

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

Supplementary Material

A. Extension to Reasoning Segmentation Tasks

Our main paper highlights the effectiveness of *SESAME* in answering queries, with or without false-premises, for open-language segmentation tasks. The method operates by “seeing” whether the referred object is present in an image, “saying” what the correct grounding is, and “segmenting” the image using the appropriate input prompt. While our initial focus is on false-premise referring segmentation tasks, our method also encompasses reasoning segmentation tasks.

Reasoning segmentation tasks [4] represent a more complex challenge compared to traditional semantic or referring segmentation tasks. These tasks necessitate advanced reasoning and world knowledge—models need to understand complex queries with intricate expressions or longer sentences. In such scenarios, models are tasked with not only identifying objects in an image but also comprehending and reasoning about the broader context and the relationships depicted within the scene.

Setup. In line with the method outlined in Sec. 4 of the main paper, we initially developed a specialized dataset for false-premise reasoning segmentation, which includes both training and validation components. This dataset, with an equal number of false-premise queries and the original true queries, was derived from the original one proposed by [4]. As shown in Fig. 1 (a), for each image within this dataset, we randomly selected language queries that correspond to other images, along with one original question and its corresponding image. These elements were then used as inputs for LLaVA [6] which was tasked with generating plausible question-and-answer pairs.

For the training of *SESAME*, we incorporated several datasets. These include ADE20K [9], COCO-stuff [2], and LVIS-PACO part segmentation [7] for semantic segmentation in addition to the reasoning segmentation dataset and the unified dataset as described in Sec. 4 of the main paper. Specifically, for the semantic segmentation datasets, we utilized a template-based method to create false-premise query-answer pairs, as depicted in Fig. 1 (b). The training split for the reasoning segmentation dataset followed the previously detailed procedure.

The architecture of the *SESAME* model is the same as described in Sec. 4. We allocated the training dataset in a ratio of 10:3:1, spanning the semantic segmentation dataset, the unified dataset (as detailed in Sec. 4), and our augmented reasoning segmentation dataset. We deliberately employed a 9 times higher sampling rate for true premise queries compared to false-premise ones to ensure effective training for

segmentation tasks. The training adheres to the same hyper-parameters and procedures as detailed in Sec. 5 of the main paper.

Method	See		Segment
	FP Recall	TP Recall	cIoU
LISA	0.5	100.0	42.40
LLaVA + LISA (Ours)	7.0	99.5	44.96
<i>SESAME</i> (Ours)	86.5	90.0	51.43

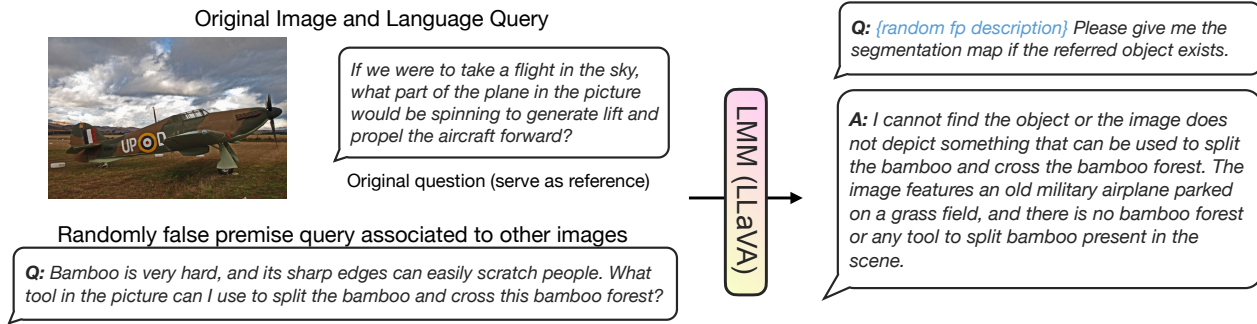
Table 1. In addition to referring segmentation tasks detailed in Tab. 2 of our main paper, *SESAME* (ours) also exhibits significant gain in reasoning segmentation tasks, hugely surpasses both our chained model methods and the LISA baseline by a large margin.

Quantitative Evaluation. When evaluating LMMs with our augmented official validation set of the reasoning segmentation dataset, we primarily focus on assessing the “See” and “Segment” components of our method. Given the inherent challenges in evaluating the “Say” component using the CLAIR score [3] — which typically relies on explicit expressions similar to captions (e.g., “a trash can”) rather than the implicit references characteristic of reasoning segmentation tasks (e.g., “the place we throw the garbage”) — we decided not to include this component in our quantitative analysis. However, to illustrate this aspect of our model’s capabilities, qualitative examples are provided in the subsequent section.

In evaluating the “seeing” ability of our model, we reported the recall for both true and false premise queries concerning object existence in the image. Segmentation performance was measured using the cIoU metric, consistent with the metrics utilized in Tab. 2 and Tab. 3 of our main paper. The results of this evaluation are presented in Tab. 1, where we compare our *SESAME* method with both the LISA baseline and our proposed chained method.

The data reveals that while the chained method outperforms the LISA baseline, our *SESAME* method shows even more significant advancements. Specifically, it achieved an impressive 12 times increase in object detection (See) scores for false premise queries with a slight decrease in performance for true premise queries. *SESAME* also demonstrated a notable improvement in segmentation performance, exceeding the chained method and the LISA baseline by over 7% in terms of cIoU. These findings highlight the critical importance of incorporating false-premise queries into model training and illustrate the wide-ranging effectiveness and adaptability of our method across various segmentation subtasks.

(a) False-premise Reasoning Segmentation Dataset



(b) False-premise Semantic Segmentation Dataset

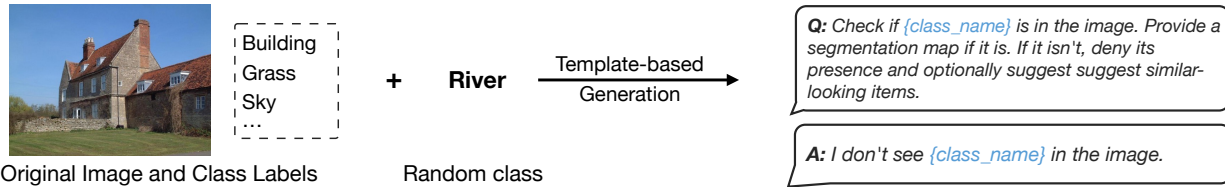


Figure 1. (a) We employ LLaVA to generate false-premise question-answer pairs for the reasoning segmentation dataset. (b) We use a template-based method to create false-premise semantic segmentation data from well-known datasets, including ADE20K, COCO-stuff, and LVIS-PACO part segmentation. Leveraging these datasets, along with the unified dataset which encompasses our curated FP-RefCOCO(+g) datasets as detailed in Sec. 4, *SESAME* effectively demonstrates its proficiency in handling false-premise queries within the realm of reasoning segmentation tasks as shown in Tab. 1.

Qualitative Results. As illustrated in Fig. 2, we showcase a selection of question-answer pairs generated by *SESAME* on the reasoning segmentation dataset. These examples demonstrate the model’s proficiency in the “Say” component, particularly in addressing complex queries.

In the first and second rows, *SESAME* successfully identifies that the referred objects do not exist in the image and provides appropriate suggestions in response. This ability to accurately deny the existence of objects and offer relevant alternatives showcases a significant advancement over prior methods.

In the third example, when presented with an entirely irrelevant query, *SESAME* demonstrates its ability to appropriately deny the query without offering unnecessary suggestions. This highlights the model’s nuanced understanding and response generation capability. In contrast, the LISA baseline method, when presented with false premise queries, tends to generate segmentation maps without the capacity to offer alternative responses in all cases.

However, as indicated in the fourth row, there is room for improvement. While *SESAME* exhibits a strong ability to deny non-existent objects, it occasionally generates hallucinated corrected results. This means that the model can occasionally generate a “corrected” premise that is still false. Addressing this issue will be a focus of our future work, aiming to enhance the model’s accuracy and reliability in

generating corrected premises.

B. More Qualitative Results

In Fig. 3, we present additional visual comparisons between *SESAME* and previous methods in the context of referring segmentation tasks. A notable strength of *SESAME* lies in its capacity to reject and, where appropriate, correct queries founded on erroneous assumptions. This capability is particularly evident in instances where the actual object bears a certain relevance to the falsely presumed item in the query, encompassing similarities in objects, attributes, or actions.

C. Discussion on the Generated Dataset

In Fig. 3, we conceptually illustrate the differences in the generation process between our FP-RefCOCO(+g) referring dataset and the traditional R-RefCOCO(+g) dataset. In addition to enhancing the context-awareness in generating false premises, our dataset, with the aid of LLM-based augmentation, is an “open set” with the following definition: it includes diverse objects and concepts that extend beyond the scope of the COCO dataset and even surpass those in RefCOCO(+g). This marks a departure from the “close-set” datasets generated through simple word replacement as seen in the R-RefCOCO(+g) dataset. As illustrated through concrete examples in Fig. 4, LLMs are capable of substituting multiple instances within sentences, introducing novel and

plausible false-premise items such as elf hats, cotton candy, and platypuses, which are absent in existing datasets.

Regarding the quality of our LLM-generated dataset, we have taken meticulous steps to mitigate generating incorrect phrases as detailed in Tab. 3. By feeding all existing referring expressions per image into the LLMs and utilizing train-of-thought prompting techniques [8], the LLMs can accurately recognize these expressions and effectively prevent the generation of "false false-premise" phrases. Although some noise might remain within the dataset, our SESAME model consistently outperforms in all evaluated tasks, as demonstrated in Tab. 2. This achievement underscores the high potential of our SESAME method, although we acknowledge that further refinement of the dataset's quality remains a valuable direction for future research.

D. Our Prompts to the LLM and LMM

Our method is anchored by two specialized prompts. The first, as elaborated in Tab. 2, underlies our cascading method. Specifically, we input the prompt into LLaVA-v1.5 [5] to obtain the "see and say" results. A key aspect of this process involves identifying non-existent predictions: if the output begins with "No, there is no...", we classify these as non-existent predictions and collect their corresponding sentences to assess the "say" score. This prompt, designed to elicit explanatory responses through chain-of-thought prompting techniques [8], has shown significantly greater efficacy compared to simpler user prompts like "Please help me segment [the referring expression] in the image?"

The second prompt used in our dataset development is outlined in Tab. 3. This prompt includes carefully crafted instructions and utilizes a chain-of-thought approach [8] in its in-context examples. This design is important to minimize the generation of false premise sentences that either deviate significantly from the input object or reference an existing one.

```
1 Analyze the image and verify if there are any
  referred objects in the image. Yes or no
  with explanations. Here's an example. When
  asked "is there any green car behind the man
  in the image", you can answer the question
  in ways like:
2
3 1. If the object exists, confirm it:
4   - "Yes, the green car behind the man is
      present in the image."
5
6 2. If not, deny the existence of the object and
  optionally provide alternative suggestions:
7   - "No, there is no green car in the image. Did
      you mean the red car in front of the man?"
8
9 Now, my question is: "Is there [the referring
  expression] in the image?" I value a precise
  and detailed analysis. Please inspect the
  image thoroughly and respond according to
  the guidelines provided above.
```

Table 2. Our full prompt for the LLaVA-v1.5 model [5] to obtain the result of see and say.

```

1 ## Your Role: Prank Expert
2
3 ## Objective
4 Turn real object descriptions of an image into fictional but relevant ones by altering specific
  elements.
5
6 ## Guidelines
7 - Change only one word in each sentence.
8 - Use unique word replacements in each sentence.
9 - Focus on modifying the main subject, its attributes, actions, or relationships to another relevant
  counterpart.
10 - Ensure altered descriptions do not coincide with any real objects in the original description.
11 - Be cautious when changing adjectives related to position (e.g., near/far, left/right) and size
    (e.g., small/large) to avoid ambiguity and inadvertent overlap with existing items.
12
13 ## Example 1:
14 Original: ["The red ball to the left of the blue toy.", "The man in a white shirt standing next to a
  woman with an umbrella.", "The smallest dog in the group, near the tree."]
15 Altered: ["The yellow ball to the left of the blue toy.", "The woman in a blue shirt standing next to
  a woman with an umbrella.", "The smallest cat in the group, near the tree."]
16 Reasoning: Changes focus on the main subject (ball color, person's gender, animal type) while ensuring
  uniqueness and avoiding overlap with real objects.
17
18 ## Example 2:
19 Original: ["a man getting ready to cut a cake", "guy in green with a knife in the right hand picture",
  "woman pointing at ice cream", "a woman in a blue shirt with floral print", "the man standing up
  and pointing"]
20 Altered: ["a kangaroo getting ready to cut a cake", "guy in purple with a knife in the right hand
  picture", "woman pointing at a pizza", "a woman in a blue shirt with stripes", "the man standing
  up and stretching"]
21 Reasoning: Each alteration (animal for person, color change, relevant object swap, pattern change,
  action change) ensures a distinct and fictional transformation while maintaining the sentence
  structure.
22
23 ## Example 3:
24 Original: ["the vase on the right", "a woman wearing a blue shirt", "a woman in a grey blue sweatshirt
  painting a figure onto a vase", "man on right", "man in red shirt", "a large brown urn being
  decorated by a woman", "a large urn on a green table and the lady is drawing on it"]
25 Altered: ["the vase in the back", "a baby wearing a blue shirt", "a woman in a grey blue sweatshirt
  jogging on the sidewalk", "man on the top", "man in orange shirt", "a large school bus being
  decorated by a woman", "a large urn on a couch and the lady is drawing on it"]
26 Reasoning: The changes made here (position, subject identity, action, location) are carefully chosen
  to create fictional scenarios without referring to other existing objects in the original
  description.
27
28 Your Turn Now! Adhere to the guidelines and answer the question!
29
30 Original: [A list of referring sentences of a single COCO image]
31 Altered:

```

Table 3. Our full prompt to the GPT-3.5-turbo model [1] for augmenting false-premise referring expression. LLM starts completion from “Altered:”.


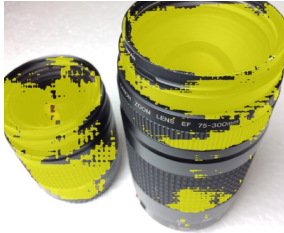










Input	Prior Work	SESAME (Ours)
		
<p>Please segment something that produces pollen in this image if it exists.</p>	<p>Sure, [SEG].</p>	<p>I can't find something that produce pollen, but something that takes picture is certainly present.</p>
		
<p>Please segment the place where piano players should sit in this image if it exists. Otherwise, tell me the object doesn't exist and optionally offer an alternative.</p>	<p>Sure, it is [SEG].</p>	<p>I can't confirm piano players should sit, but there's car players should sit.</p>
		
<p>Please segment the objects that can protect the snail and prevent it from getting injured in this image if it exists.</p>	<p>Sure, it is [SEG].</p>	<p>I can't find the objects that can protect the snail and prevent it from getting injured in this image.</p>
		
<p>When admiring a beautiful sunset, what part of the picture would we most likely focus on? Please output segmentation mask if it exists. Otherwise, describe the image in detailed.</p>	<p>Sure, [SEG].</p>	<p>I cannot find a beautiful sunset in the image. The image features a woman with curly hair looking out of a window.</p>

Figure 2. In contrast to prior work (the output of the LISA [4] is shown above), *SESAME* typically succeeds in rectifying complex input queries and refrains from generating a segment when unnecessary.









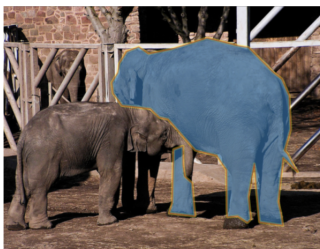
Type	Input	Prior Work	SESAME (Ours)
<i>Relevant Objects</i>			
	Help me locate the cookies on the table.	Sure, it is [SEG].	There are no cookies here, but the image does have fries .
<i>Irrelevant Objects</i>			
	Can you find the people in this image?	[SEG].	No, there are no people in this image.
<i>Similar Attributes</i>			
	Please segment the purple bus in the image.	Sure, [SEG].	The purple school bus isn't in the image, but the yellow school bus is.
<i>Associated Actions</i>			
	Help me locate the man in yellow in the image.	Sure, it is [SEG].	The man in yellow isn't in the image, but the man in blue is.
<i>Associated Actions</i>			
	Please segment the person playing basketball .	Sure, it is [SEG].	The person playing basketball isn't in the shot, but the person skateboarding is.

Figure 3. *SESAME* not only robustly identifies and refrains from predicting segmentation masks for non-existent objects or concepts in an image but also intelligently suggests and outputs segmentation masks for similar, relevant instances through commonsense reasoning.



Real: biggest elephant in the image
LLM: biggest **platypus** in the image



Real: skier in teal blue pants
LLM: skier in **lavender** pants



Real: woman in santa hat
LLM: woman in **elf hat**



Real: woman holding a pastry
between two men eating ice cream
LLM: woman holding a **donut**
between two men eating **cotton candy**



Real: a computer monitor to the
right of two other monitors
LLM: a computer **printer** to the
right of two other **printers**



Real: a woman holding a tennis
racket in the right image
LLM: a woman holding a **badminton**
racket in the right image

Figure 4. Some concrete examples in our FP-RefCOCO(+g) dataset.

References

- [1] 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](#)
- [2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, 2018. [1](#)
- [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. [1](#)
- [4] 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](#), [5](#)
- [5] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. [3](#)
- [6] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv:2304.08485*, 2023. [1](#)
- [7] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *CVPR*, 2023. [1](#)
- [8] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. [3](#)
- [9] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. [1](#)