Table R1. Performance on referring expression segmentation benchmarks and grounded conversation generation benchmarks. The evaluation metric of RES benchmarks is cIoU. "ft" indicates finetuning on the referring expression datasets or grounded conversation datasets.

Method	Freeze Decoder	Visual Encoder	refCOCO			refCOCO+			refCOCOg		GCG			
			Val	TestA	TestB	Val	TestA	TestB	Val	Test	METEOR	CIDEr	AP_{50}	mIOU
LISA	×	2	74.1	76.5	71.1	62.4	67.4	56.5	66.4	68.5	-	-	-	-
LISA(ft)	×	2	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6	12.9	32.2	24.8	61.7
PixelLM	×	1	73.0	76.5	68.2	66.3	71.7	58.3	69.3	70.5	-	-	-	-
GSVA(ft)	×	2	77.2	78.9	73.5	65.9	69.6	59.8	72.7	73.3	-	-	-	-
GLaMM(ft)†	×	2	79.5	83.2	76.9	72.6	78.7	64.6	74.2	74.9	14.6	37.9	27.2	64.6
OMG-LLaVA	✓	1	76.3	77.7	71.2	67.7	69.7	58.9	70.7	70.2	13.5	33.1	26.1	62.8
OMG-LLaVA(ft)	×	1	78.0	80.3	74.1	69.1	73.1	63.0	72.9	72.9	14.5	38.5	28.6	64.7
OMG-LLaVA(ft)	✓	1	77.2	79.8	74.1	68.7	73.0	61.6	71.7	71.9	14.5	37.5	28.9	64.6

Table R2. Performance on the image-level benchmarks. † denotes using the IntermLM2-7B as the LLM, the same as OMG-LLaVA.

Method	MME↑	MMBench↑	SEED-Bench↑	POPE↑	AI2D↑
Training only with LLaVA 1.5 dataset					
LLaVA 1.5	1510	64.3	66.1	85.9	55.5
LLaVA 1.5†	1689 (1422/267)	68.5	65.9	86.7	56.6
OMG-LLaVA	1731 (1448/282)	67.5	68.9	89.7	61.7
Co-training with LLaVA dataset and segmentation datasets					
LISA	2 (1/1)	0.4	-	0.0	0.0
PixelLM	444 (309/135)	17.4	-	0.0	0.0
LaSagnA	0 (0/0)	0.0	-	0.0	0.0
GLaMM	23 (14/9)	36.8	-	0.94	28.2
OMG-LLaVA	1412 (1177/235)	47.9	56.5	80.0	42.9

Table R3. Performance with different LLMs.

LLM	refCOCO		refCOCO+		MME		MMBench	SEED-Bench	POPE	AI2D	MMstar	SQA
	CIoU	GIoU	CIoU	GIoU	perception	reasoning						
Phi3-3.8B	76.5	78.0	67.8	70.0	1291.6	265.0	59.6	60.6	86.7	56.9	37.1	64.7
InternLM2-7B	76.3	77.8	67.7	69.9	1177.1	235.4	47.9	56.5	80.0	42.9	33.1	57.8
Qwen2-7B	76.7	78.2	69.1	71.2	1215.7	251.1	62.8	60.7	84.3	52.6	37.2	66.4

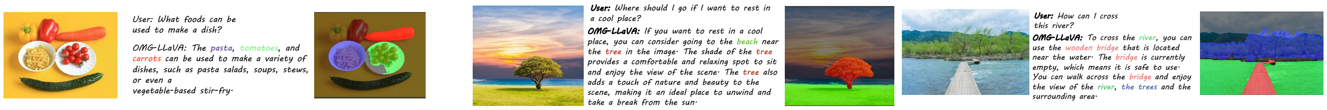
Table R4. Ablation study of projector for pixel-centric and object-centric visual tokens. "Cross Attn." indicates incorporating an additional cross-attention layer into the projector to facilitate interaction with image features. "O" denotes that a unified projector is employed for all object-centric tokens, encompassing object queries and visual prompt tokens. "O&P" signifies the utilization of a shared projector for both object-centric and pixel-centric tokens.

Methods	Cross Attn.	Share	refCOCO		refCOCO+		refCOCOg		refCOCOg(C)
			cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	METEOR
M0			72.3	74.1	60.8	63.5	65.4	68.6	13.1
M1		O	74.5	75.9	63.6	65.9	68.7	71.0	13.6
M2		O&P	71.1	72.8	60.3	63.2	64.8	68.4	12.4
M3	✓	O	72.3	73.7	60.6	63.0	66.5	69.2	13.2

Table R5. Ablation study on perception prior embedding.

Methods	refCOCO		refCOCO+		refCOCOg	
	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU
None	58.7	61.0	52.6	55.0	55.8	58.1
SoftMax	72.5	74.3	63.2	65.4	67.8	70.6
ArgMax	72.4	74.1	63.3	65.3	67.6	70.5
L1 Norm	72.0	73.8	62.9	65.0	67.4	70.3

Table R6. OMG-LLaVA can visualize LLM intentions through segmentation masks. This feature is crucial for intelligent assistants, enabling more convenient human-AI interactions.



User: What foods can be used to make a dish?

OMG-LLaVA: The pasta, tomatoes, and carrots can be used to make a variety of dishes, such as pasta salads, soups, stews, or even a vegetable-based stir-fry.

User: Where should I go if I want to rest in a cool place?

OMG-LLaVA: If you want to rest in a cool place, you can consider going to the beach near the tree in the image. The shade of the tree provides a comfortable and relaxing spot to sit and enjoy the view of the scene. The tree also adds a touch of nature and beauty to the scene, making it an ideal place to unwind and take a break from the sun.

User: How can I cross this river?

OMG-LLaVA: To cross the river, you can use the wooden bridge that is located near the water. The bridge is currently empty, which means it is safe to use. You can walk across the bridge and enjoy the view of the river, the trees and the surrounding area.