InstructSeg: Unifying Instructed Visual Segmentation with Multi-modal Large Language Models

Supplementary Material

A. Additional Implementation Details

We utilize Phi-2 [14] with 2.7B parameters as our Large Language Model, SigLIP [48] as our CLIP Encoder, Swin-B [27] as our Visual Encoder, and pre-trained Maks2Former [5] as our Segmentation Decoder. We keep both CLIP Encoder and Visual Encoder frozen while applying LoRA [13] with the rank of 8 to finetune the Large Language Model. In contrast, the OVP, VMTF, and Segmentation Decoder components are fully finetuned. We use AdamW optimizer with the learning rate and weight decay set to 0.00004 and 0, respectively. In addition, we adopt Cosine Decay for the learning rate schedule, where the warmup steps are set to 1680.

B. Segmentation Decoder Structure

We illustrate the architecture of the segmentation decoder module in Fig. 5. Consistent with previous methods [5, 51], our approach integrates both a pixel decoder and a transformer decoder to extract pixel-level visual information and instance-level object information. Distinctively, we compute the similarity between mask embeddings and multi-grained text embeddings to derive mask scores, which are then utilized for the selection of mask proposals.

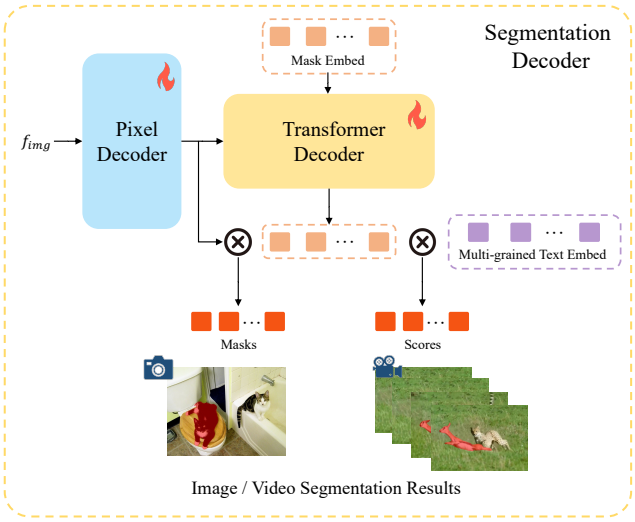

Figure 5. The structure of the Segmentation Decoder module. Following [5], we adopt the pixel decoder and transformer decoder to excavate pixel-level visual information and instance-level object information. In contrast, we calculate the similarity between mask embeddings and multi-grained text embeddings as the mask scores for mask proposals' selection.

C. Different Numbers of Mask Tokens

In this section, we evaluate the performance of different numbers of mask tokens used in InstructSeg. As shown in Tab. 10, InstructSeg exhibits stronger reasoning and segmentation capability as the numbers of mask tokens increase. A greater number of mask tokens provides more candidate mask proposals and retains richer semantic information from the pre-trained decoder.

Why Multiple masks? Firstly, our decoder generates multiple masks and their corresponding mask scores. This design allows our model to fully utilize the semantic information and segmentation capabilities of pre-trained segmentation predictors such as Mask2Former [5] (like SAM [17] used in VISA [45]). Secondly, our paradigm can be easily extended in the future to process multiple objects in referring and reasoning segmentation tasks even combined with instance-level tasks, enabling it to handle more challenging and realistic scenarios. Finally, multiple masks can also bring significant performance enhancement as illustrated in Tab. 10

Table 10. Analysis of different numbers of mask tokens.

Mask Tokens	RefCOCO val	ReVOS
	cIoU	$\mathcal{J}\&\mathcal{F}$
2	81.4	46.0
5	81.9	46.6
10	82.7	47.1
50	84.5	50.4
100	**85.8**	**51.9**

D. Task-specific Instructions Design

In this section, we illustrate the text prompt with task-specific instructions for all the Instructed Visual Segmentation tasks. As shown in Tab. 11, we design different instruction templates for all four segmentation tasks along with corresponding text prompts.

E. More Qualitative Results

E.1. Referring Expression Segmentation (Image-level)

Fig. 6 shows the visualization of InstructSeg on referring segmentation task.

Table 11. Task-specific language instructions for all the Instructed Visual Segmentation tasks..

Task	Visual Type	Dataset	Instruction Template	Text Prompt
Referring Expression Segmentation	Image	RefCOCO/+/g	*You need to perform Referring Expression Segmentation on the image according to the Text Prompt.*	"A baseball catcher with an open mitt"
Reasoning Segmentation	Image	ReasonSeg	*You need to perform Reasoning Segmentation on the image according to the Text Prompt.*	"The person who appears to have already won in the battle"
Referring Video Object Segmentation	Video	Ref-YouTube-VOS, etc.	*You need to perform Referring Video Object Segmentation on the video according to the Text Prompt.*	"A duck is held by a person with her both hands"
Reasoning Video Object Segmentation	Video	ReVOS	*You need to perform Reasoning Video Object Segmentation on the video according to the Text Prompt.*	"Which person is in the leading position?"

Figure 6. Qualitative results of InstructSeg's capability in referring expression segmentation.

E.2. Reasoning Segmentation (Image-level)

Fig. 7 presents the effectiveness of our InstructSeg in understanding the complex question and performing segmentation according to the reasoning process.

E.3. Reasoning and Referring Video Object Segmentation (Video-level)

Fig. 8 shows the effectiveness of InstructSeg in comprehending both the reasoning questions and temporal coherence. InstructSeg is capable of producing segmentation masks that maintain consistency across temporal sequences.

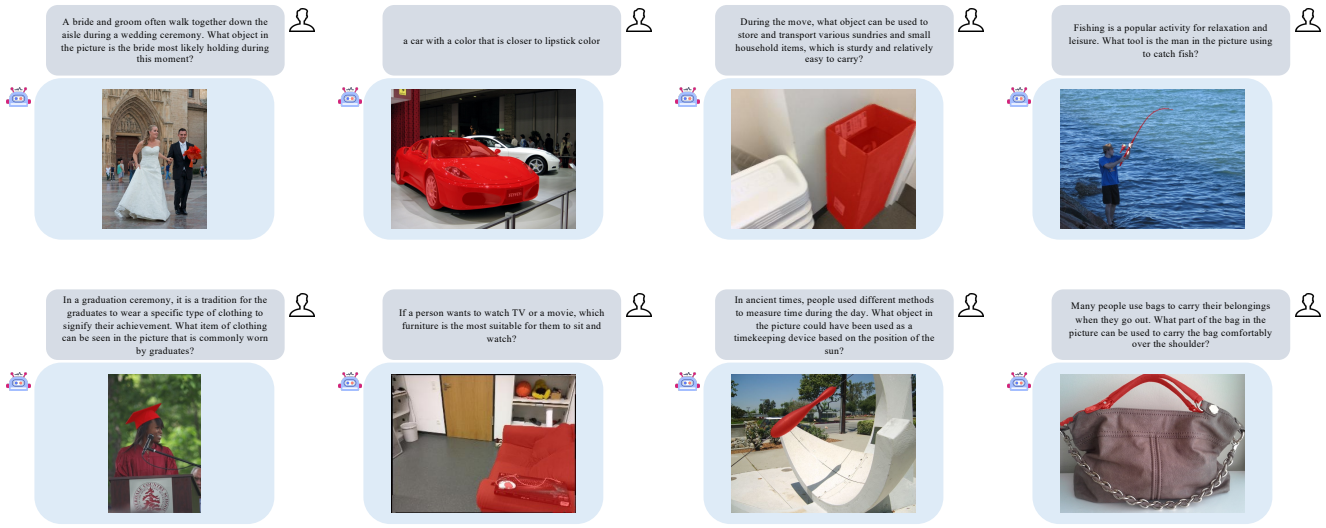

Figure 7. Qualitative results of InstructSeg in reasoning segmentation.

Figure 8. Qualitative results of InstructSeg demonstrate its capability in the complex reasoning video object segmentation task and referring video object segmentation task.