LISA

Language
Instructed
Segmentation
Assistant

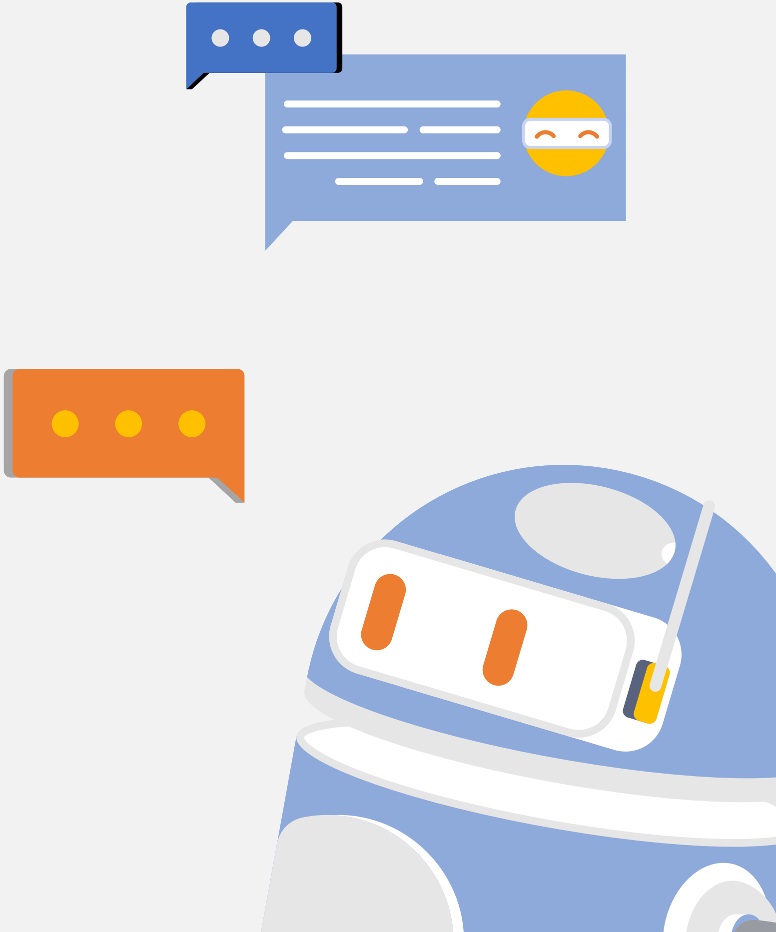
Davide Cavicchini

Matteo Ranzetti

01 →

LISA

Language-**I**nstructed **S**egmentation **A**ssistant



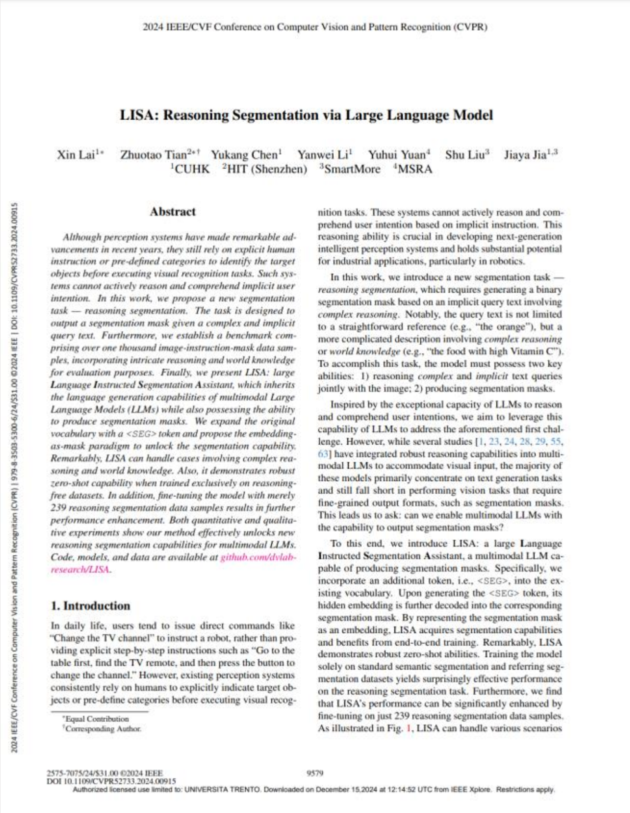
Reasoning Segmentation

Given an input image and a query instruction, output the binary segmentation mask corresponding to the query.

Unlike the referring segmentation task, these queries should involve **reasoning**.

REFERRING SEGMENTATION
"Where is the orange in this image?"

REASONING SEGMENTATION
"Which fruit in the image has the most Vitamin C?"



2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

LISA: Reasoning Segmentation via Large Language Model

Xin Lai[1*] Zhuotao Tian[2*†] Yukang Chen[1] Yanwei Li[1] Yuhui Yuan[4] Shu Liu[3] Jiaya Jia[1,3]
[1]CUHK [2]HIT (Shenzhen) [3]SmartMore [4]MSRA

2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-5300-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/CVPR52733.2024.00915

Abstract

*Although perception systems have made remarkable advancements in recent years, they still rely on explicit human instruction or pre-defined categories to identify the target objects before executing visual recognition tasks. Such systems cannot actively reason and comprehend implicit user intention. In this work, we propose a new segmentation task — reasoning segmentation. The task is designed to output a segmentation mask given a complex and implicit query text. Furthermore, we establish a benchmark comprising over one thousand image-instruction-mask data samples, incorporating intricate reasoning and world knowledge for evaluation purposes. Finally, we present LISA: large **L**anguage **I**nstructed **S**egmentation **A**ssistant, which inherits the language generation capabilities of multimodal Large Language Models (LLMs) while also possessing the ability to produce segmentation masks. We expand the original vocabulary with a <SEG> token and propose the embedding-as-mask paradigm to unlock the segmentation capability. Remarkably, LISA can handle cases involving complex reasoning and world knowledge. Also, it demonstrates robust zero-shot capability when trained exclusively on reasoning-free datasets. In addition, fine-tuning the model with merely 239 reasoning segmentation data samples results in further performance enhancement. Both quantitative and qualitative experiments show our method effectively unlocks new reasoning segmentation capabilities for multimodal LLMs. Code, models, and data are available at github.com/dvlab-research/LISA.*

1. Introduction

In daily life, users tend to issue direct commands like "Change the TV channel" to instruct a robot, rather than providing explicit step-by-step instructions such as "Go to the table first, find the TV remote, and then press the button to change the channel." However, existing perception systems consistently rely on humans to explicitly indicate target objects or pre-define categories before executing visual recognition tasks. These systems cannot actively reason and comprehend user intention based on implicit instruction. This reasoning ability is crucial in developing next-generation intelligent perception systems and holds substantial potential for industrial applications, particularly in robotics.

In this work, we introduce a new segmentation task — *reasoning segmentation*, which requires generating a binary segmentation mask based on an implicit query text involving *complex reasoning*. Notably, the query text is not limited to a straightforward reference (e.g., "the orange"), but a more complicated description involving *complex reasoning* or *world knowledge* (e.g., "the food with high Vitamin C"). To accomplish this task, the model must possess two key abilities: 1) reasoning *complex* and *implicit* text queries jointly with the image; 2) producing segmentation masks.

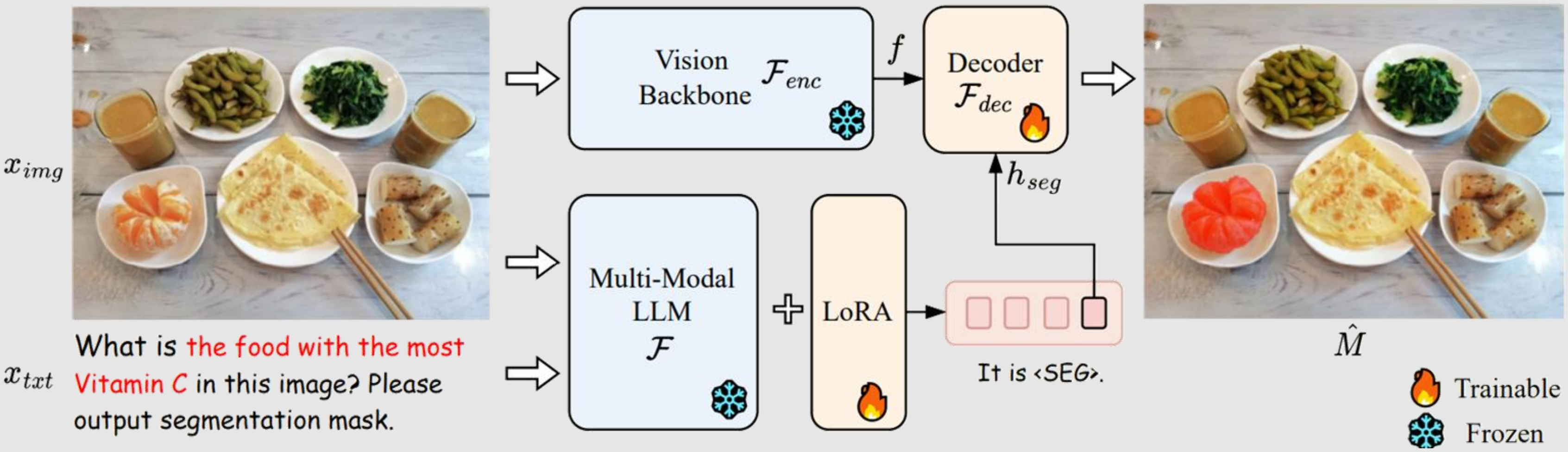
Inspired by the exceptional capacity of LLMs to reason and comprehend user intentions, we aim to leverage this capability of LLMs to address the aforementioned first challenge. However, while several studies [1, 23, 24, 28, 29, 55, 63] have integrated robust reasoning capabilities into multimodal LLMs to accommodate visual input, the majority of these models primarily concentrate on text generation tasks and still fall short in performing vision tasks that require fine-grained output formats, such as segmentation masks. This leads us to ask: can we enable multimodal LLMs with the capability to output segmentation masks?

To this end, we introduce LISA: a large **L**anguage **I**nstructed **S**egmentation **A**ssistant, a multimodal LLM capable of producing segmentation masks. Specifically, we incorporate an additional token, i.e., <SEG>, into the existing vocabulary. Upon generating the <SEG> token, its hidden embedding is further decoded into the corresponding segmentation mask. By representing the segmentation mask as an embedding, LISA acquires segmentation capabilities and benefits from end-to-end training. Remarkably, LISA demonstrates robust zero-shot abilities. Training the model solely on standard semantic segmentation and referring segmentation datasets yields surprisingly effective performance on the reasoning segmentation task. Furthermore, we find that LISA's performance can be significantly enhanced by fine-tuning on just 239 reasoning segmentation data samples. As illustrated in Fig. 1, LISA can handle various scenarios

*Equal Contribution
†Corresponding Author.

2575-7075/24/$31.00 ©2024 IEEE
DOI 10.1109/CVPR52733.2024.00915

LISA: The Model



- Extend the vocabulary with a <SEG> token
- Feed a query and an image to the multimodal LLM
- If the LLM generates a <SEG> token, decode it to a binary mask

$$
\begin{aligned}
\mathcal{L}_{txt} &= \mathbf{CE}(\hat{\mathbf{y}}_{txt}, \mathbf{y}_{txt}), \\
\mathcal{L}_{mask} &= \lambda_{bce}\mathbf{BCE}(\hat{\mathbf{M}}, \mathbf{M}) + \lambda_{dice}\mathbf{DICE}(\hat{\mathbf{M}}, \mathbf{M}).
\end{aligned} \tag{4}
$$

Training **Data**

Semantic Segmentation

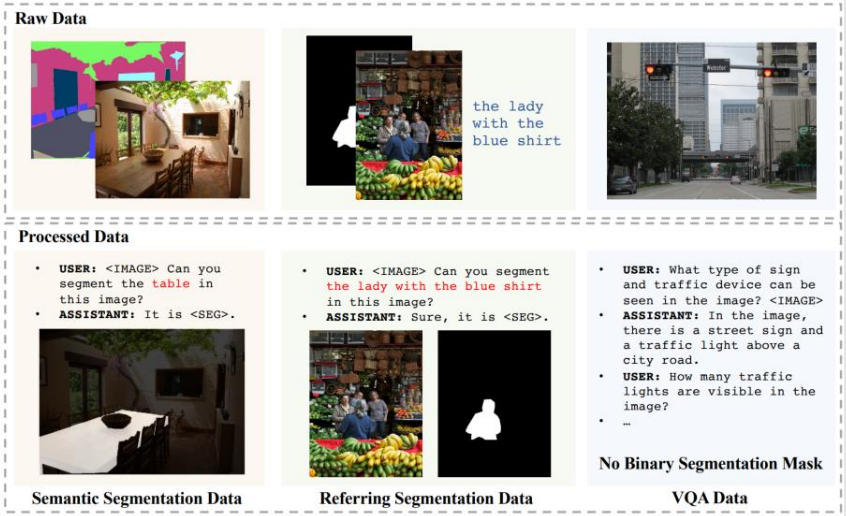
- ADE20K
- COCO-Stuff
- LVIS-PACO

Referring Segmentation

- refCOCO
- refCOCO+
- refCOCOg
- refCLEF

Visual Question Answering

- LLaVa-Instruct-150k
- LLaVa-V1.5-mix665K



ReasonSeg

Images with questions and segmentation mask output for each image

- 239 train images (with additional text responses)
- 200 val images
- 779 test images

Used to fine-tune LISA and released for benchmarking purposes

- **USER:** <IMAGE> What is the place where the driver can observe the speed in this image? Please output segmentation mask.
- **ASSISTANT:** <SEG>.

- **USER:** <IMAGE> Who is the loser in this image? Please output segmentation mask and explain why.
- **ASSISTANT:** Sure, it is <SEG>. ... The man on top is likely the winner of the match, as he has control over the other fighters body...

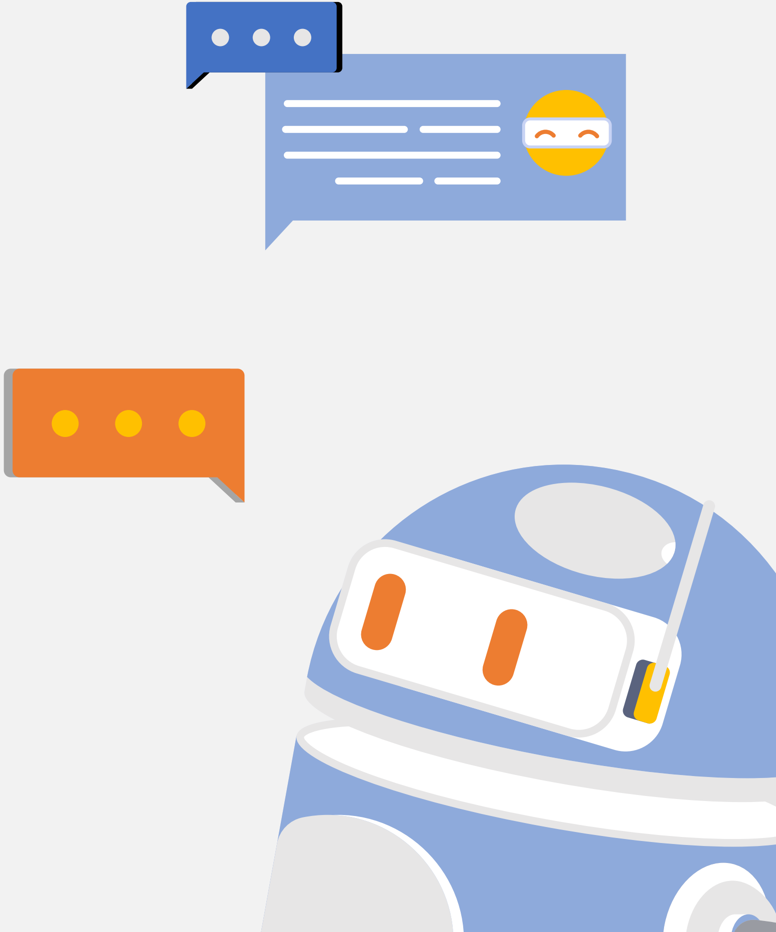
- **USER:** <IMAGE> Can you segment the food that tastes not spicy in this image?
- **ASSISTANT:** <SEG>.

- **USER:** <IMAGE> Can you segment the unusual part in this image and explain why?
- **ASSISTANT:** Sure, it is <SEG>. In the image, the unusual part is the dog wearing a reindeer antler headband.

02 →

LISA

Limitations

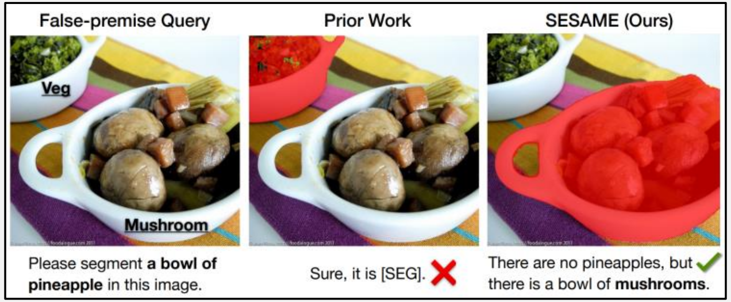


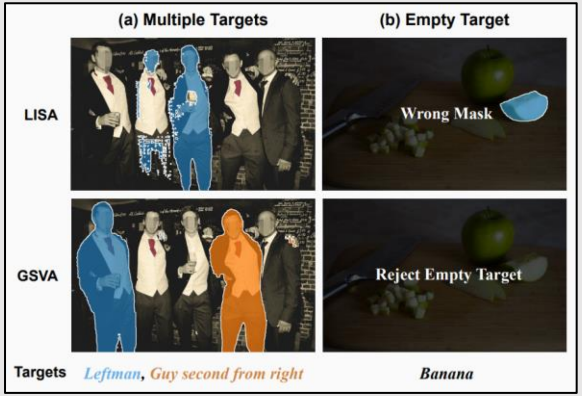
Explored Issues

Wu, Tsung-Han et al. "See, Say, and Segment: Teaching LMMs to Overcome False Premises." *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023): 13459-13469.

FALSE PREMISES

when queries imply the existence of something that is not actually present in the image.





Xia, Zhuofan et al. "GSVA: Generalized Segmentation via Multimodal Large Language Models." *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023): 3858-3869.

MULTIPLE TARGETS

It also has some problems with generating multiple targets.

Proposed Solutions

Better dataset

[REJ] token

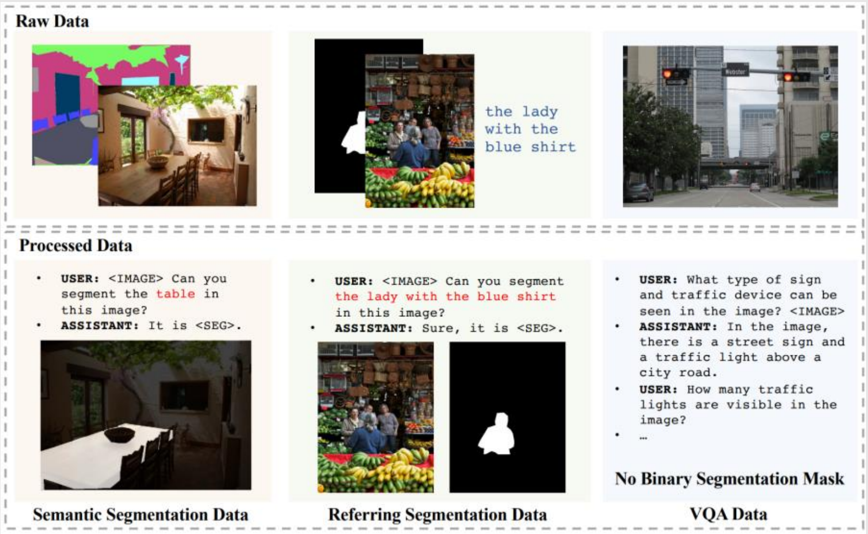
Prompting

Chain of Reasoning

Proposed **Solutions**

Better dataset

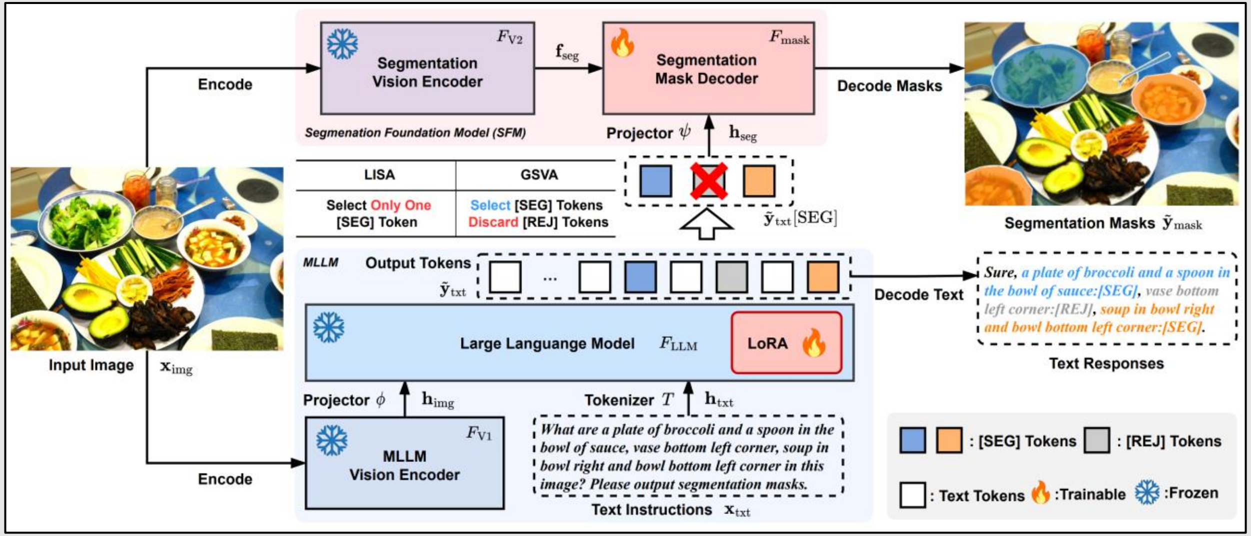
From	Segment A	→	Here it is <SEG>
To	Segment A and B	→	Here is A <SEG> and B <SEG>
	Segment C	→	C is not in the image



Proposed Solutions

[REJ] token

Train the model to refuse the generation using:
- Improved Dataset
- New [REJ] Token



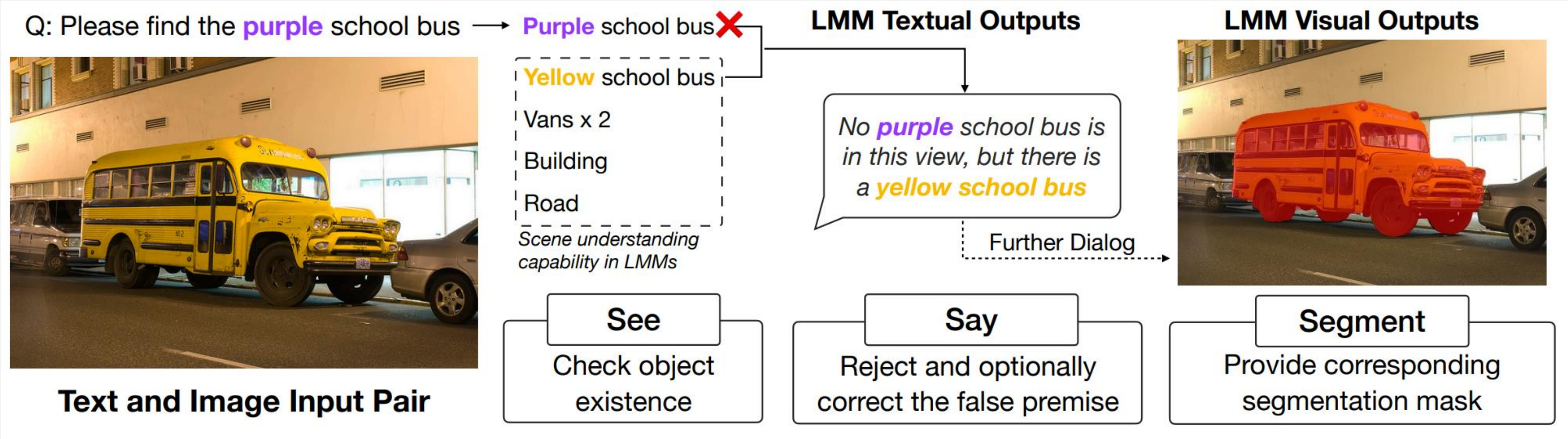
Xia, Zhuofan et al. "GSVA: Generalized Segmentation via Multimodal Large Language Models." *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023): 3858-3869.

Proposed **Solutions**

Prompting

Walk the model to the solution:

- SEE
- SAY
- SEGMENT

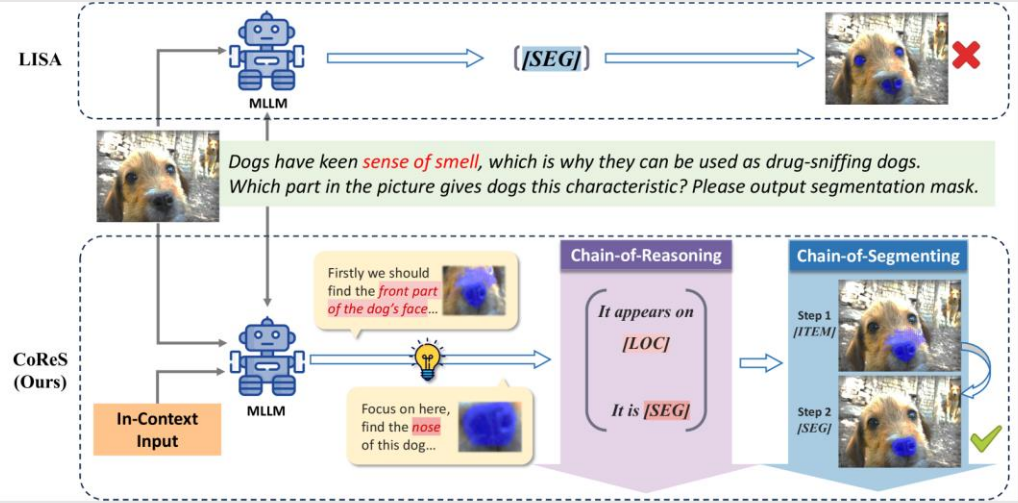


Wu, Tsung-Han et al. "See, Say, and Segment: Teaching LMMs to Overcome False Premises."

Proposed Solutions

Chain of Reasoning

Uses Chain of Thought in the segmentation domain:
- Coarse Location of the object
- Fine Grained Segmentation



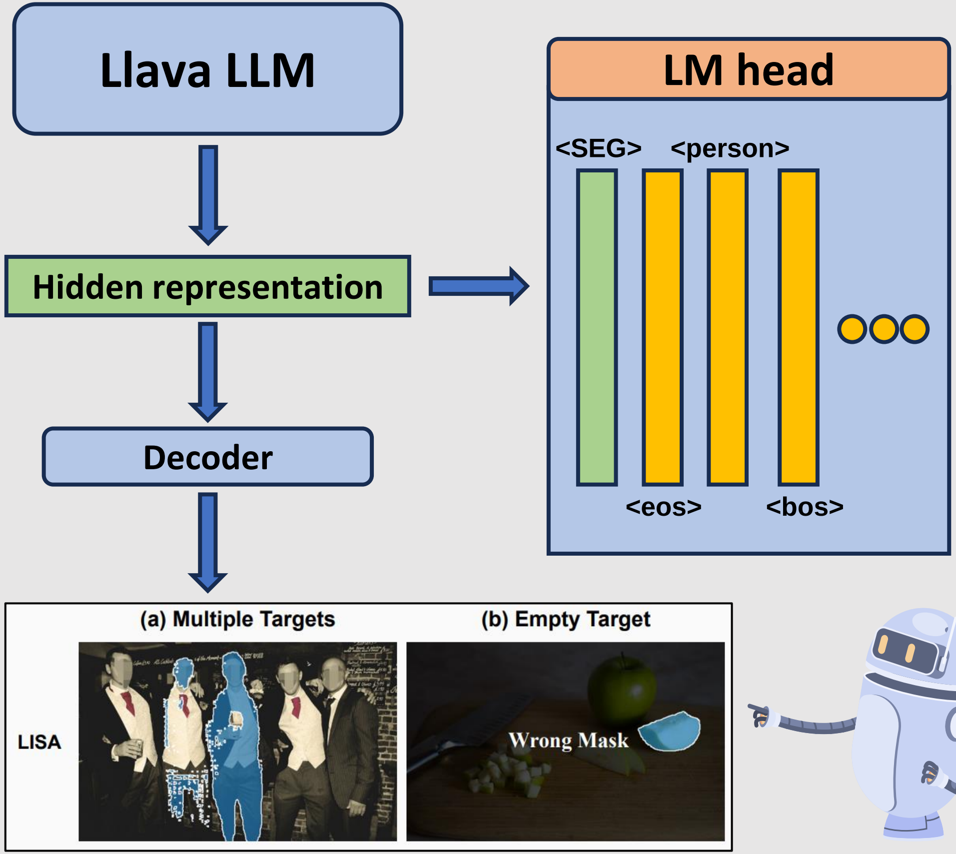
Bao, Xiaoyi et al. "CoReS: Orchestrating the Dance of Reasoning and Segmentation."

Limited Expressivity

$$\max_i(HiddenRep \cdot TokenEmb_i) = TokenID(< SEG >)$$

➕

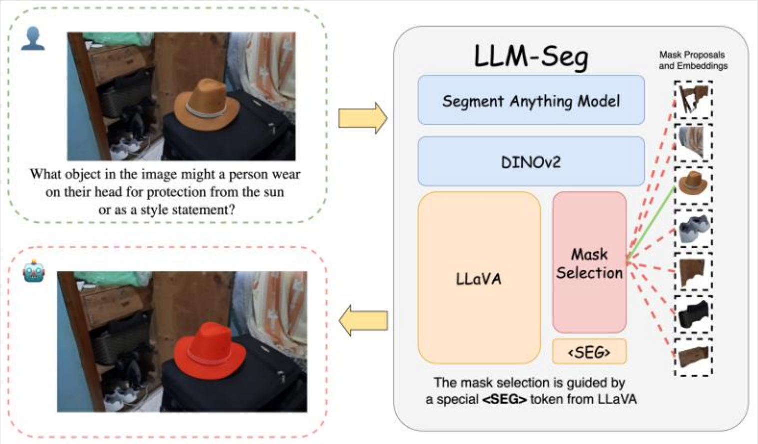
$HiddenRep$ encode for the position in the image



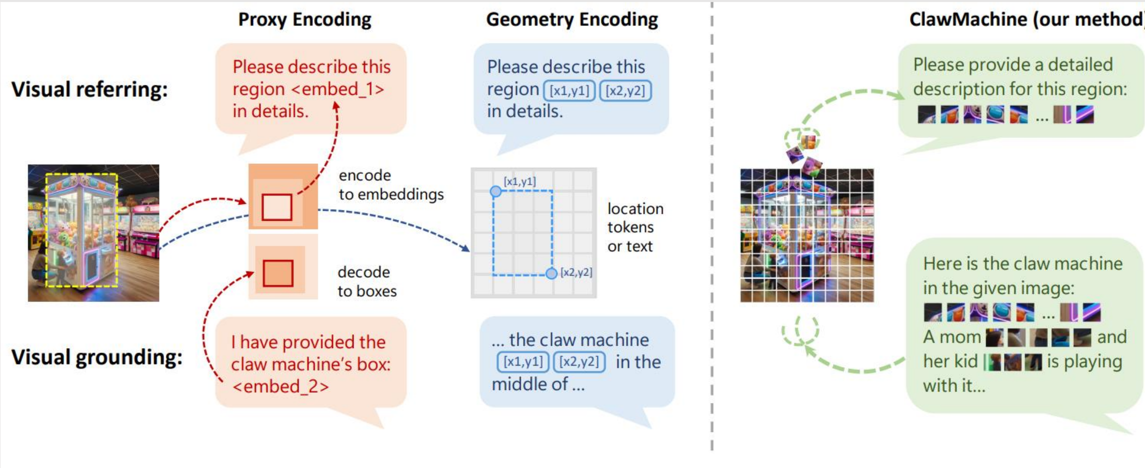
Different Approaches

LLM-Seg:
Extract Segmentation
Classify <SEG> token

ClawMachine:
Use Image Patches as tokens the model can generate



Wang, Junchi and Lei Ke. "LLM-Seg: Bridging Image Segmentation and Large Language Model Reasoning." *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2024): 1765-1774.



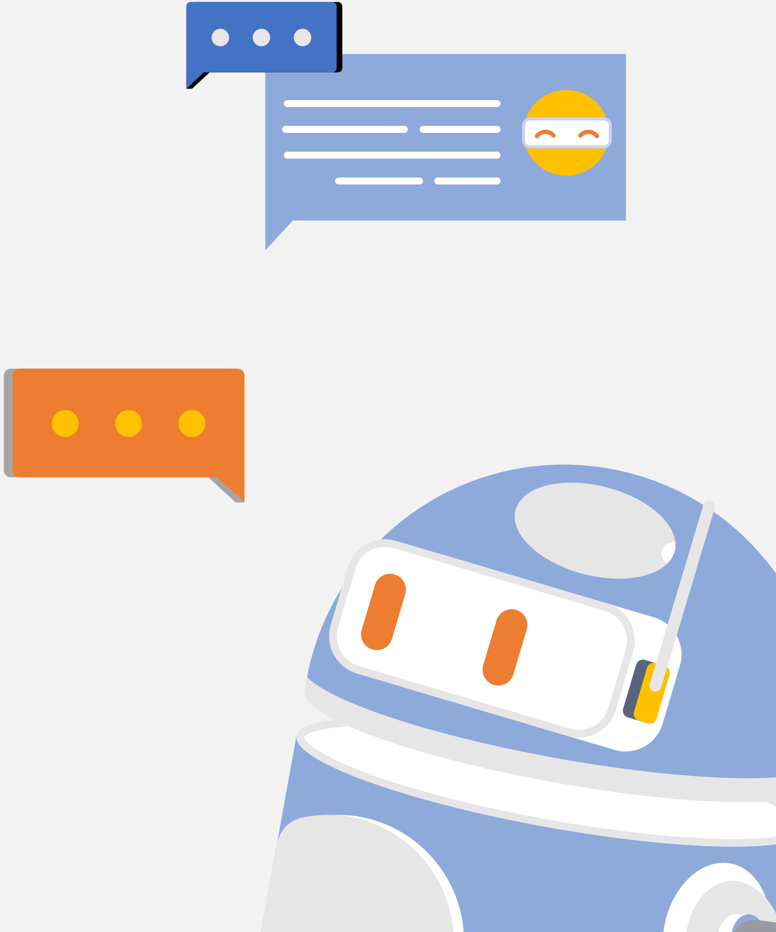
Ma, Tianren et al. "ClawMachine: Fetching Visual Tokens as An Entity for Referring and Grounding."

03 →

ELISA

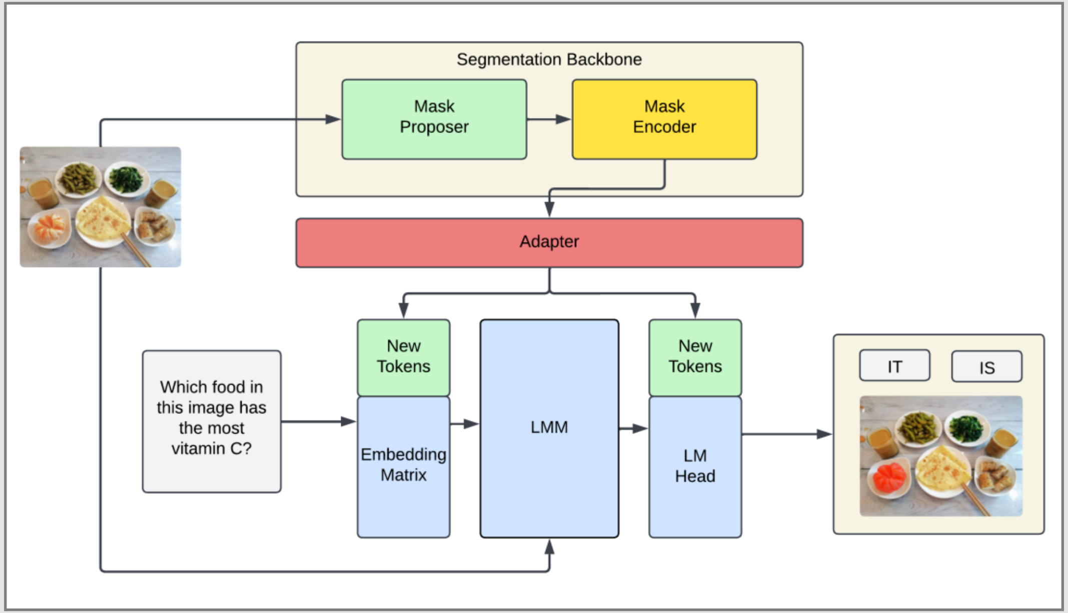
Extractive
Language-**I**nstructed
Segmentation **A**ssistant

> Thought about acronym for segmentation model for 6 seconds
>
> A fitting evolution of the name could be "**ELISA**", standing for Extractive Language-Instructed Segmentation Assistant. This acronym preserves the connection to the original LISA framework while highlighting the key addition—the extractive pipeline—used to dynamically enrich the model's vocabulary and reasoning capabilities.



Extracting not Generating

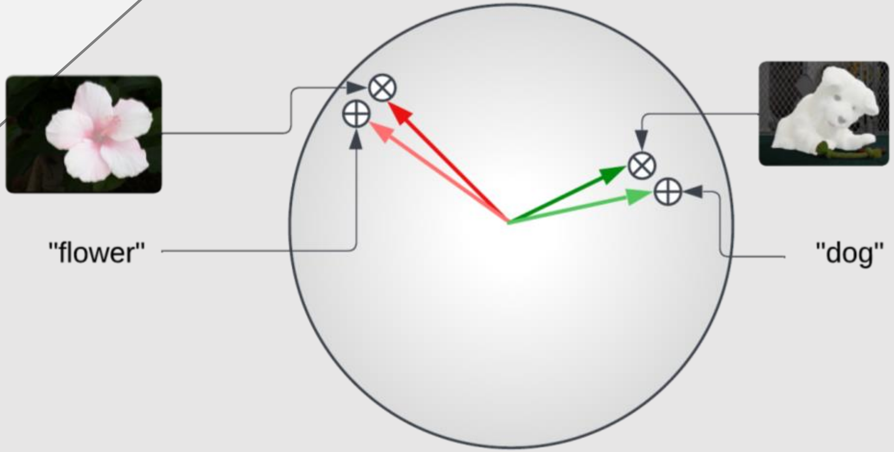
By compelling the model to **choose** from the masks we propose, we plan on circumventing the **limited expressivity** problem.

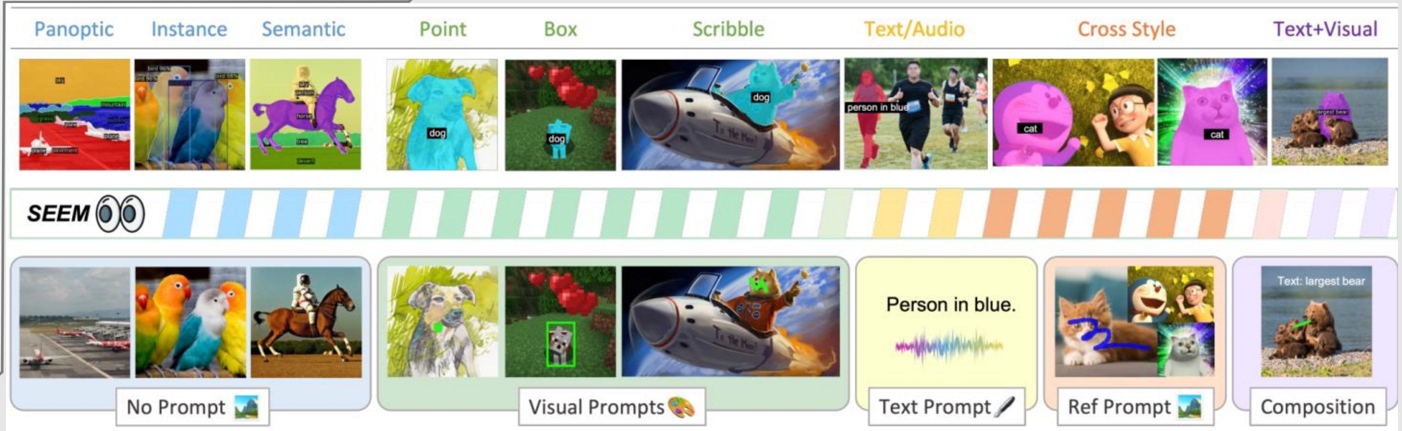


We chain some easily replaceable components to propose the masks, then **dynamically extend** the vocabulary of the chosen LLM.

Hopes & Dreams

We hope that the model will overlap the embeddings of the masks with the language concepts that they represent, allowing touse the tokens **interleaved** the text.

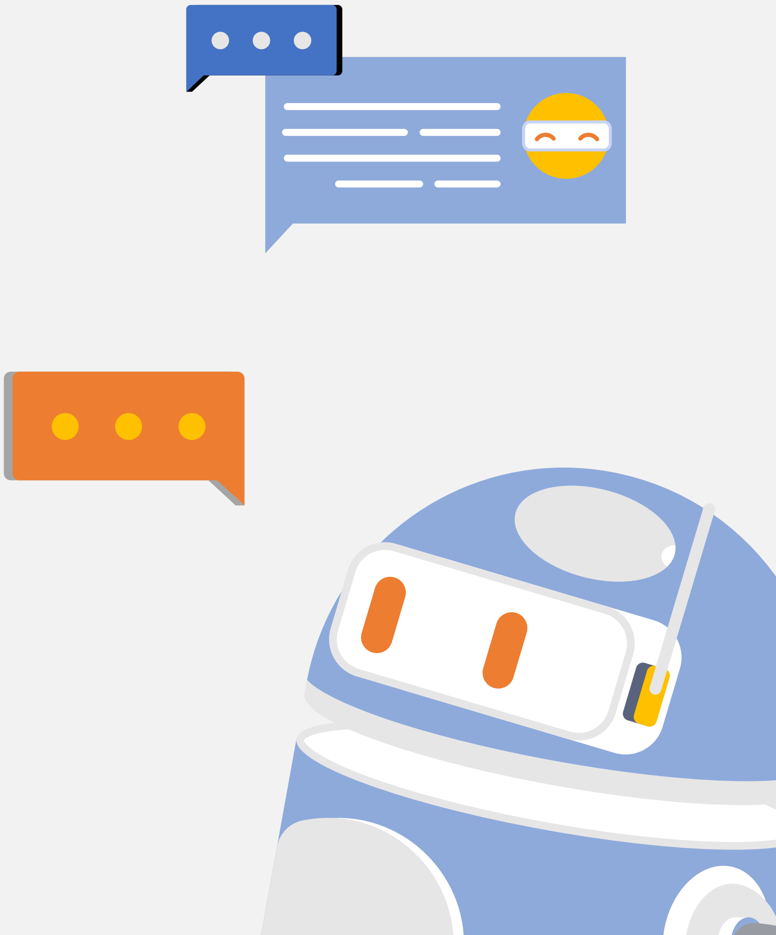




Zou, Xueyan et al. "Segment Everything Everywhere All at Once." *ArXiv* abs/2304.06718 (2023): n. pag.

04 →

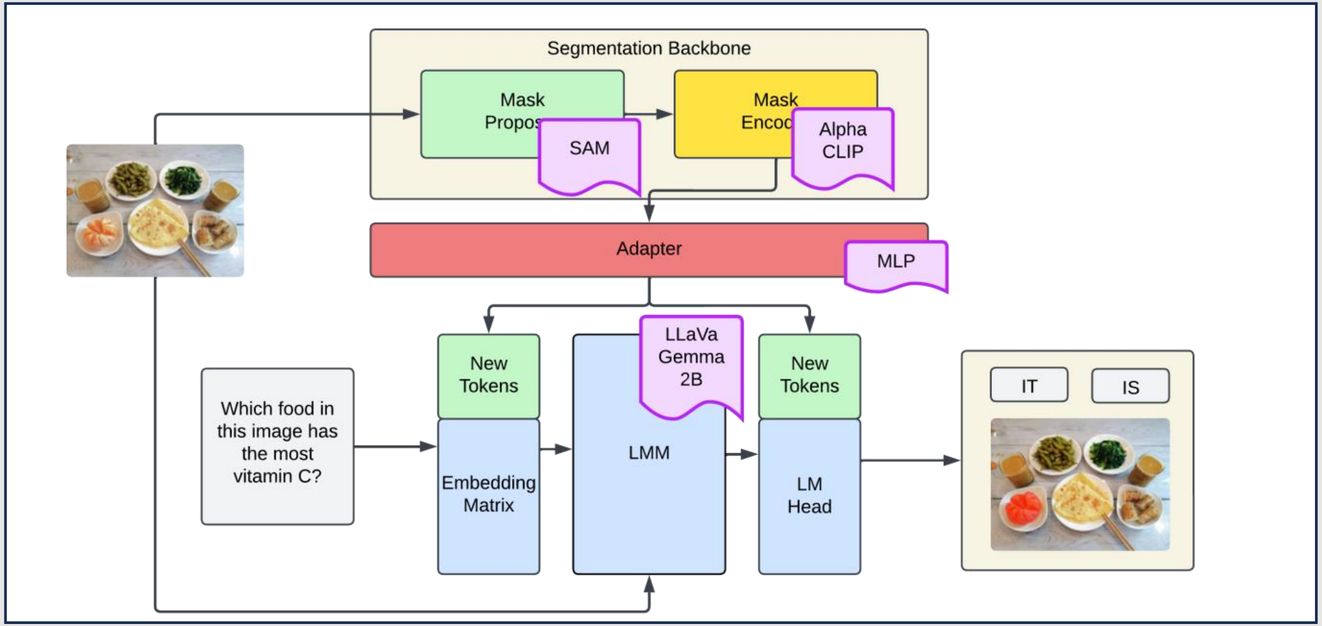
ELISA Implementation



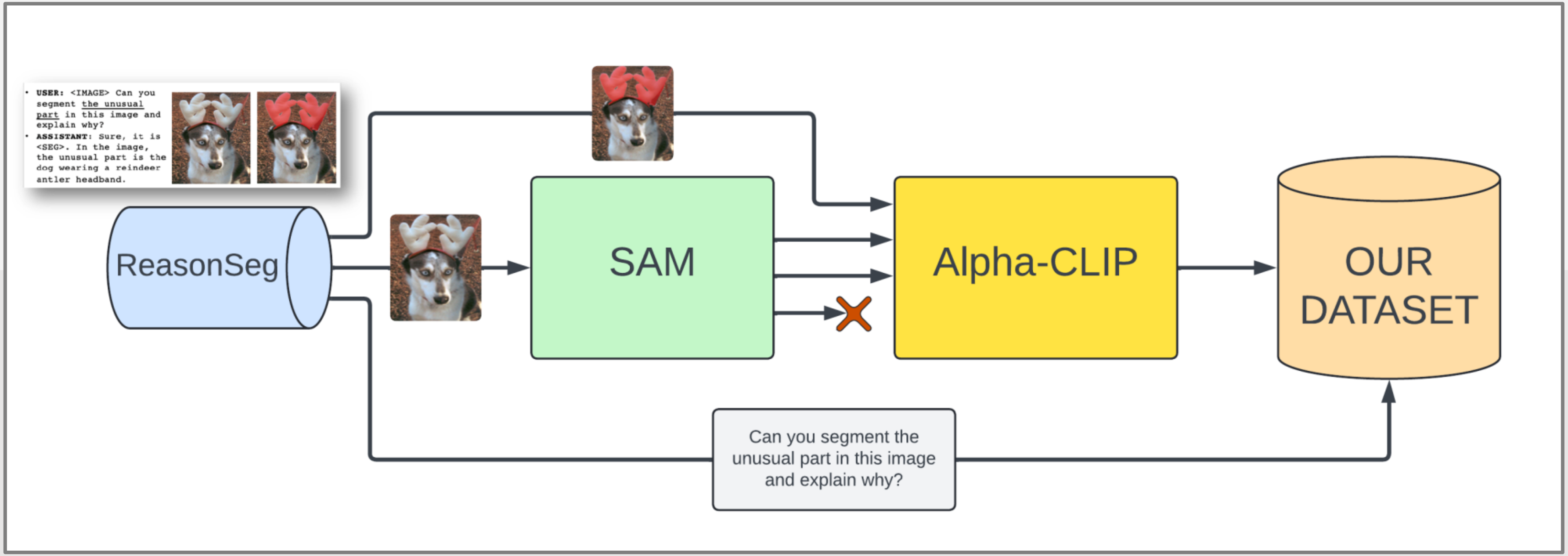
Our Implementation Choices

Throwback to the architecture page:

Our choices for the parts of this model are shown over the components.



Dataset & Preprocessing

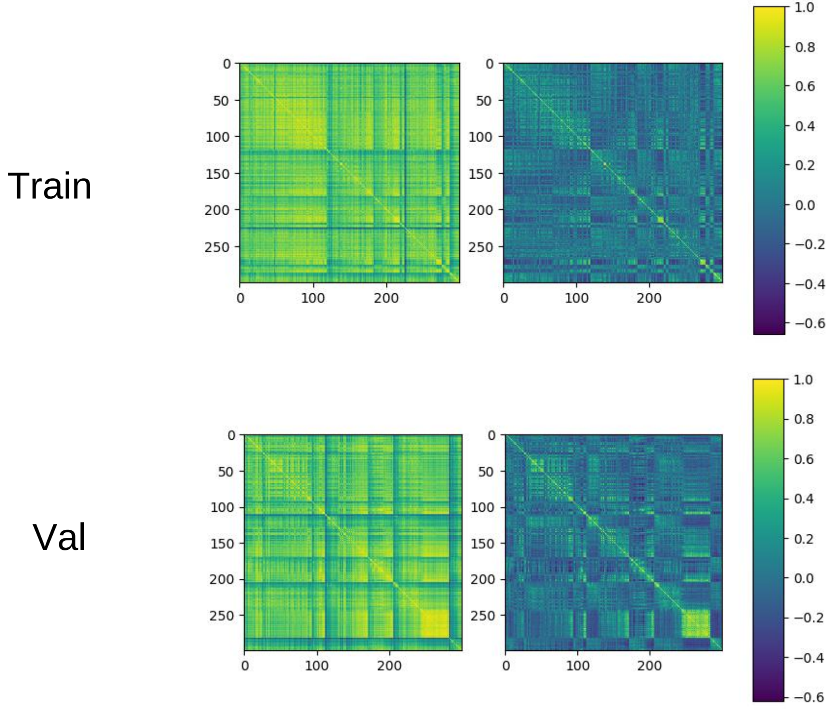


AlphaCLIP Issues

The embeddings of all the masks in an image are very similar to each other.

We normalize them using the mean of the masks in the training set to address this issue.

Cosine Similarities between masks, before and after normalization



Training setup

GPU

- RTX 3090 (24GB)

Optimizer:

- ADAM

Scheduler

- Cosine Annealing

We still couldn’t fit everything in the GPU, so we had to quantize stuff.

Training setup

GPU

- RTX 3090 (24GB)

Optimizer:

- ADAM
- Quantized to 8bit

Scheduler

- Cosine Annealing

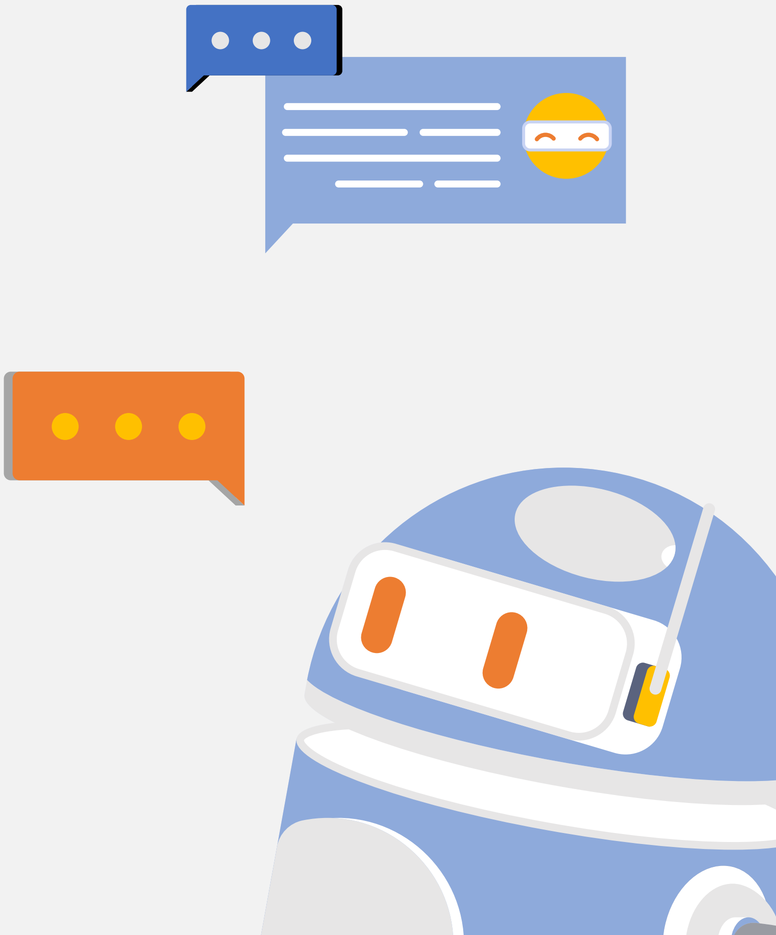
We still couldn’t fit everything in the GPU, so we had to quantize stuff.

LLaVa-Gemma-2B

- Quantized to 4bit

05 →

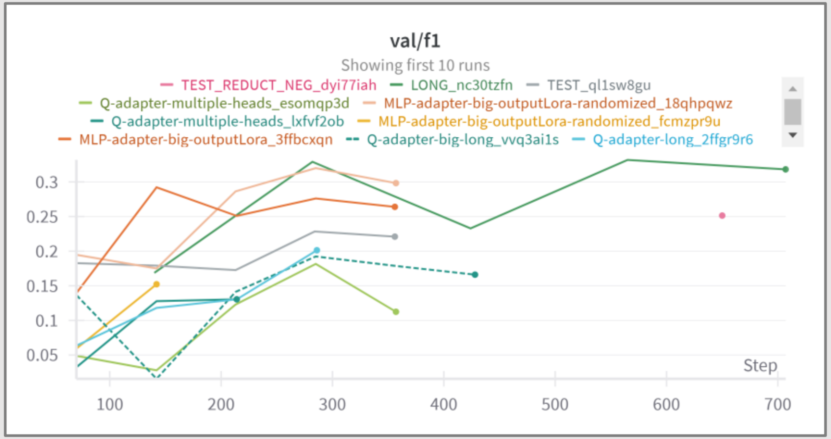
ELISA Results

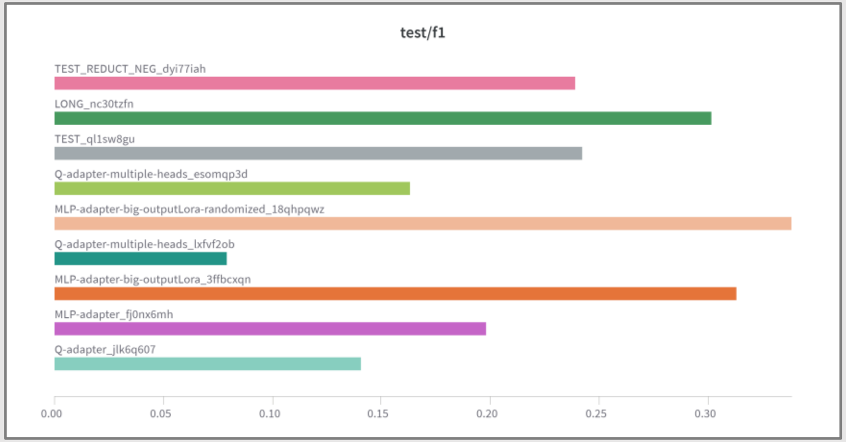


Results

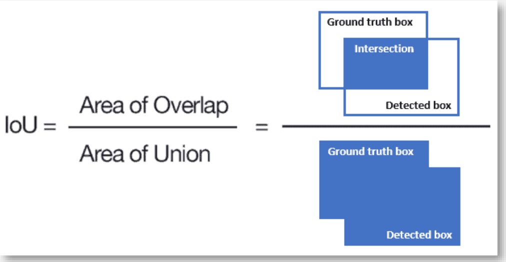
Model	Val Precision	Val Recall	Val F1
Rand50	6,1	50	10,2
RandClass	3,9	3,9	4,0
ELISA	33,2	39,0	32,4

Model	Test Precision	Test Recall	Test F1
Rand50	4,5	50	7,8
RandClass	3,7	4,1	3,9
ELISA	30,5	34,5	30,1





Results - comparison

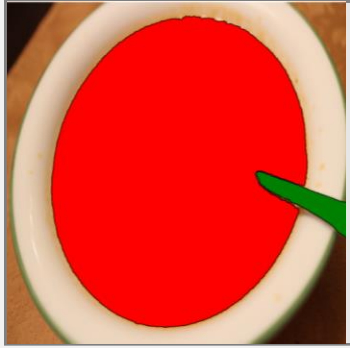


Method	val		test					
	overall		short query		long query		overall	
	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU
OVSeg [26]	28.5	18.6	18.0	15.5	28.7	22.5	26.1	20.8
GRES [27]	22.4	19.9	17.6	15.0	22.6	23.8	21.3	22.0
X-Decoder [65]	22.6	17.9	20.4	11.6	22.2	17.5	21.7	16.3
SEEM [66]	25.5	21.2	20.1	11.5	25.6	20.8	24.3	18.7
Grounded-SAM [30]	26.0	14.5	17.8	10.8	22.4	18.6	21.3	16.4
LISA-7B	44.4	46.0	37.6	34.4	36.6	34.7	36.8	34.1
LISA-7B (ft)	52.9	54.0	40.6	40.6	49.4	51.0	47.3	48.4
LISA-13B	48.9	46.9	39.9	43.3	46.4	46.5	44.8	45.8
LISA-13B (ft)	56.2	62.9	44.3	42.0	54.0	54.3	51.7	51.1
LLaVA1.5-7B + OVSeg	38.2	23.5	24.2	18.7	44.6	37.1	39.7	31.8
LISA-7B-LLaVA1.5	53.6	52.3	47.1	48.5	49.2	48.9	48.7	48.8
LISA-7B-LLaVA1.5 (ft)	61.3	62.9	48.3	46.3	57.9	59.7	55.6	56.9
LLaVA1.5-13B + OVSeg	37.9	26.4	27.1	19.4	46.1	40.6	41.5	34.1
LISA-13B-LLaVA1.5	57.7	60.3	50.8	50.0	54.7	50.9	53.8	50.8
LISA-13B-LLaVA1.5 (ft)	**65.0**	**72.9**	**55.4**	**50.6**	**63.2**	**65.3**	**61.3**	**62.2**

	Validation		
	cIoU	**gIoU**	**IoU**
ELISA (top50 SAM)	15,3	6,2	12,5
ELISA (top50 SAM + GT)	36,8	13,8	27,7

Promising Results

It **learned** to insert them into the text!
Meaning it's somewhat learning the actual **meaning** and **aligning** them to the text embeddings of the original model.



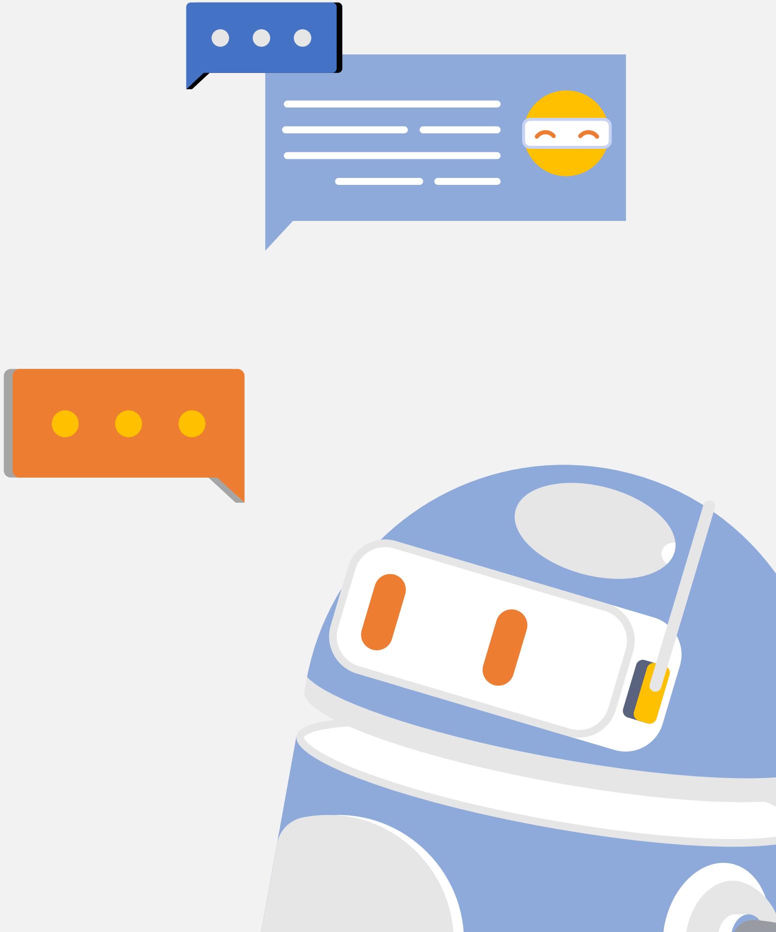
<SEG_MASK_6>, when eating a bowl of soup, a spoon would most likely be used to scoop and consume the soup.**<SEG_MASK_1>**. In the image, there is a spoon placed next to the bowl of soup, which suggests that it is the most likely utensil to be used for eating the soup. [...]

The object that can be used to store and transport various sundries and small household items during the move is a **<SEG_MASK_1>**. In the image, there is a cardboard box sitting next to **<SEG_MASK_1>**. Cardboard boxes are known for their sturdiness and ease of transportation. They can be used to securely store […]

06 →

ELISA

Future Work



Known Problems & Solutions

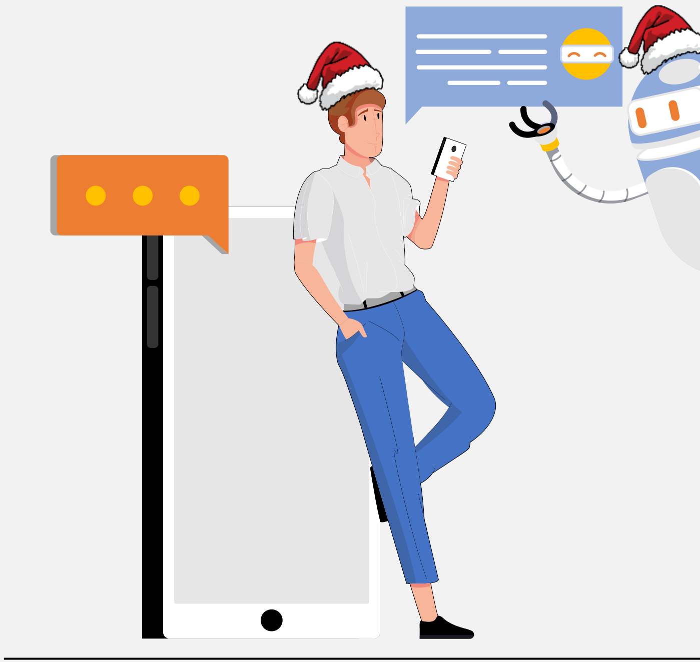
- Use better dataset
- Use better models or techniques to get embeddings of masks (e.g. masked pooling)
- Pre-train the adapter to align the segmentations to the text embedding of the content of the images.
- Augment the data with a GPT to have less repetitive query/answer structure

Thanks & Happy Holidays

References:

- Lai, Xin et al. "LISA: Reasoning Segmentation via Large Language Model."
- Xia, Zhuofan et al. "GSVA: Generalized Segmentation via Multimodal Large Language Models."
- Wei, Cong et al. "LaSagnA: Language-based Segmentation Assistant for Complex Queries."
- Wu, Tsung-Han et al. "See, Say, and Segment: Teaching LMMs to Overcome False Premises."
- Bao, Xiaoyi et al. "CoReS: Orchestrating the Dance of Reasoning and Segmentation."
- Wang, Junchi and Lei Ke. "LLM-Seg: Bridging Image Segmentation and Large Language Model Reasoning."
- Ma, Tianren et al. "ClawMachine: Fetching Visual Tokens as An Entity for Referring and Grounding."
- Zou, Xueyan et al. "Segment Everything Everywhere All at Once." ArXiv abs/2304.06718 (2023)

Github repo: https://github.com/DavidC001/LISA_ACV.git



Davide Cavicchini

Matteo Ranzetti