lai, Yixin Chen, Shu Liu, and Jiaya Jia. Learning context-aware classifier for semantic segmentation. *AAAI*, 2023. 2

[62] Andeep S Toor, Harry Wechsler, and Michele Nappi. Question action relevance and editing for visual question answering. *Multimedia Tools and Applications*, 78:2921–2935, 2019. 2

[63] Jianming Wang, Enjie Cui, Kunliang Liu, Yukuan Sun, Jiayu Liang, Chunmiao Yuan, Xiaojie Duan, Guanghao Jin, and Tae-Sun Chung. Referring expression comprehension model with matching detection and linguistic feedback. *IET Computer Vision*, 14(8):625–633, 2020. 1, 3

[64] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *CVPR*, 2022. 8

[65] Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision*, pages 148–166. Springer, 2022. 3

[66] Jianzong Wu, Xiangtai Li, Xia Li, Henghui Ding, Yunhai Tong, and Dacheng Tao. Towards robust referring image segmentation. *arXiv preprint arXiv:2209.09554*, 2022. 1, 3, 5, 6

[67] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *CVPR*, 2019. 2

[68] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 2

[69] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, 2018. 2

[70] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, 2022. 2, 8

[71] Yanzhi Yi, Hangyu Deng, and Jinglu Hu. Improving image captioning evaluation by considering inter references variance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 985–994. Association for Computational Linguistics, 2020. 3

[72] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity, 2023. 2

[73] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 2

[74] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 2, 4

[75] Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual object detection with multimodal large language models. *arXiv preprint arXiv:2305.18279*, 2023. 2

[76] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv:2307.03601*, 2023. 2

[77] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *NeurIPS*, 2021. 2

[78] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 2

[79] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *ECCV*, 2018.

[80] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *ECCV*, 2018. 2

[81] Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*, 2023. 2

[82] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *ICCV*, 2019. 2

[83] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *CVPR*, 2023. 2, 8

[84] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *arXiv:2304.06718*, 2023. 2, 5, 6, 8