Appendix

A Implementation Details

For training datasets, we follow LISA to utilize a combination of datasets for semantic segmentation, COCO-Stuff, PACO-LVIS, PASCAL-Part, and Mapillary Vistas), referring segmentation(refCLEF, refCOCO, refCOCO+, and refCOCOg), and the ReasonSeg training set for fine-tuning. We abandon using the visual question answering (LLaVA-Instruct-150k) dataset employed by the baseline method LISA. The rationale lies in CoReS's control over the chain-of-reasoning process through the fixed templates of LLM output formats. The textual outputs from the VQA dataset, which may conflict with the logic of template-based outputs and fine-grained tasks, could adversely affect the training process.

For the test dataset, we conduct experiments on the ReasonSeg. It comprises image-question pairs with reasoning difficulty and the ground truth segmentation masks. Its training and validation set includes 239 and 200 image-instruction pairs, respectively.

As for the LoRA tuning of the MLLM, the LoRA rank is set to 8 and the LoRA alpha and dropout are set to 16 and 0.05, respectively. We apply deepspeed as the code engine, with 8 NVIDIA A100 for training. The optimizer for CoReS is AdamW and its learning rate is set to 0.0003, with a 100-step warm-up-decay.

B Ablation Experiments

As for different design alternatives of CoReS, we conduct experiments in Tab.S1. **Text-Only CoT**: We first compare with a CoT-version of LISA which performs explicit text-form reasoning before segmentation, which is trained with the additional text generated by GPT as supervision. This Text-Only CoT decreases the performance of LISA. One possible reason is that textual outputs weaken the info in [SEG], consequently impacting the subsequent segmentation process. **Supervised C1**, we choose two ways to provide supervision for C1, ①'*C1-GT*' is using the final mask for C1 loss calculation at a low weight scale. The decrease in performance can be attributed to the undermining of the CoT formation, regardless of weight differences. ②'*Pseudo*' refers to generating larger masks as C1 gt. It can be observed that leading MLLM to generate without constraints is better than rigid constraints. **Query-Specific Textual Prompts**: We adopt query-specific textual prompts (instead of irrelevant prompts) as inputs. Although it may not be practical in applications due to time and cost considerations, the improvement, as a rough upper bound, indirectly validates the effectiveness of in-context input.

We conduct experiments on the hyperparameters in Tab.S2. Placing excessive emphasis on the reasoning chain loss leads to a decrease in model performance. This is because it merely serves as an implicit constraint on the output of intermediate text, while the ultimate segmentation chain loss deserves more

Table S1: Implementation strategies of CoT. 'LISACoT' refers to the text-only version of CoT, and 'relevant' refers to using the trained CoReS-7B with query-related prompts generated by GPT. 'C1-GT' and 'pseudo' are the above two C1 supervisions.

Metric	CoReS	LISA	LISACoT	relevant	C1-GT	pseudo
gIoU	59.4	52.9	44.7	60.1	55.9	58.1
cIoU	62.1	54.0	45.9	62.8	55.8	62.1

attention. Furthermore, increasing the weight of the DICE loss also has a positive effect on the results, which aligns with empirical intuition.

Table S2: Hyperparameters analysis of loss weights.

Hyper-	$\lambda_{COS}/\lambda_{COR}$			$\lambda_{CE}/\lambda_{DICE}$		
param	2:1	1:1	1:2	4:1	3:2	2:3
gIoU	**59.4**	58.8	55.5	57.6	59.0	**59.4**
cIoU	**62.1**	61.1	58.3	62.2	**66.2**	62.1

We use gIoU as our primary metric, rather than the cIoU with a bias towards large-area targets. To clarify, we divide target objects based on their sizes, as shown in Tab.B. cIoU is dominated by the performance on large objects, which dilutes the segmentation advantage of CoReS on more challenging small targets.

Table S3: Ablation of different metrics. '$cIoU_s$', '$cIoU_m$' and '$cIoU_l$' refer to cIoU of objects with small, medium, and large sizes, respectively.

Method	LLaVA-v1.5-13B				
	gIoU	cIoU	$cIoU_s$	$cIoU_m$	$cIoU_l$
LISA	65.0	67.8	22.4	53.6	78.7
CoReS	**68.1**	**68.2**	**35.7**	**58.5**	**80.5**

Table S4: Performance comparison on part segmentation datasets. 'PIN', 'PP', and 'RP' are PartImageNet, PascalPart, and ReasonPart, respectively.

Method	PIN	PP	RP	
	mIoU	mIoU	gIoU	cIoU
LISA	33.3	14.8	14.1	18.9
CoReS	**39.0**	**19.6**	**20.9**	**33.1**

C Experiments on other benchmarks

Additionally, we evaluate the ability of CoReS on general fine-grained segmentation datasets. The mean-intersection-over-union (mIoU) serves as the evaluation metric here. On PartImageNet, the zero-shot performance of CoReS shows a 5.7% improvement over LISA. Similarly, on the validation set of PascalPart, an approximate 5% enhancement is observed. These results demonstrate that the multimodal reasoning chain structure of CoReS not only elevates the understanding of reasoning tasks but also augments the capability to comprehend details in general dense visual tasks.

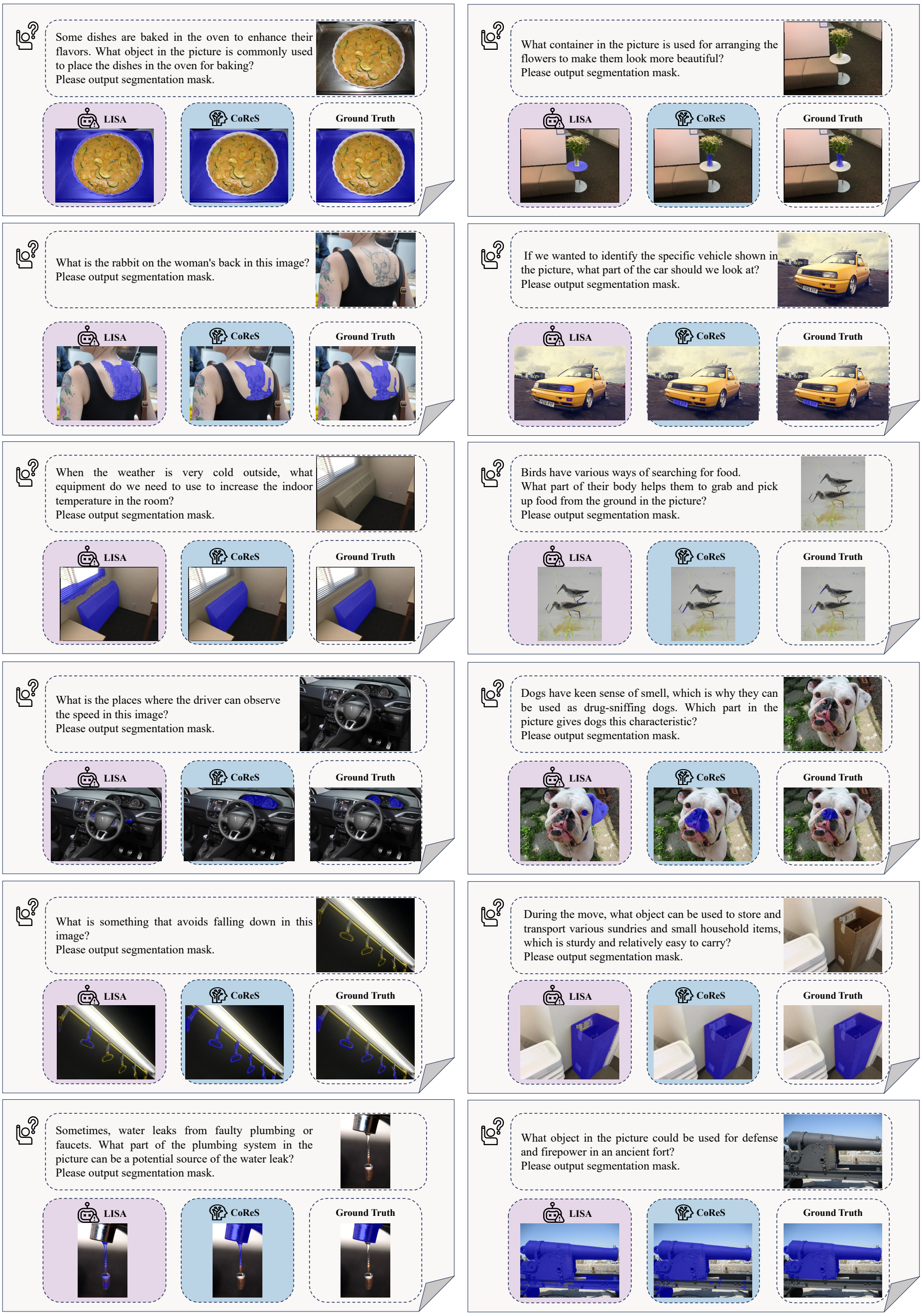
In addition, we construct two reasoning-based part segmentation benchmarks based on the aforementioned two component-level datasets. We employ in-context learning with chatGPT to generate questions about category names, yielding questions such as "In the oceanic scene, which functional feature aids fish in steering and maintaining stability in the currents?" ReasonPart is established on 2,957 validation set images from PartImageNet and 4,465 images from PascalPart. As Tab.S4 indicates, on the challenging benchmark ReasonPart, which features complex queries and fine-grained segmentation parts, CoReS outperforms LISA by 6.8% in terms of gIoU and approximately 15% in cIoU, substantiating the efficacy of its dual-chain structure.

D Analysis of C1 masks

Obtaining mask annotations for intermediate logical layers is impractical due to the logic subjectivity, so MLLM with world knowledge is used for a unified unsupervised representation. The scene-level info-injection of C1 is firstly based on MLLM's in-context learning ability, where in-context inputs are provided for logical prompts. Secondly, the MLLM's semantically coherent and reasonable output helps the formation of the hierarchy. Supervised templates, like 'It appears on [LOC],' constrain the output semantics after 'appears on' naturally carry the scene-level or item location meaning. Thirdly, since a rough mask is required as SAM input, the gradient backpropagation also pushes [LOC] toward the scene-level info.

E Quantitative Results

Qualitative results also prove the effectiveness of CoReS in Fig.S2 and Fig.S1. When facing complex reasoning problems, LISA exhibits errors in grounding the exact instance referred by the reasoning query, while CoReS achieves the correct answers. Whether it's objects that are difficult to segment, such as a fork inserted into rice, or challenging reasoning tasks like determining the source of power for carts in an image, the chain-like hierarchy of CoReS results in correct and fine-grained segmentation.



Some dishes are baked in the oven to enhance their flavors. What object in the picture is commonly used to place the dishes in the oven for baking?
Please output segmentation mask.

What container in the picture is used for arranging the flowers to make them look more beautiful?
Please output segmentation mask.

What is the rabbit on the woman's back in this image?
Please output segmentation mask.

If we wanted to identify the specific vehicle shown in the picture, what part of the car should we look at?
Please output segmentation mask.

When the weather is very cold outside, what equipment do we need to use to increase the indoor temperature in the room?
Please output segmentation mask.

Birds have various ways of searching for food. What part of their body helps them to grab and pick up food from the ground in the picture?
Please output segmentation mask.

What is the places where the driver can observe the speed in this image?
Please output segmentation mask.

Dogs have keen sense of smell, which is why they can be used as drug-sniffing dogs. Which part in the picture gives dogs this characteristic?
Please output segmentation mask.

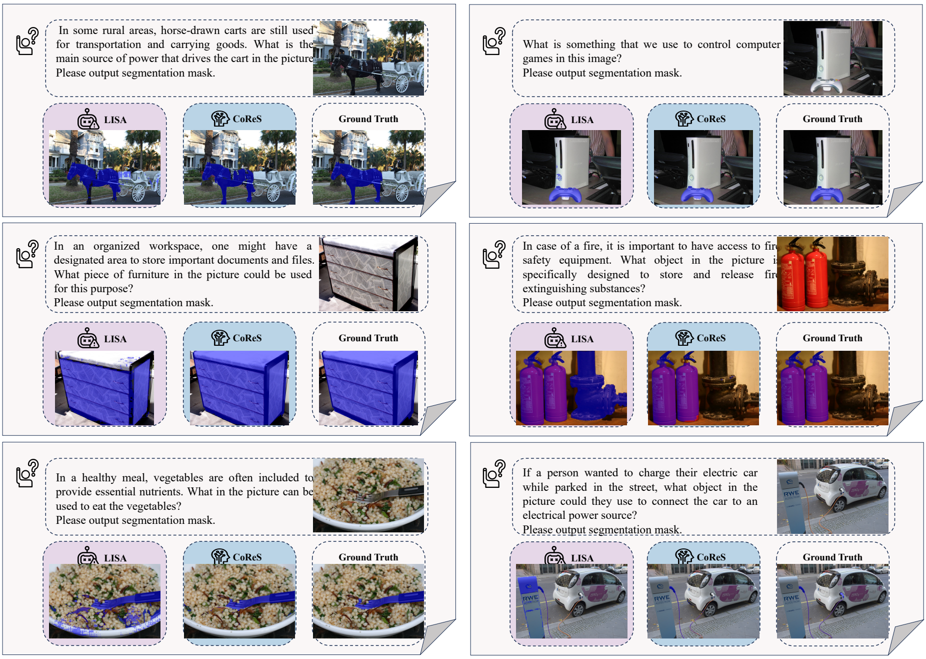
What is something that avoids falling down in this image?
Please output segmentation mask.

During the move, what object can be used to store and transport various sundries and small household items, which is sturdy and relatively easy to carry?
Please output segmentation mask.

Sometimes, water leaks from faulty plumbing or faucets. What part of the plumbing system in the picture can be a potential source of the water leak?
Please output segmentation mask.

What object in the picture could be used for defense and firepower in an ancient fort?
Please output segmentation mask.

Fig. S1: Visual comparison of CoReS and LISA.



In some rural areas, horse-drawn carts are still used for transportation and carrying goods. What is the main source of power that drives the cart in the picture
Please output segmentation mask.

What is something that we use to control computer games in this image?
Please output segmentation mask.

In an organized workspace, one might have a designated area to store important documents and files. What piece of furniture in the picture could be used for this purpose?
Please output segmentation mask.

In case of a fire, it is important to have access to fir safety equipment. What object in the picture i specifically designed to store and release fir extinguishing substances?
Please output segmentation mask.

In a healthy meal, vegetables are often included to provide essential nutrients. What in the picture can be used to eat the vegetables?
Please output segmentation mask.

If a person wanted to charge their electric car while parked in the street, what object in the picture could they use to connect the car to an electrical power source?
Please output segmentation mask.

Fig. S2: Visual comparison of CoReS and LISA.