

MMR: A Large-scale Benchmark Dataset for Multi-target and Multi-granularity Reasoning Segmentation

Donggon Jang*, Yucheol Cho*, Suin Lee, Taehyeon Kim, and Dae-Shik Kim
Department of Electrical Engineering, KAIST

* Equal Contribution



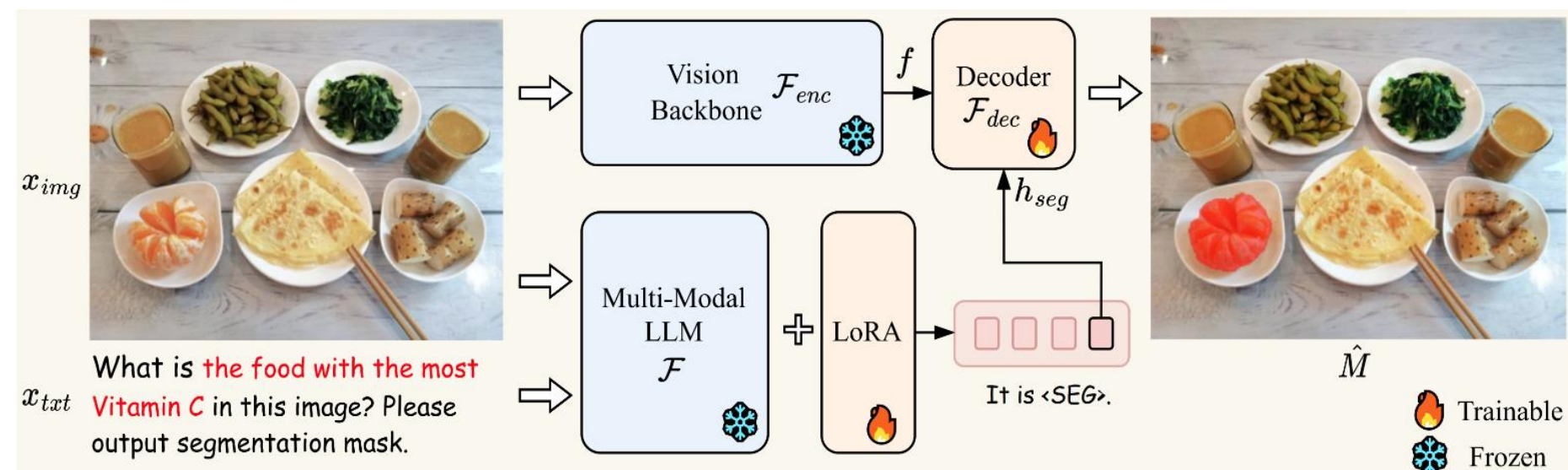
Poster session

Thu 24 Apr, 10 a.m.

iclr.cc/virtual/2025/poster/28436

❖ What is the Reasoning Segmentation?

- LISA^[1] first introduces reasoning segmentation task.
- Unlike previous tasks that rely on explicit text (e.g., “orange”), **reasoning segmentation handles implicit queries that require intricate reasoning or world knowledge** (e.g., “the food with most vitamin C”).

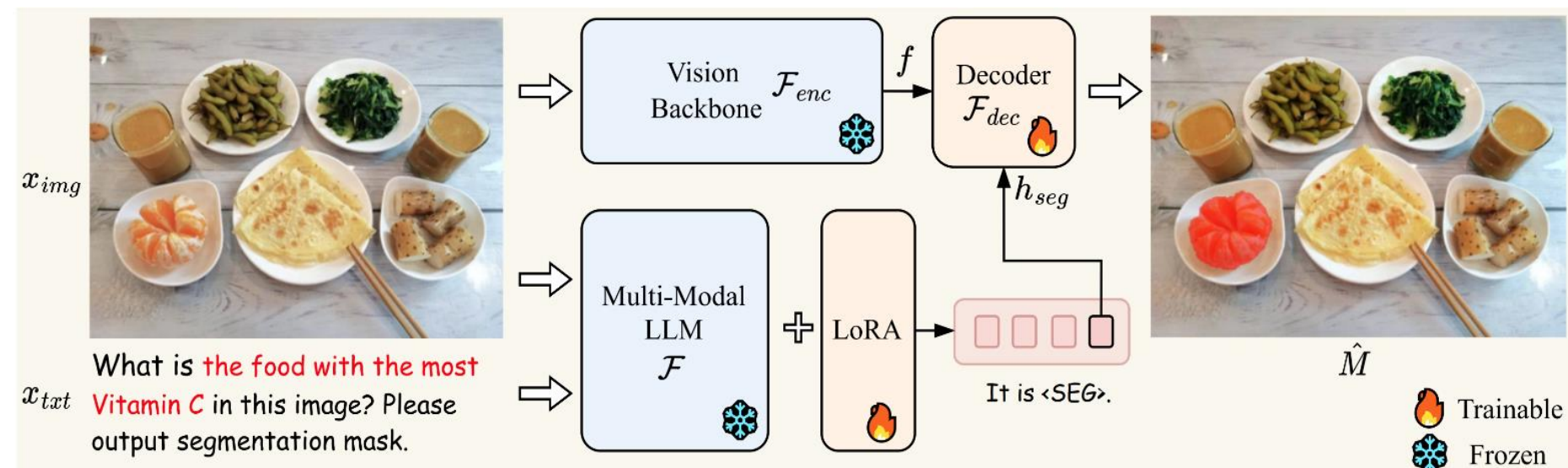


The framework of LISA^[1]

Background

❖ What is the Reasoning Segmentation?

- For reasoning segmentation task, LISA^[1] propose the **ReasonSeg** dataset which contains 1,218 implicit text question-answer pairs that involve complex reasoning for each image.



The framework of LISA^[1]



Question: "In the context of public transportation, which mode of transportation can carry many passengers and travel along designated tracks?"

Answer: It is [SEG]

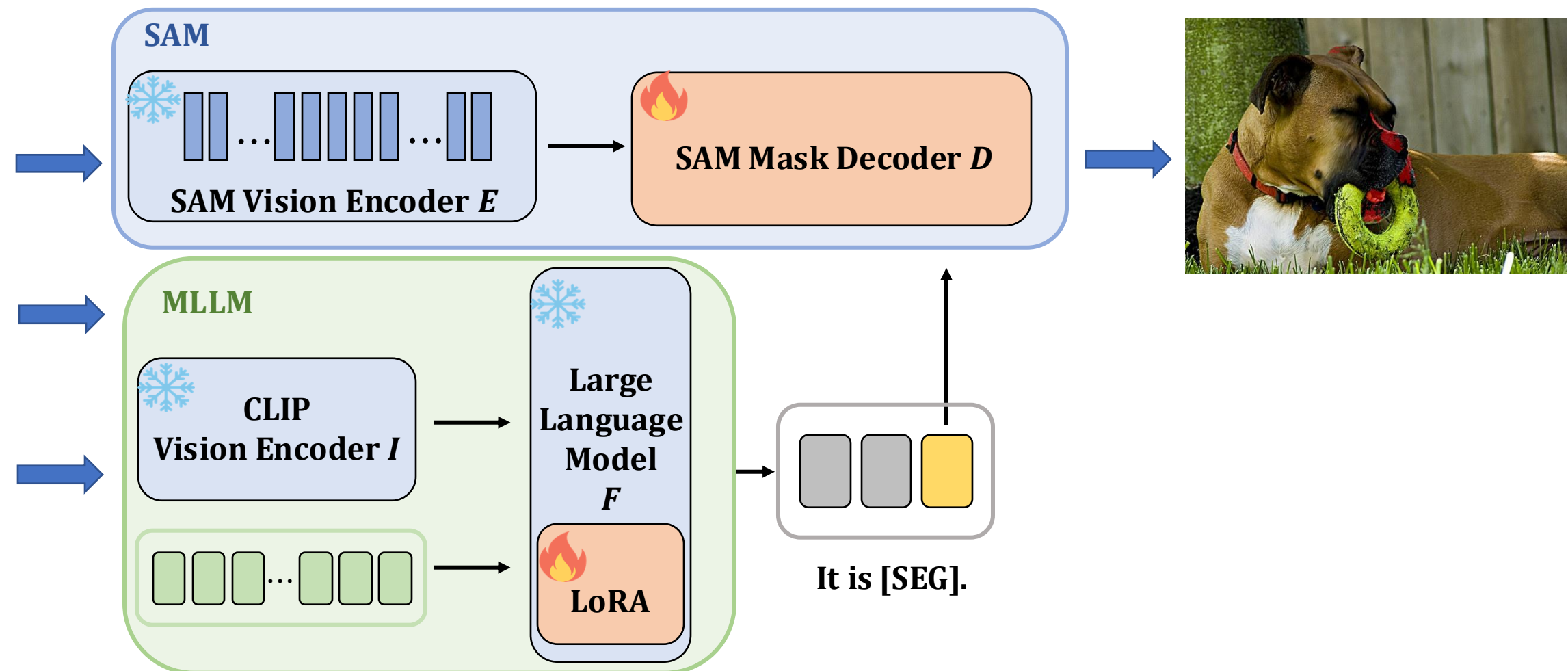
The example of ReasonSeg^[1]

Motivation

❖ Challenges



Q: In case the animal hears a sound and tries to identify its direction, which parts would most likely be actively moving?

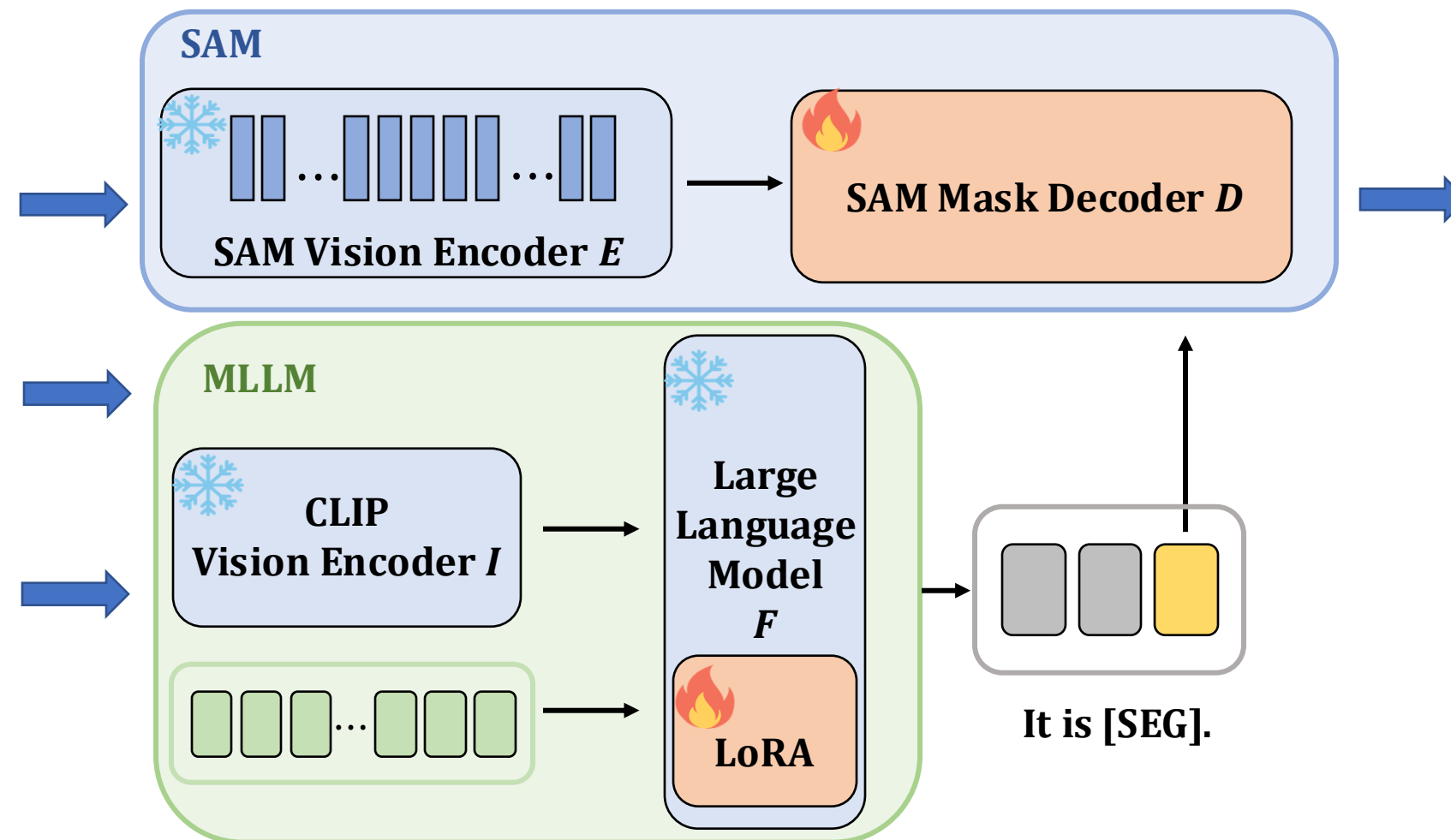


Motivation

❖ Challenges



Q: In case the animal hears a sound and tries to identify its direction, which parts would most likely be actively moving?



① Fail to provide multi-target masks
ex) dog: ears

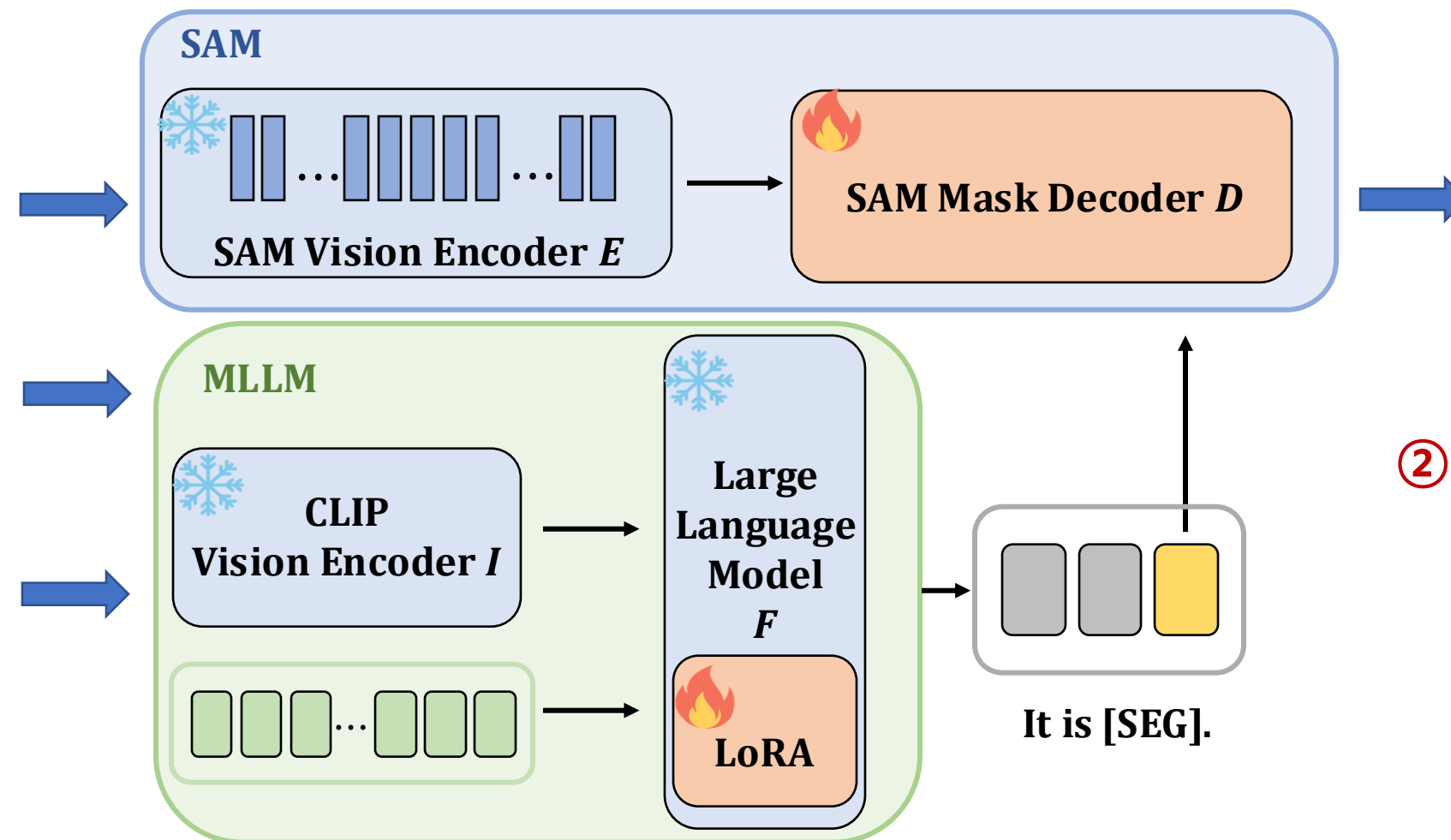


Motivation

❖ Challenges



Q: In case the animal hears a sound and tries to identify its direction, which parts would most likely be actively moving?

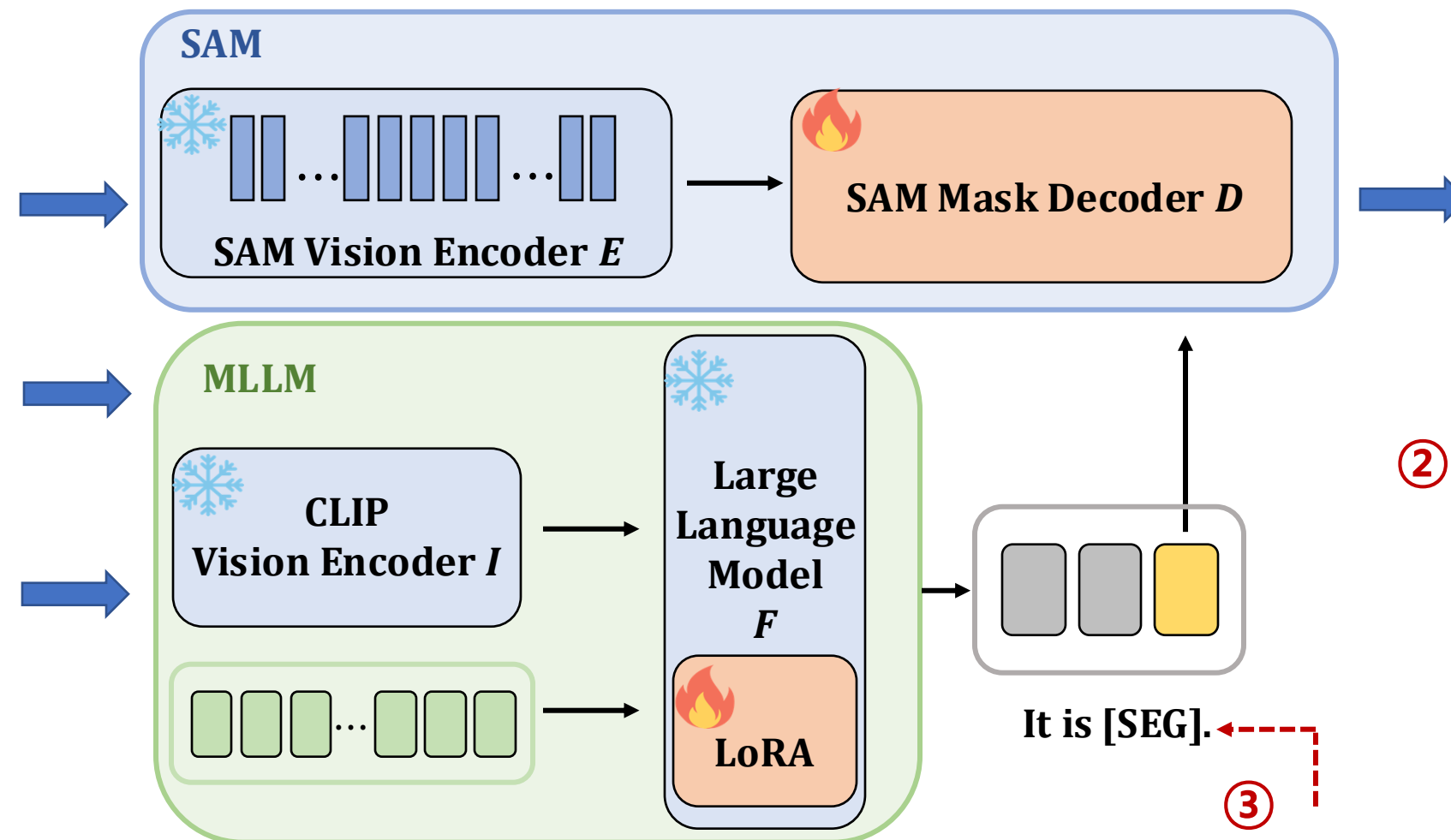


Motivation

❖ Challenges



Q: In case the animal hears a sound and tries to identify its direction, which parts would most likely be actively moving?



① Fail to provide multi-target masks
ex) dog: ears

② Inaccurate part-level masks
ex) dog: nose

③ Fail to provide text answers

Motivation

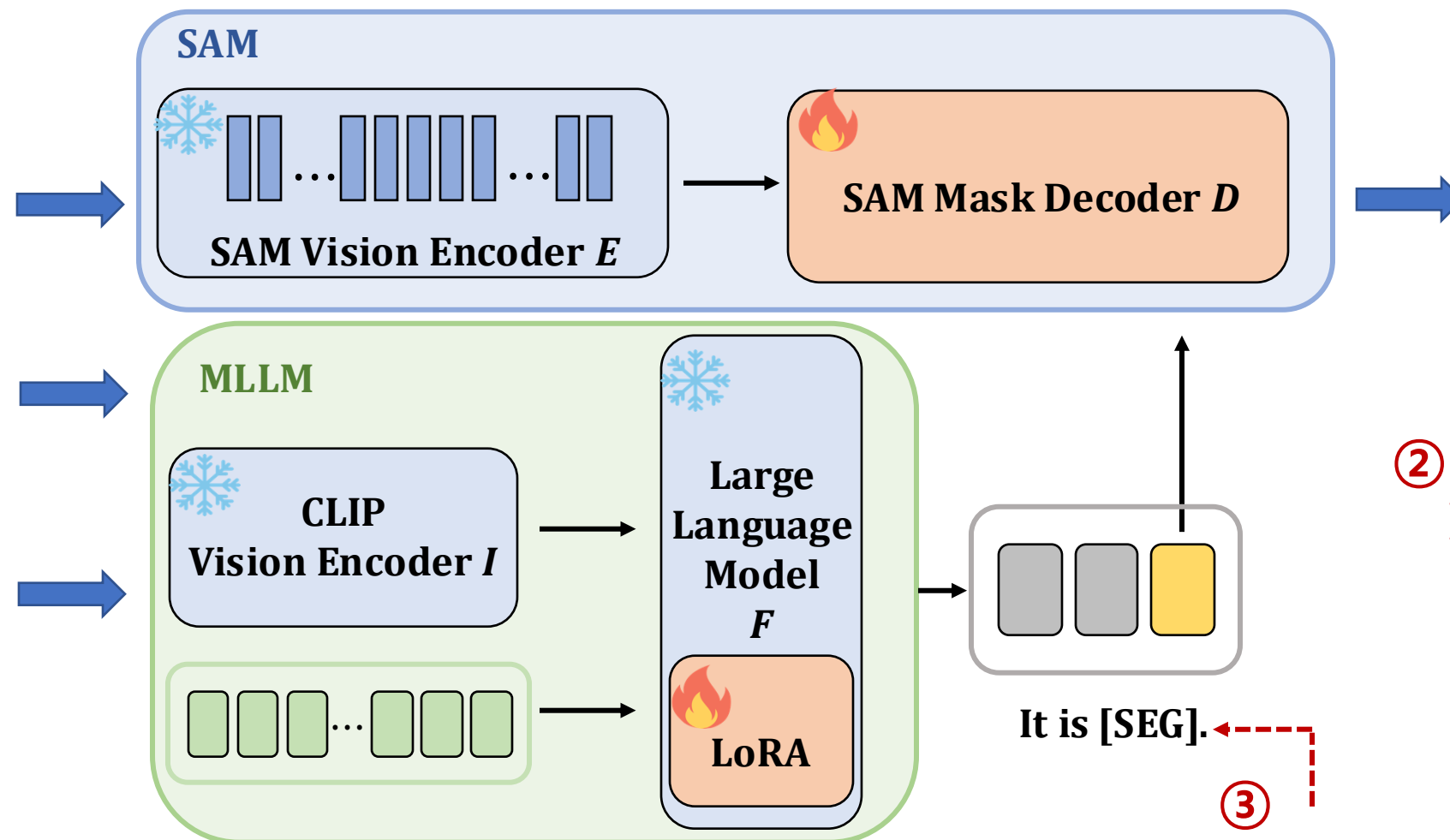
❖ Challenges

- **Absence of a dataset** considering the cases below:

- **Multi-target.**
- **Multi-granularity (object-level & part-level).**
- **Text answers.**



Q: In case the animal hears a sound and tries to identify its direction, which parts would most likely be actively moving?



① **Fail to provide multi-target masks**
ex) dog: ears

② **Inaccurate part-level masks**
ex) dog: nose

③ **Fail to provide text answers**

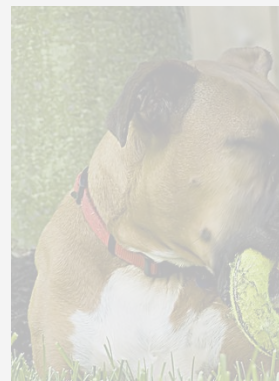
Motivation

❖ Challenges

- Absence of a dataset considering the cases below:

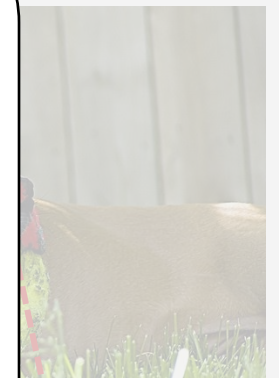
- Multi-target.
- Multi-granularity (object-level & part-level).
- Text answers

① Fail to provide multi-target masks
ex) dog: ears



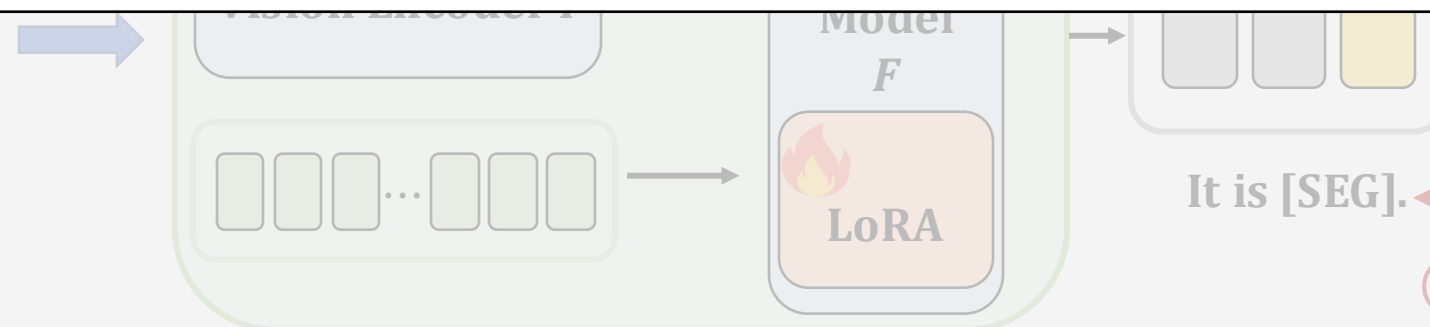
Goal:

We need to construct the new reasoning segmentation dataset which covers multi-target, object-level, and part-level reasoning!



part-level masks
Ex) dog: nose

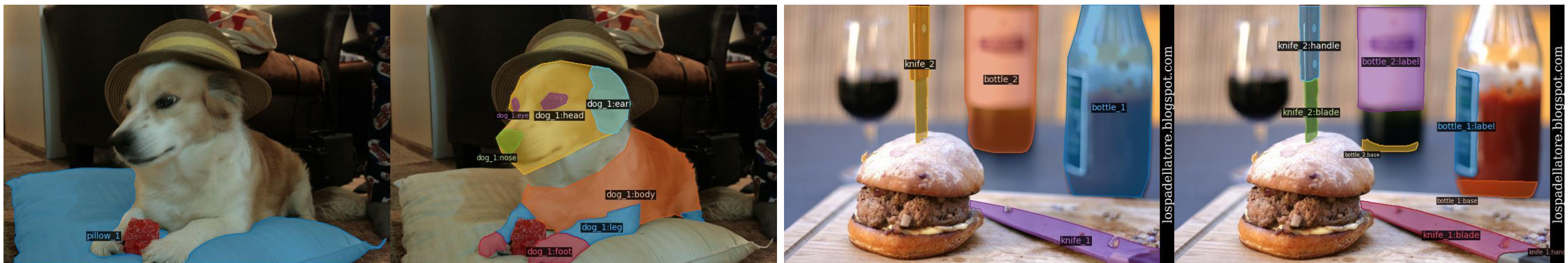
Q: In case the animal makes a sound and tries to identify its direction, which parts would most likely be actively moving?



③ Fail to provide text answers

❖ Generation pipeline

- We propose **Multi-target and Multi-granularity Reasoning segmentation (MMR) Dataset**.
- MMR dataset is based on PACO-LVIS dataset^[1].
 - PACO-LVIS includes **456 object-specific part classes** across **75 object categories**, offering **502K part-level masks** across **273K object-level masks**.
 - By utilizing these annotations, we can reduce annotation costs.
- To create intricate and implicit question-answer pairs, we **GPT API-assisted data generation scheme** similar to LLaVA^[2].



PACO-LVIS Examples

Method

❖ Generation pipeline

- To guide the GPT-4V API effectively, we carefully craft prompts.

System Message

① GPT Role

You are an AI visual assistant capable of analyzing a single image.

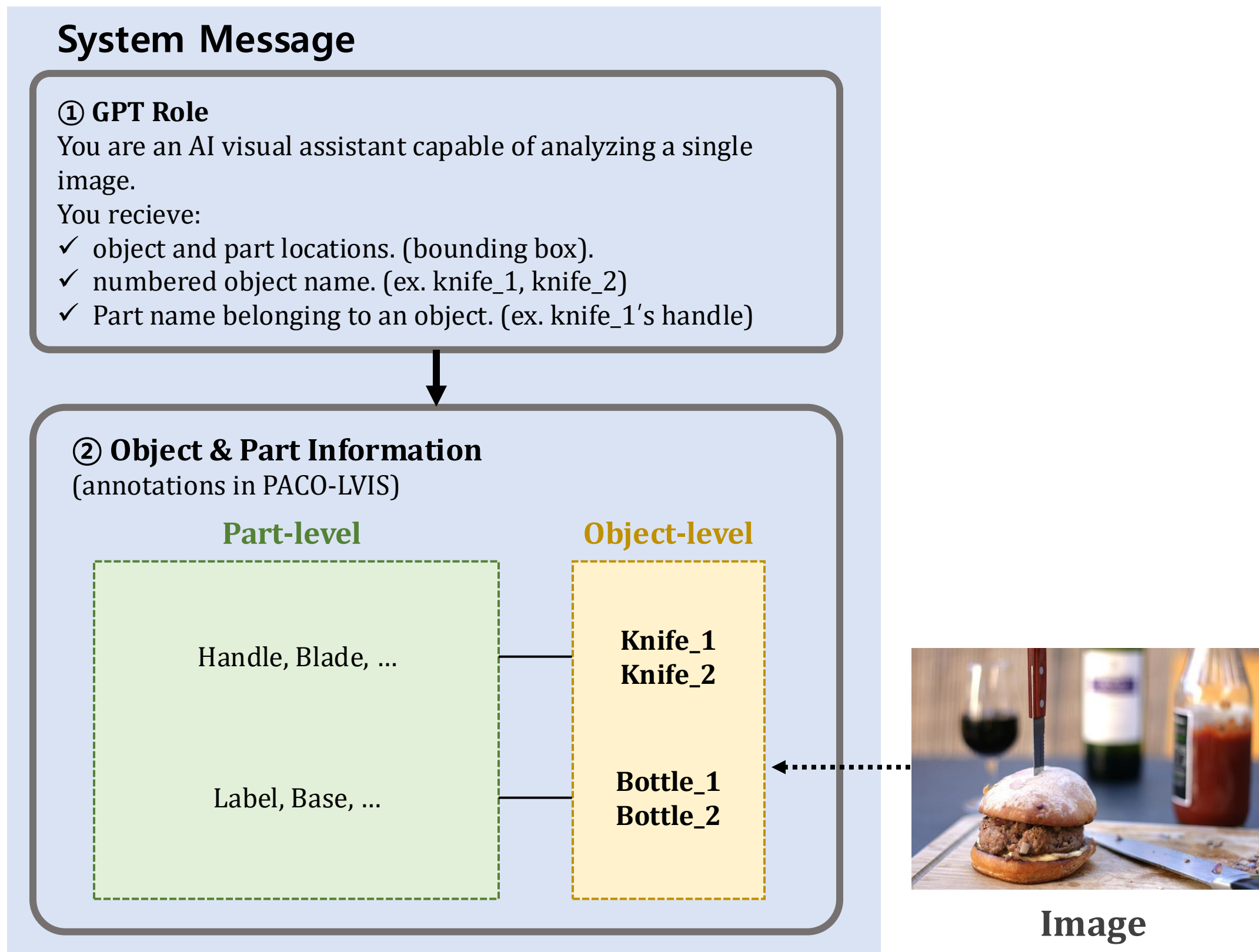
You receive:

- ✓ object and part locations. (bounding box).
- ✓ numbered object name. (ex. knife_1, knife_2)
- ✓ Part name belonging to an object. (ex. knife_1's handle)

Method

❖ Generation pipeline

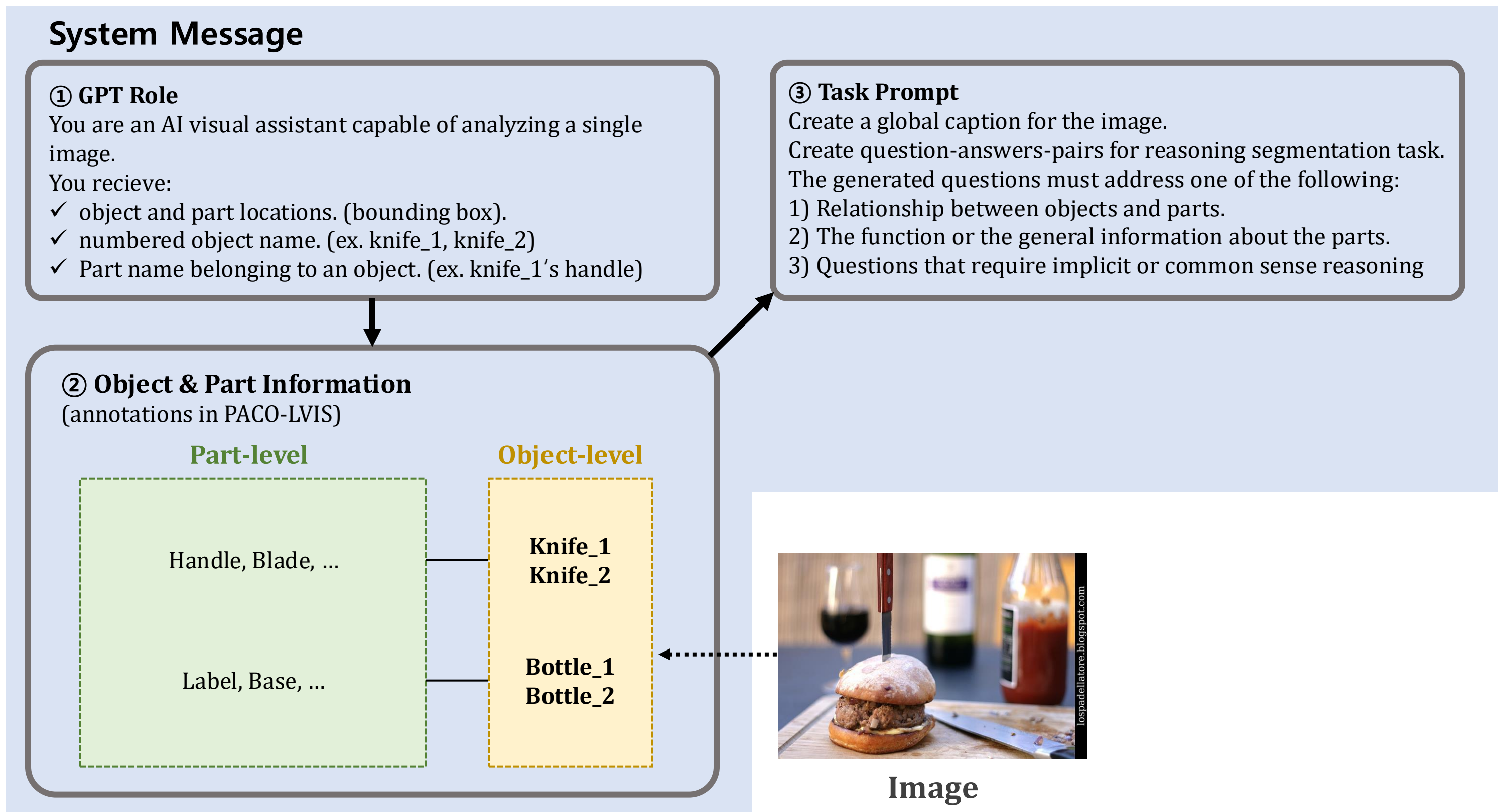
- To guide the GPT-4V API effectively, we carefully craft prompts.



Method

❖ Generation pipeline

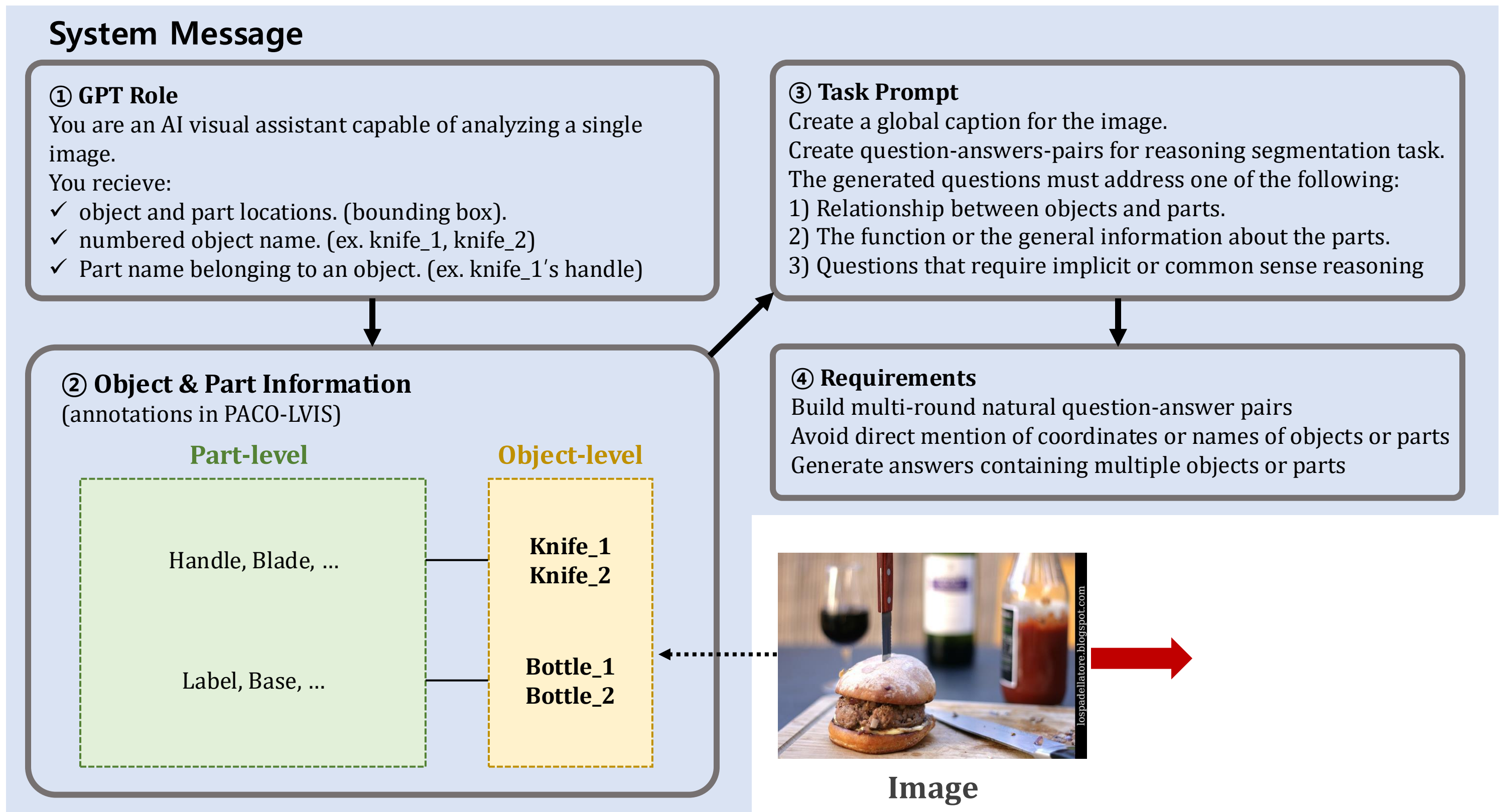
- To guide the GPT-4V API effectively, we carefully craft prompts.



Method

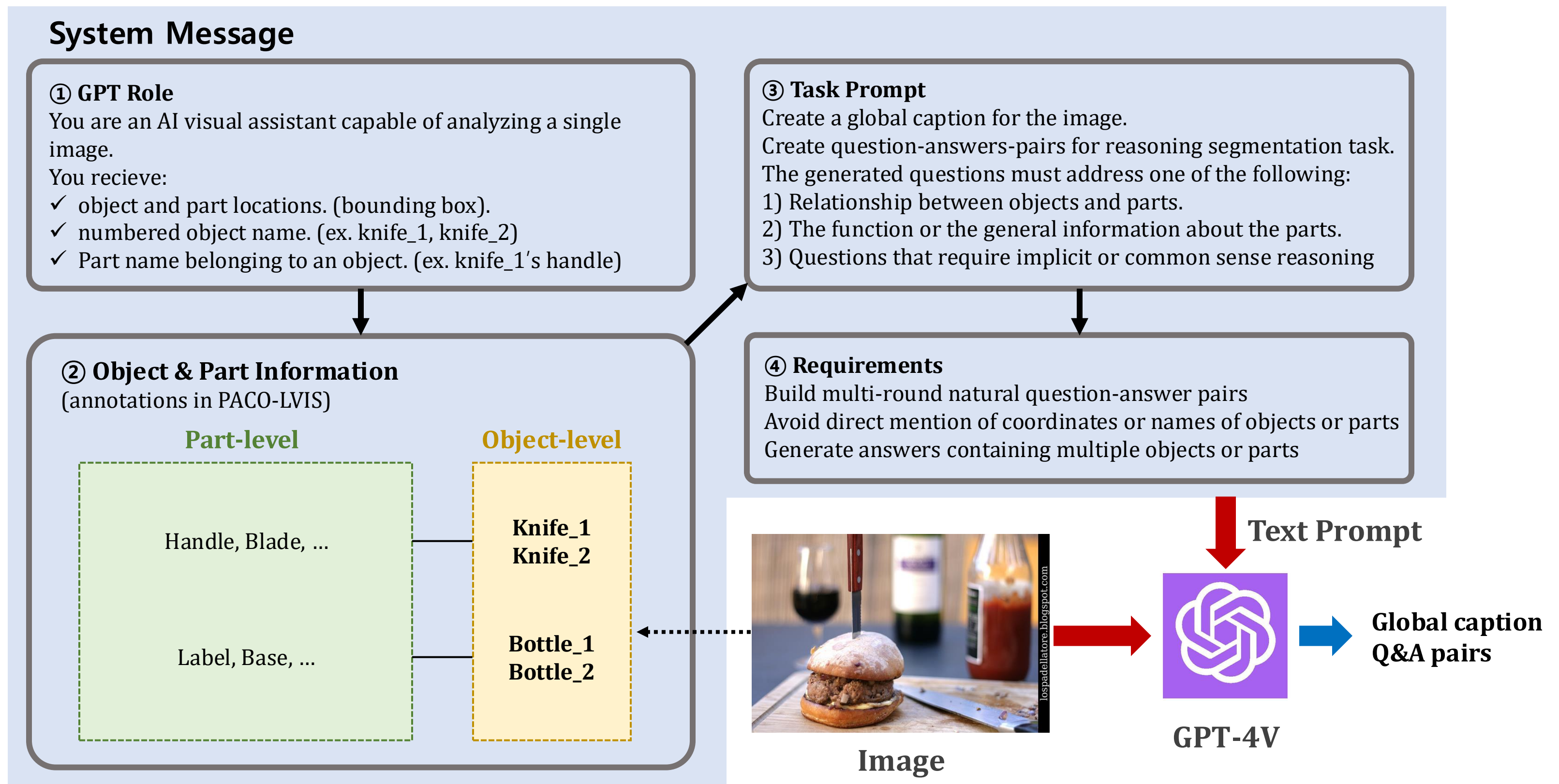
❖ Generation pipeline

- To guide the GPT-4V API effectively, we carefully craft prompts.

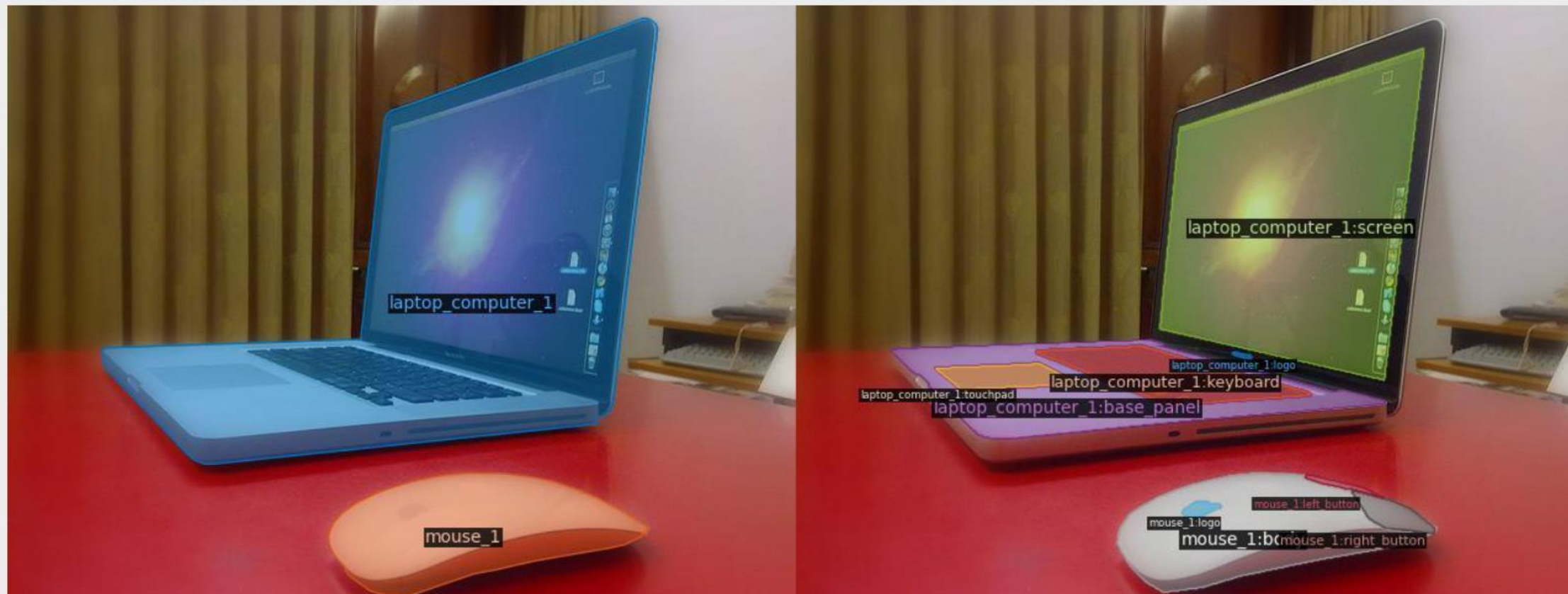


❖ Generation pipeline

- To guide the GPT-4V API effectively, we carefully craft prompts.



❖ An example of MMR dataset



Global Caption: A laptop is opened and set on a table next to a computer mouse, suggesting a typical workspace setup.

Question1: If one were to begin typing a document, which two areas of this device would they interact with first?

Answer1: They would primarily interact with the `laptop_computer_1's keyboard [195, 276, 418, 325]` to type and `laptop_computer_1's touchpad [113, 290, 231, 312]` to navigate within the document.

Question2: Where can one find the manufacturer's branding on the devices pictured here?

Answer2: The manufacturer's branding can be found on the `laptop_computer_1's logo [354, 281, 370, 288]` and on the `mouse_(computer_equipment)_1's logo [314, 403, 345, 416]`.

Question3: To move the cursor on the screen without touching the laptop, which part of the computer equipment would one use?

Answer3: One would use the `mouse_(computer_equipment)_1's body [260, 379, 516, 477]` along with either the `mouse_(computer_equipment)_1's left_button [413, 380, 480, 401]` or `mouse_(computer_equipment)_1's right_button [451, 393, 519, 429]` to click and interact with the cursor on the screen.

Question4: After finishing work and deciding to pack up, which two parts of the laptop would come into contact?

Answer4: When closing the laptop, `laptop_computer_1's screen [295, 34, 510, 305]` would come into contact with `laptop_computer_1's base_panel [77, 271, 479, 352]`.

Method

❖ MMR dataset statistics

- **MMR includes 194K reasoning questions-answer pairs** with corresponding images and masks.
 - 75 object categories and 445 part categories from PACO-LVIS.
- Fig. 3 (a) and (c) demonstrate that the question-answer pairs are grounded in common and general objects and their associated parts.

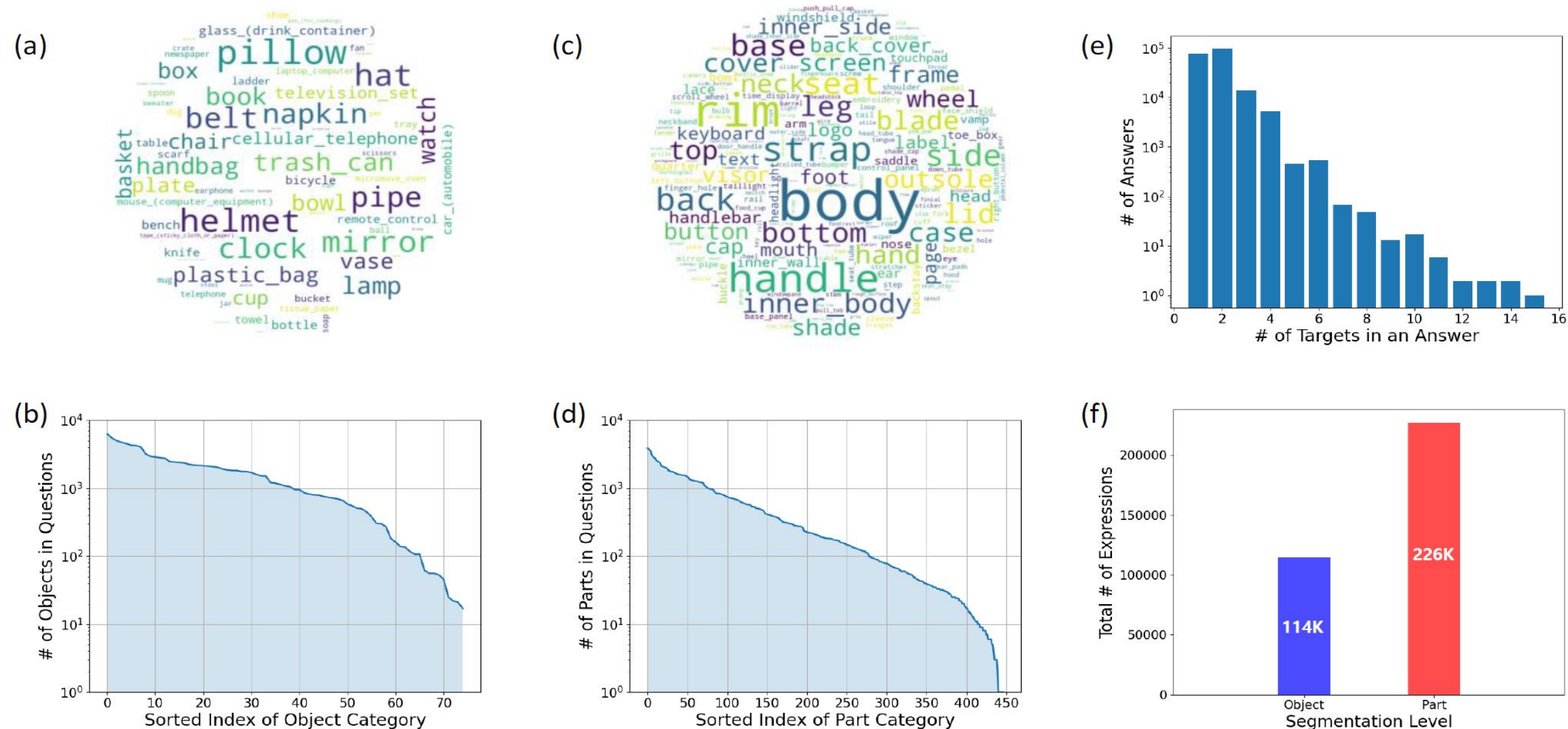


Figure. Statistics of the proposed MMR dataset. **(a)** the word cloud for the object categories, **(b)** the number of objects per each object category in questions (log scale), **(c)** the word cloud for the part categories, **(d)** the number of parts per each part category in questions (log scale), **(e)** the distribution of target count in answers, and **(f)** the total number of expressions of objects and parts.

Method

❖ MMR dataset statistics

- MMR dataset encompasses a wide range of categories, ensuring that the question-answer pairs are **not biased toward specific categories** and **exhibit a high level of diversity**.

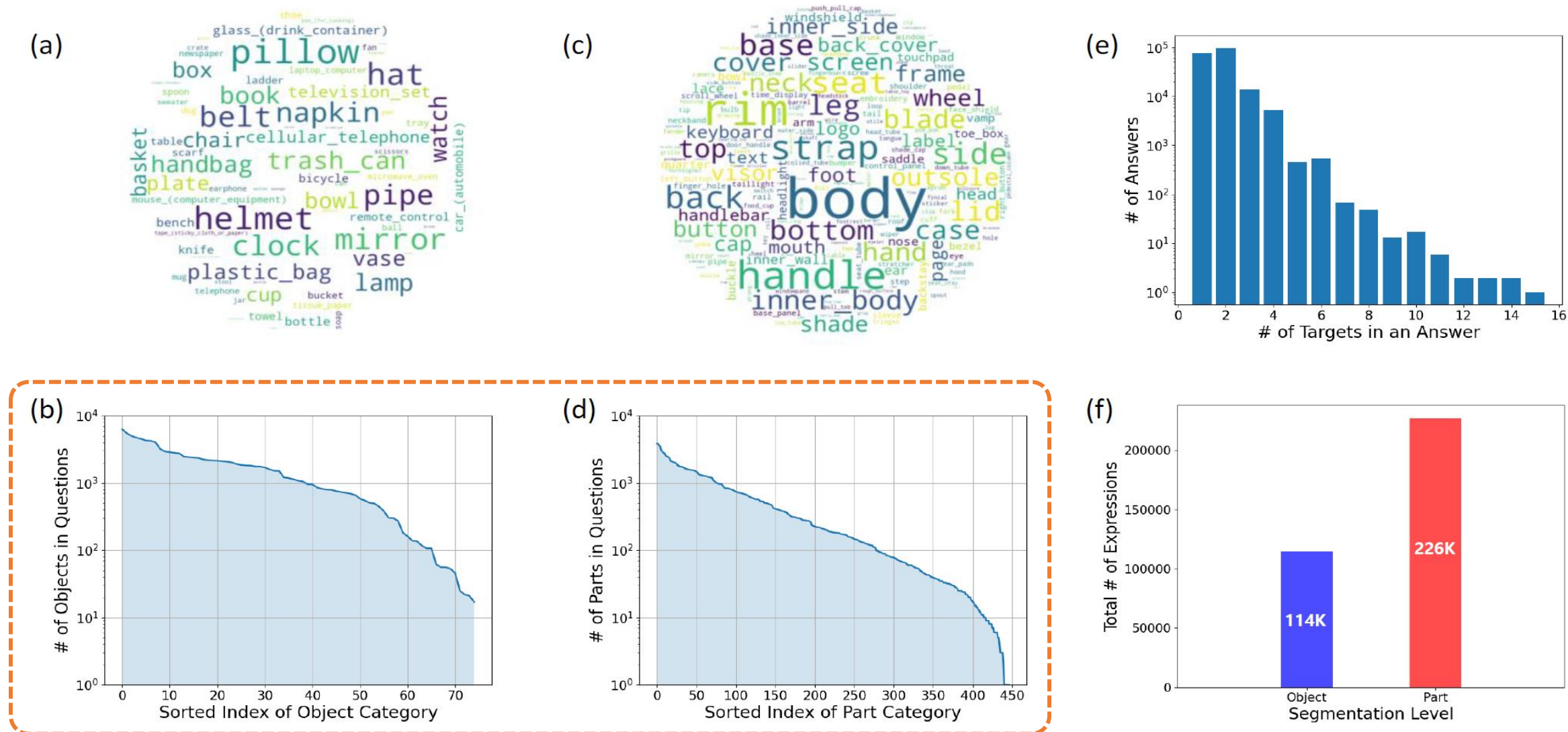
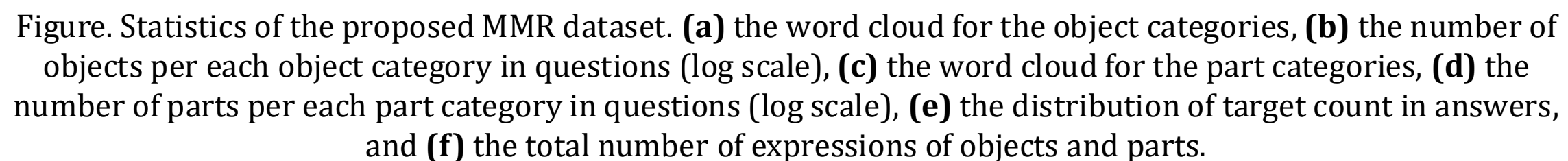



Figure. Statistics of the proposed MMR dataset. **(a)** the word cloud for the object categories, **(b)** the number of objects per each object category in questions (log scale), **(c)** the word cloud for the part categories, **(d)** the number of parts per each part category in questions (log scale), **(e)** the distribution of target count in answers, and **(f)** the total number of expressions of objects and parts.

- On average, there are **1.8 targets per answer**, with **the maximum number of targets in a single pair being 16**.
- This demonstrates that **MMR dataset can consider multiple targets and cover diverse target reasoning**.



❖ Comparison with existing reasoning segmentation datasets

	ReasonSeg ^[1]	MUSE ^[2]	MMR (Ours)
Object-level			

❖ Comparison with existing reasoning segmentation datasets

	ReasonSeg ^[1]	MUSE ^[2]	MMR (Ours)
Object-level	✓	✓	✓
Part-level	✓	✗	✓

❖ Comparison with existing reasoning segmentation datasets

	ReasonSeg ^[1]	MUSE ^[2]	MMR (Ours)
Object-level	✓	✓	✓
Part-level	✓	✗	✓
Multi-target	✗	✓	✓

[1] Lai, Xin, et al. "Lisa: Reasoning segmentation via large language model." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[2] Ren, Zhongwei, et al. "Pixellm: Pixel reasoning with large multimodal model." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024.

❖ Comparison with existing reasoning segmentation datasets

	ReasonSeg ^[1]	MUSE ^[2]	MMR (Ours)
Object-level	✓	✓	✓
Part-level	✓	✗	✓
Multi-target	✗	✓	✓
# of Q&A pairs	1.2K	214K	194K

[1] Lai, Xin, et al. "Lisa: Reasoning segmentation via large language model." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[2] Ren, Zhongwei, et al. "Pixellm: Pixel reasoning with large multimodal model." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024.

❖ Comparison with existing reasoning segmentation datasets

	ReasonSeg ^[1]	MUSE ^[2]	MMR (Ours)
Object-level	✓	✓	✓
Part-level	✓	✗	✓
Multi-target	✗	✓	✓
# of Q&A pairs	1.2K	214K	194K
GPT API	GPT-3.5	GPT-4V	GPT-4V

[1] Lai, Xin, et al. "Lisa: Reasoning segmentation via large language model." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[2] Ren, Zhongwei, et al. "Pixellm: Pixel reasoning with large multimodal model." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024.

❖ Comparison with existing reasoning segmentation datasets

- MMR first includes large-scale, multi-target, and multi-granularity question-answer pairs, strengthening real-world applicability.

	ReasonSeg ^[1]	MUSE ^[2]	MMR (Ours)
Object-level	✓	✓	✓
Part-level	✓	✗	✓
Multi-target	✗	✓	✓
# of Q&A pairs	1.2K	214K	194K
GPT API	GPT-3.5	GPT-4V	GPT-4V

[1] Lai, Xin, et al. "Lisa: Reasoning segmentation via large language model." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[2] Ren, Zhongwei, et al. "Pixellm: Pixel reasoning with large multimodal model." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

❖ M²SA Framework

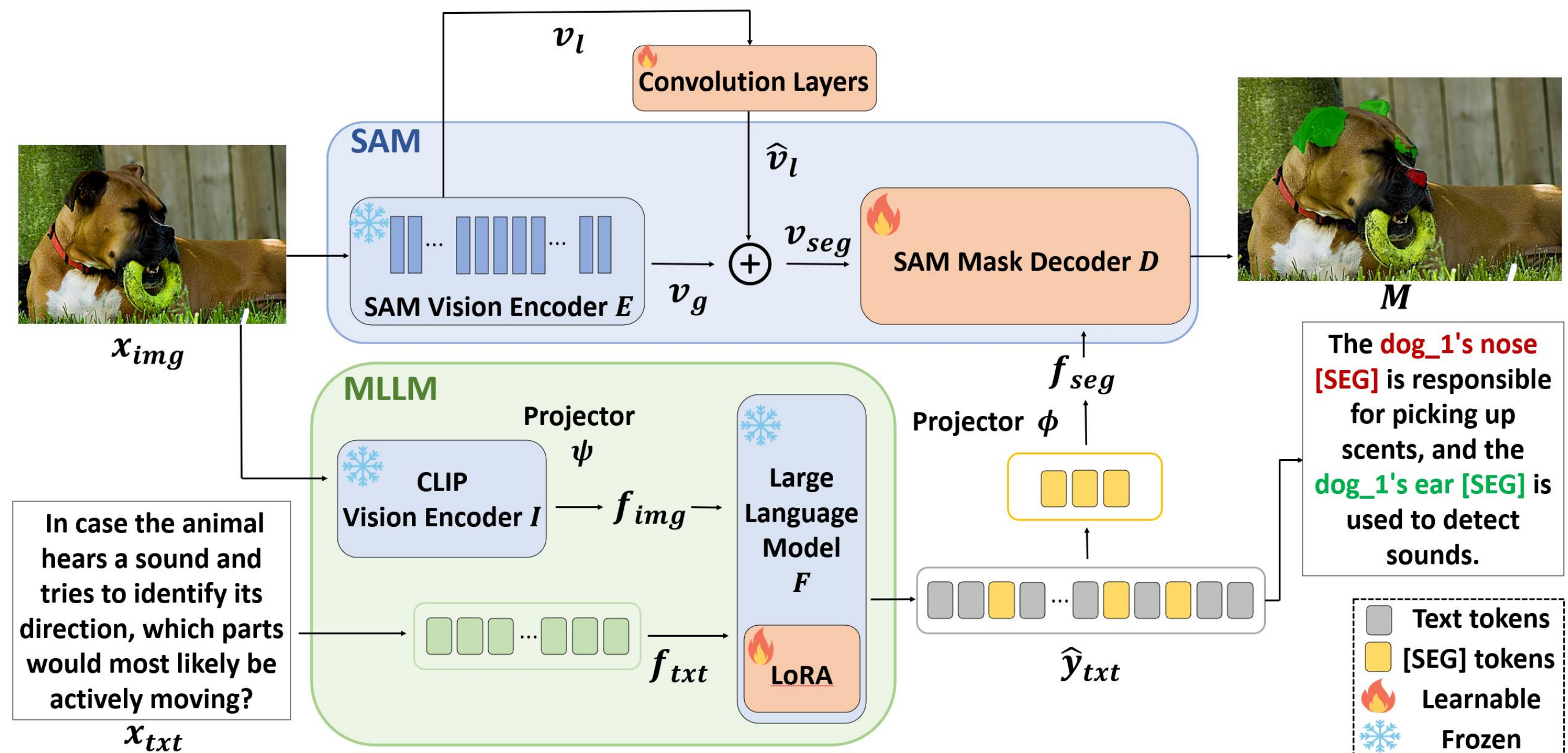


Figure. The overview of M²SA Framework

❖ M²SA Framework

1. Early Local Feature Fusion

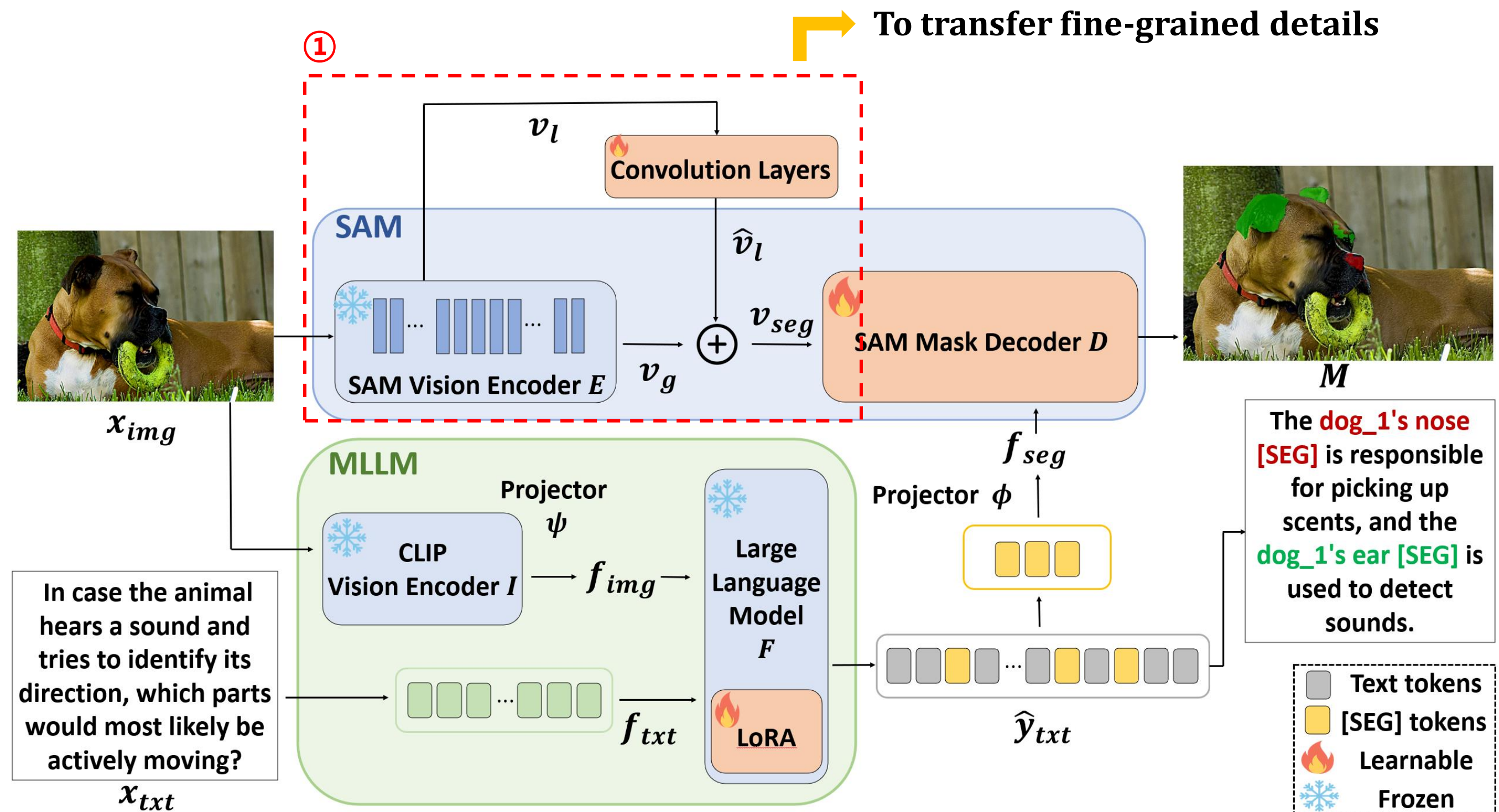


Figure. The overview of M²SA Framework

❖ M²SA Framework

1. Early Local Feature Fusion
2. Multiple [SEG] tokens

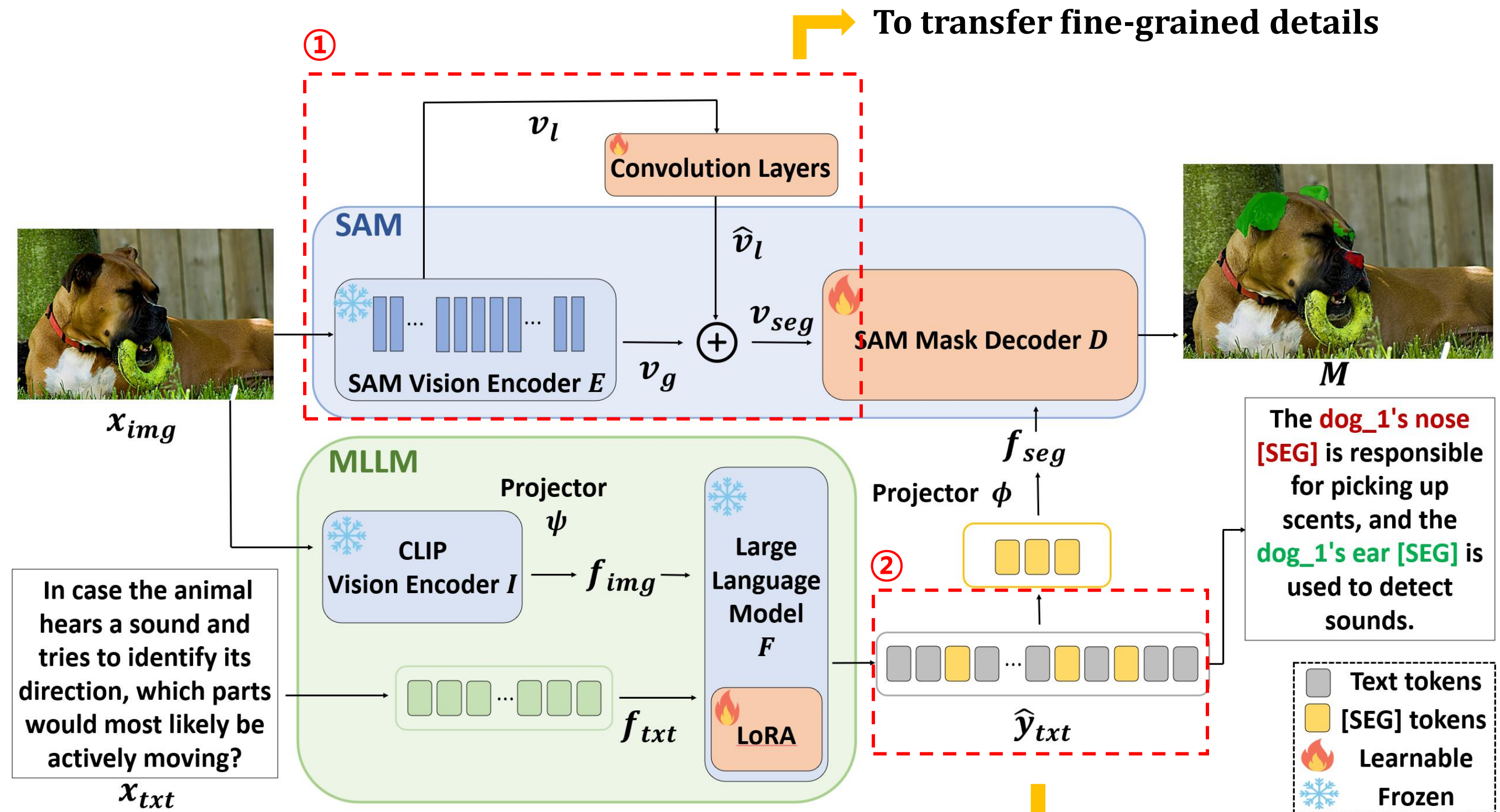


Figure. The overview of M²SA Framework

To predict multiple target masks

Experimental Results

❖ Comparison on MMR dataset

• Key Results:

- The previous reasoning segmentation models perform poorly on the proposed MMR dataset, particularly struggling with the part-only set due to the lack of detailed part-level understanding.
- M²SA shows highly competitive performance, showcasing its strength in comprehensive reasoning segmentation.

Methods	val				test			
	Obj & Part		Obj		Part		Obj & Part	
	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU
LISA-7B	13.8	18.3	23.5	25.1	6.6	7.9	14.5	17.9
LISA-7B _{tr}	19.4	31.6	34.7	41.8	8.0	13.1	19.5	27.1
GSVA-7B	14.6	25.1	26.4	34.3	6.0	11.6	15.5	24.8
GSVA-7B _{tr}	19.8	38.9	30.2	41.1	8.0	18.6	21.2	34.5
GLaMM	12.6	19.2	23.7	31.9	3.9	6.4	13.3	18.7
GLaMM _{tr}	26.9	47.1	40.3	54.2	12.1	25.5	30.3	45.0
M²SA-7B	27.8	48.6	41.0	55.6	13.5	27.0	30.9	46.8
LISA-Llama2-13B	15.4	20.0	26.1	27.9	7.4	8.4	16.1	19.8
LISA-Llama2-13B _{tr}	22.3	33.4	40.2	45.2	10.7	16.4	23.0	29.2
M²SA-Llama2-13B	28.4	49.1	42.3	57.6	13.6	27.2	31.6	47.6

Table. Reasoning segmentation results on MMR validation and test sets. The gIoU and cIoU metrics are reported for the comparison. *Obj & Part*, *Obj*, and *Part* denote multi-granularity, object-only, and part-only evaluation settings.

Conclusion

❖ Contribution

- We construct **the MMR dataset**, which includes **194K complex and implicit question pairs for multi-target and multi-granularity reasoning segmentation**.
 - This dataset facilitates advanced reasoning segmentation tasks in open-world scenarios.
- We propose **M²SA** for multi-target and multi-granularity reasoning segmentation. It incorporates an **early local feature fusion** and **multiple [SEG] tokens** to improve fine-grained visual understanding and segment multiple targets.

❖ For datasets and codes, please visit our github:



Thank you!