

Reasoning to Attend: Try to Understand How <SEG> Token Works

Supplementary Material

We provide supplementary material related to the main paper, arranged as follows:

1. Additional implementation details (Appendix A)
2. Additional Analysis (Appendix B)
3. Additional ablation study (Appendix C)
4. Additional qualitative results (Appendix D)
5. Discussion (Appendix E)

A. Additional Implementation Details

A.1. Grid Search for Optimal Threshold

Given that the threshold for the foreground mask has a significant impact on the IoU, to eliminate the bias introduced by manually setting the threshold (*e.g.*, 0.5), we perform a grid search over the similarity map for each image with a step size of 0.01 to identify the optimal foreground mask. For each threshold t , we convert the similarity map into a binary mask by applying

$$\hat{M}(x, y) = \begin{cases} 1 & \text{if } \mathcal{S}(x, y) \geq t \\ 0 & \text{if } \mathcal{S}(x, y) < t \end{cases}, \quad (14)$$

where $\mathcal{S}(x, y)$ is the similarity score for each pixel at position (x, y) , $\hat{M}(x, y)$ is the binary mask at that pixel. We calculate cIoU for all threshold values in the grid, and choose the threshold t' that maximizes the cIoU for the image

$$t' = \arg \max_t (\text{cIoU}(t)). \quad (15)$$

Once the optimal threshold is selected for each image, we use it to generate the final binary masks for evaluation, which ensures that the comparison is fair and threshold-invariant.

A.2. Model Architecture and Training

As for reasoning segmentation, we trained two models: READ-7B and READ-13B. For READ-7B, we initialize the parameters using the released SESAME model [42] to accelerate training, with the training dataset allocated in a 10:1:1:1:1:10 ratio. We employ LoRA [9] for efficient fine-tuning, using $\text{lor}_a_r = 8$, and conduct end-to-end joint training. For READ-13B, we train it from scratch, using LLaVA 1.5-13B as the base model. Initially, we train it on the full dataset in a 10:10:2:3:1:1 ratio for about 8 epochs, and then fine-tune it with a ratio of 3:10:2:3:1:10, using a learning rate of 0.0001 and $\text{lor}_a_r = 64$. As for referring segmentation, we maintain the same settings as those used for READ-7B in reasoning segmentation. All our code will be publicly available at <https://github.com/rui-qian/READ>.

B. Additional Analysis

(1) Fig. 4 shows qualitative analysis of the <SEG> token on the ReasonSeg *val* set. Points derived from (a) serve as

prompts with original SAM in (c). Similarity between the <SEG> token and image token embeddings stemming from the last hidden layer is computed by Eq.(5), w.r.t. LLaVA encoder in (a) and SAM decoder in (b). The consistency in (a), (b) indicates that the <SEG> token in LMMs learns semantics similar to direct mentions in text, as observed in CLIP [31]. Note that 1st column in (b) shows failure cases, indicating the existence of misalignment between the LLaVA encoder in (a) and SAM decoder in (b). Such observation sheds light on the interpretability of semantic alignment issues, where the LLaVA encoder generates accurate textual responses even in scenarios where the SAM decoder fails at segmentation, when eliciting LISA [16] for reasoning explanations. In future work, we aim to further investigate the underlying connections behind this phenomenon. (2) Fig. 6 shows a qualitative analysis of $\mathcal{P}_{\text{prompt}}$ on the ReasonSeg *val* set. We first select several points with the highest similarity scores as positives (red in (b)) and an equal number of points with the lowest similarity scores as negatives (blue in (b)). These points are then directly used as prompts instead of the <SEG> token, and are input into the original SAM model to generate the segmentation mask. Columns in (b) demonstrate that only relying on the selected similarity points as prompt can still generate a segmentation mask potentially.

C. Additional Ablation Study

Effect of points ratios. To explore how the ratios of positive, negative, and neutral points impact the performance of READ, we vary the positive and negative thresholds (t_{pos} and t_{neg}) as well as the number of points $|\mathcal{P}|$. As the positive sample ratio (t_{pos}) increases, model performance improves, particularly when fewer points are used ($|\mathcal{P}|=10$). Also, increasing the number of points generally enhances performance, with the most significant improvements observed at $|\mathcal{P}|=60$, regardless of the t_{pos} setting.

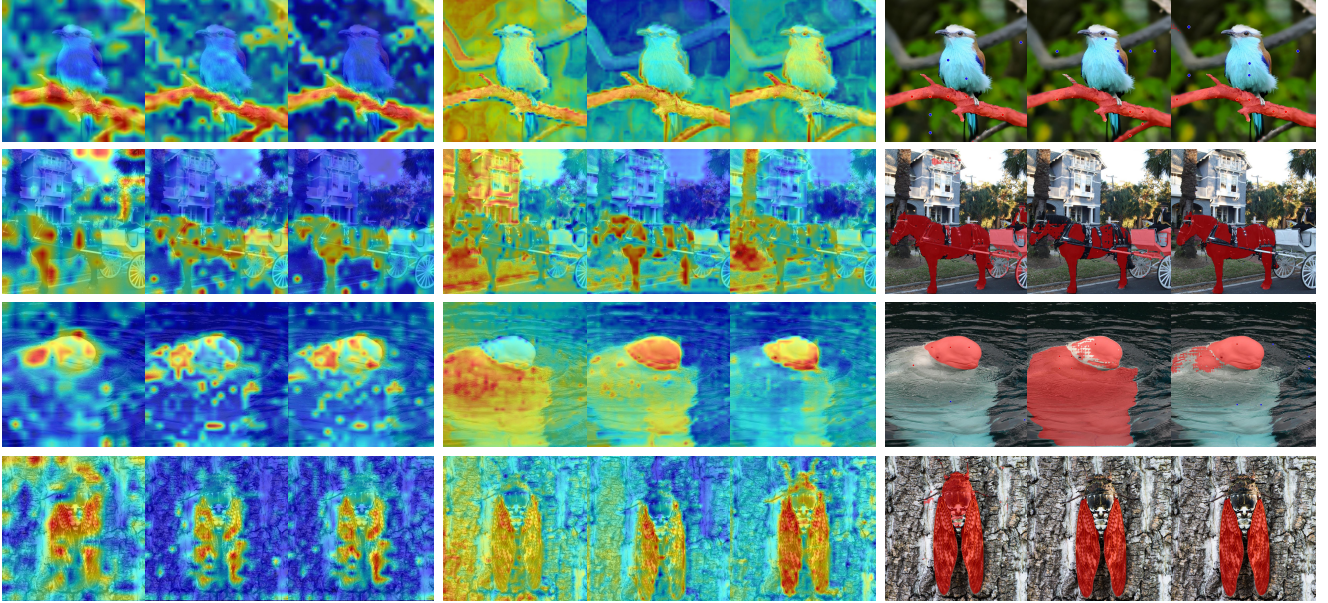
Table 8. Ablation study on points ratios.

t_{pos}	t_{neg}	$ \mathcal{P} =10$		$ \mathcal{P} =30$		$ \mathcal{P} =60$	
		gIoU	cIoU	gIoU	cIoU	gIoU	cIoU
0.8	0.2	58.94	65.16	59.75	67.62	59.71	68.17
0.7	0.3	58.48	64.00	58.82	65.32	59.20	67.70
0.6	0.4	58.59	64.27	58.66	65.00	58.93	66.69

D. Additional Qualitative Results

Fig. 7 shows qualitative results on the FP-RefCOCO(+g) *val* set. Also, READ retains the conversational ability of LLMs while performing segmentation tasks and can refuse to output a mask when the queried object doesn't exist.

Fig. 8 shows the qualitative results of READ on the ReasonSeg *val* set. LISA and SESAME exhibit various defects to some extent when handling the displayed cases, whereas our approach delivers more desirable segmentation results.



(a) <SEG> with LLaVA

(b) <SEG> with SAM

(c) Points as prompt with SAM

Figure 4. Qualitative analysis of the <SEG> token on the ReasonSeg val set. The 1st, 2nd, and 3rd columns of (a), (b), and (c) are LISA, SESAME, and READ (Ours) for comparisons, respectively. Points derived from (a) serve as prompts with original SAM in (c).

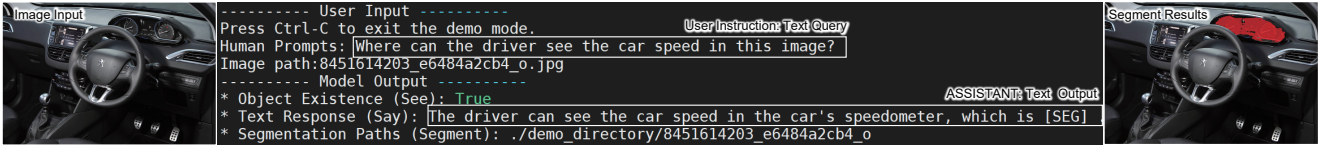


Figure 5. Showcase of complex reasoning and world knowledge.

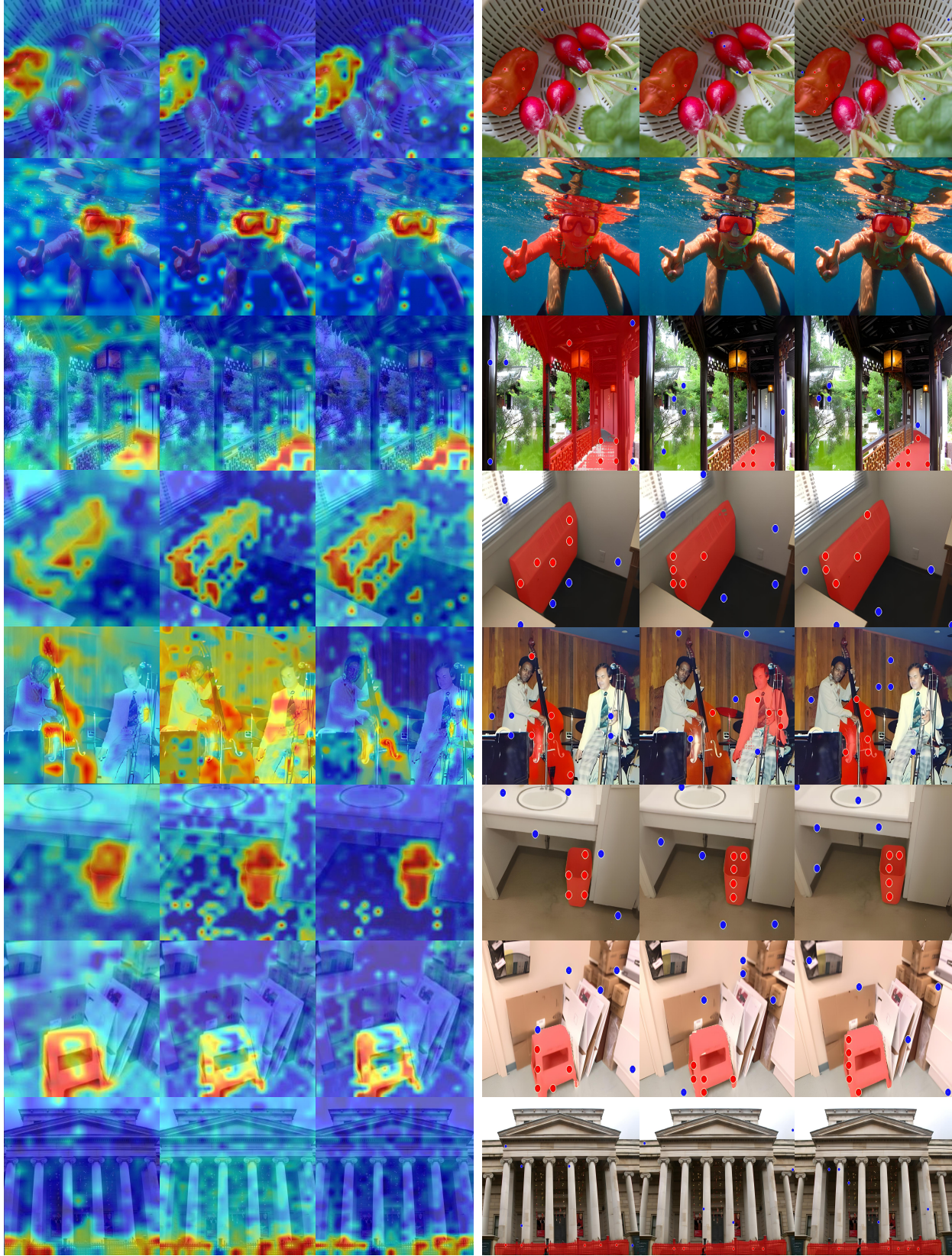
E. Discussion

Applicability. To showcase the broader applicability of our approach, we discuss how READ can be integrated with other methods. For LLM-based referring segmentation, such as LISA [16], GSVA [44], and GlaMM [33], our SasP module can be seamlessly incorporated with negligible effort, as they share the same <SEG> token pipeline as READ (ours). For non-LLM-based referring segmentation, such as MMCA [47], we compute the similarity between the output state of the <SEG>-like token and the image tokens derived from the last hidden layer in transformers to obtain a similarity map. We then select highly activated points for sparse embedding representations or use these points to interpolate features from a CNN-based (ResNet) feature map, similar to the lightweight RoI pooling operation in object detection tasks. The resulting embeddings can then be employed for downstream vision tasks. Beyond segmentation, as long as a vision task involves generating an attention map, our Discrete-to-Continuous (DtoC) strategy (Sec. 4.1) can be applied to edit the attention map.

Necessity. This raises two pivotal issues for consideration. First, is the <SEG> token (or a <SEG>-like placeholder) truly necessary? Moreover, what advantages does

the <SEG> token offer (why <SEG> token)? For the former, if the <SEG> token merely serves as a connector role for downstream tasks, then it is not necessary. For tasks that only involve segmenting positive samples where the object to be segmented is expected to exist (as in LISA), one could alternatively use the embeddings derived from the LLMs’ output text to tap into the LLMs’ capabilities. However, if the <SEG> token functions as a decision indicator of whether segmentation should be performed, then its inclusion becomes necessary. For instance, when it comes to false premises where the target objects might not exist, it is crucial to rely on the LLMs’ prediction (specifically, whether the output contains the <SEG> token) to determine if segmentation should take place.

For the latter, the <SEG> token infuses LLMs’ world knowledge into downstream tasks, compared to non-LLM-based methods such as MMCA [47] and M-DGT [5]. As illustrated in Fig. 5, solving the text query “Where can the driver see the car speed?” requires the model to possess world knowledge, since the query itself does not explicitly contain semantics that point to the answer (“speedometer”). In contrast, MMCA and M-DGT use BERT and ResNet as backbones, regardless of how effective their feature embeddings are, they inherently lack additional world knowledge.



(a) <SEG> with LLaVA

(b) Points as prompt with SAM

Figure 6. Qualitative analysis of $\mathcal{P}_{\text{prompt}}$ (points as prompt) on the ReasonSeg *val* set. The 1st, 2nd, and 3rd columns of (a), (b) are LISA, SESAME, and READ (Ours) for comparisons, respectively. Points derived from (a) serve as prompts with original SAM in (b).



Figure 7. Visualization on the FP-RefCOCO(+g) val set.














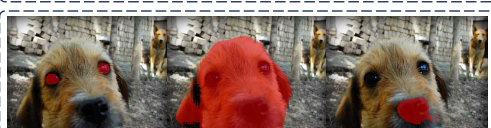







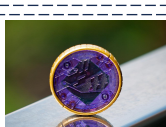
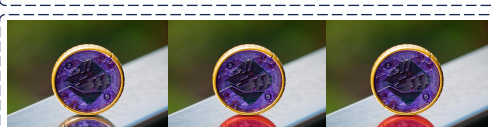
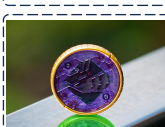












<p>Fishing is a popular activity for relaxation and leisure. What tool is the man in the picture using to catch fish?</p>					
<p>Please check this image for the stronger mario. If it's present, supply the segmentation map.</p>					
<p>Dogs use their mouths to perform various tasks, including eating and vocalizing. What part of the dog's body is primarily responsible for these actions?</p>					
<p>Dogs often like to find a comfortable place to rest. What object in the picture can offer a soft and comfortable surface for the dog to lie on?</p>					
<p>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?</p>					
<p>In some regions, people raise certain animals for their milk, meat, and skin. What animal in the picture could be domesticated for such purposes?</p>					
<p>When we need to access or store things above our reach, what would be helpful to stand on?</p>					
<p>Examine for the reflection of the object in this image.</p>					
<p>What container in the picture is most likely to be used next for pouring hot water to make tea?</p>					
<p>What object in the picture could be used for defense and firepower in an ancient fort?</p>					
<p>In the image, is the ship that is most likely to carry a fleet commander evident?</p>					
<p>Can you confirm the existence of something that the persons use to cross the water in this image?</p>					
Query	Image	LISA	SESAME	READ(Ours)	Ground Truth

Figure 8. Visual comparison among READ (ours) and prior works on the ReasonSeg *val* set.