北京大学
PEKING UNIVERSITY

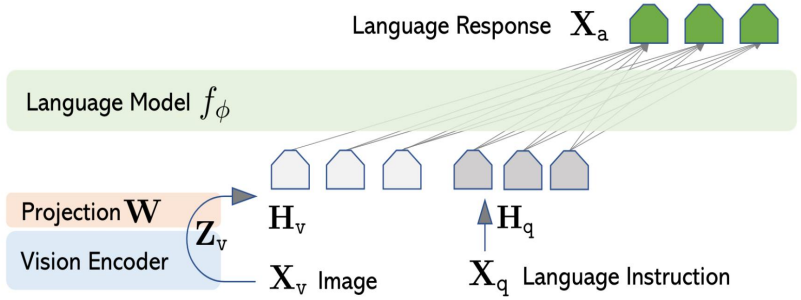
NEURAL INFORMATION PROCESSING SYSTEMS

UFO: A Unified Approach to Fine-grained Visual Perception via Open-ended Language Interface

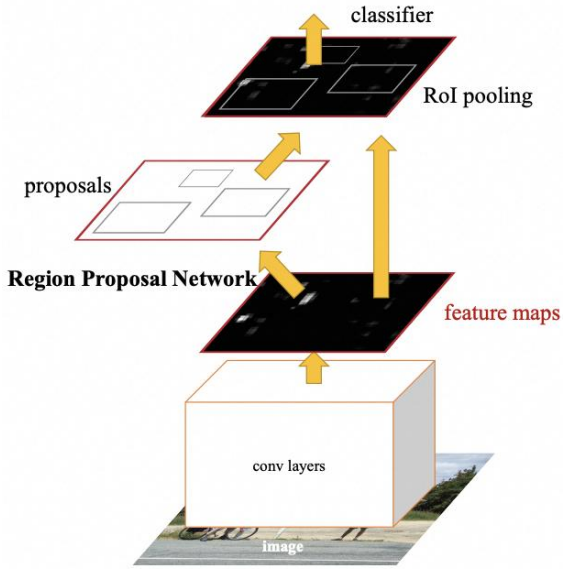
Hao Tang[1,2] **Chenwei Xie**[2] **Haiyang Wang**[1] **Xiaoyi Bao**[2,3]
Tingyu Weng[2] **Pandeng Li**[2] **Yun Zheng**[2†] **Liwei Wang**[1,4,5†]
[1]Center for Data Science, Peking University [2]Alibaba Group
[3] CASIA [4] Center for Machine Learning Research, Peking University
[5] State Key Laboratory of General Artificial Intelligence, Peking University, Beijing, China

Motivation

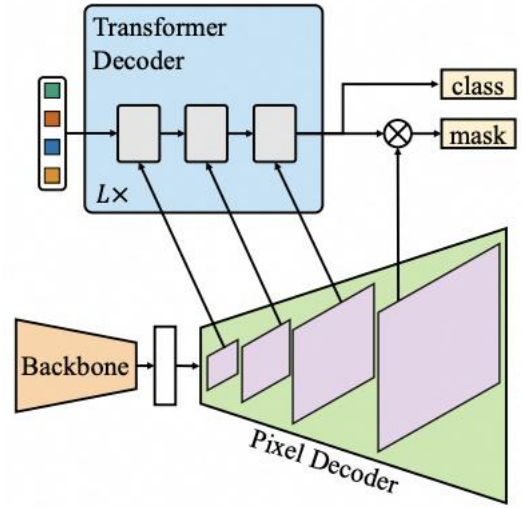
- Vision-language tasks have been unified in MLLMs
  - Transformer + Next token prediction



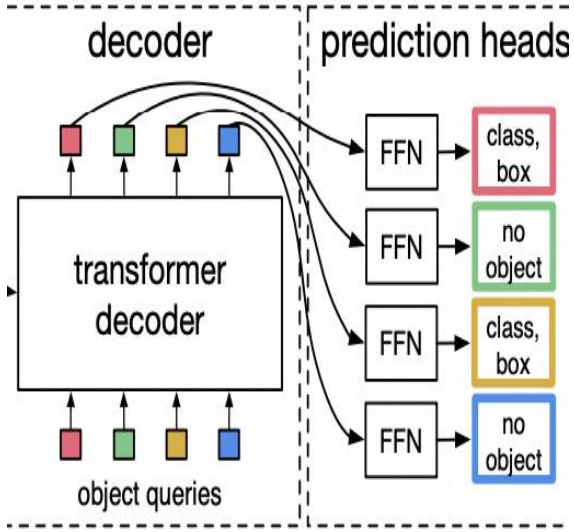
- Fine-grained visual perception: detection, segmenttion, depth estimation…
  - Task-specific modules



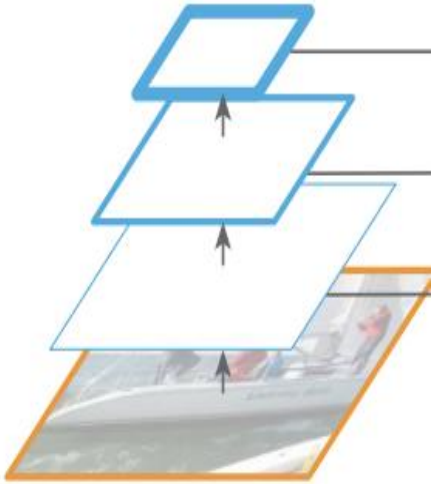
RPN



Mask Decoder



Object Queries

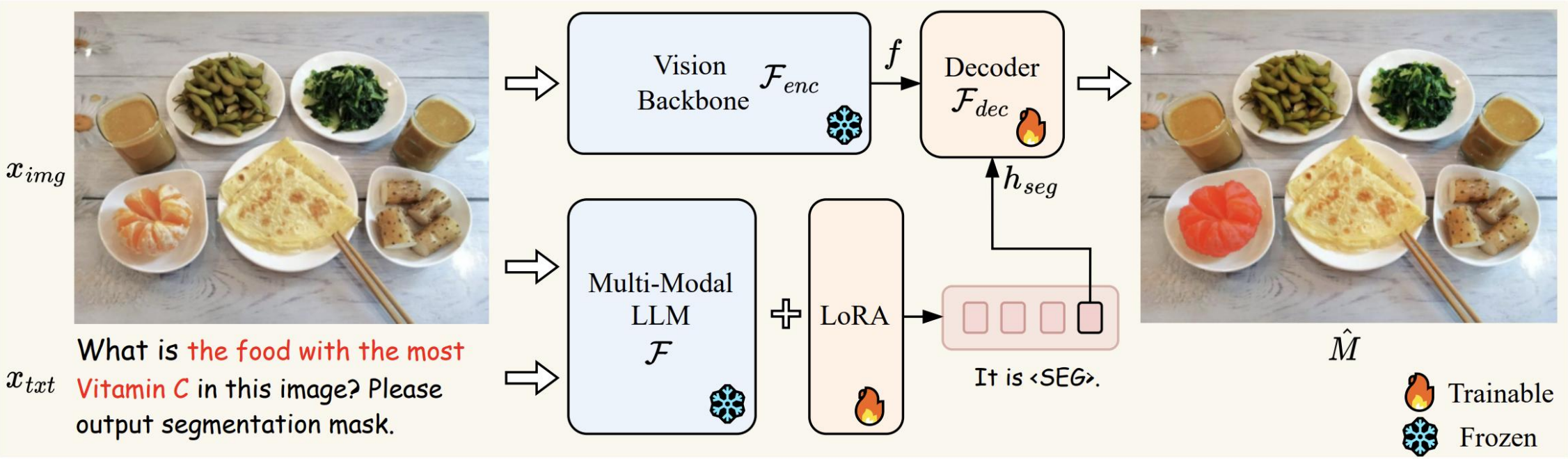


FPN

Motivation

- How to extend MLLMs with fine-grained perception?

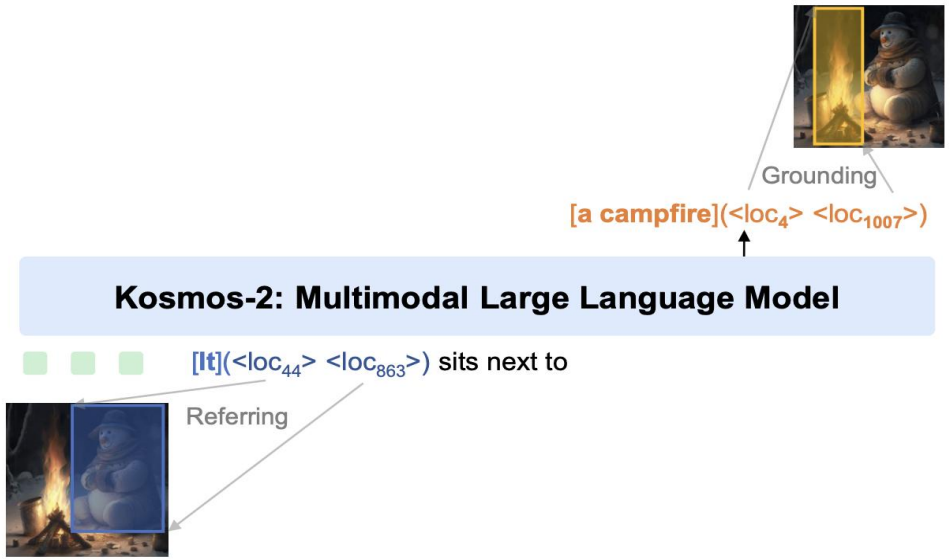
- MLLM + task decoder
  - SAM, Grounding DINO
- Complex architectures and training

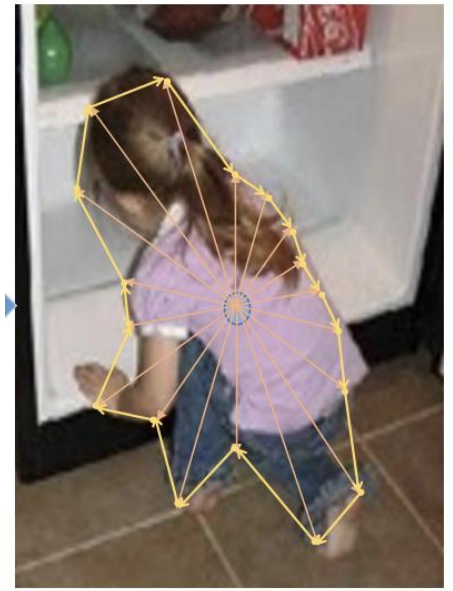


Motivation

- How to extend MLLMs with fine-grained perception?

- Conver box and mask to text
  - polygon for mask
- Hard to support multi-object detection
- Low mask performance by quantization errors



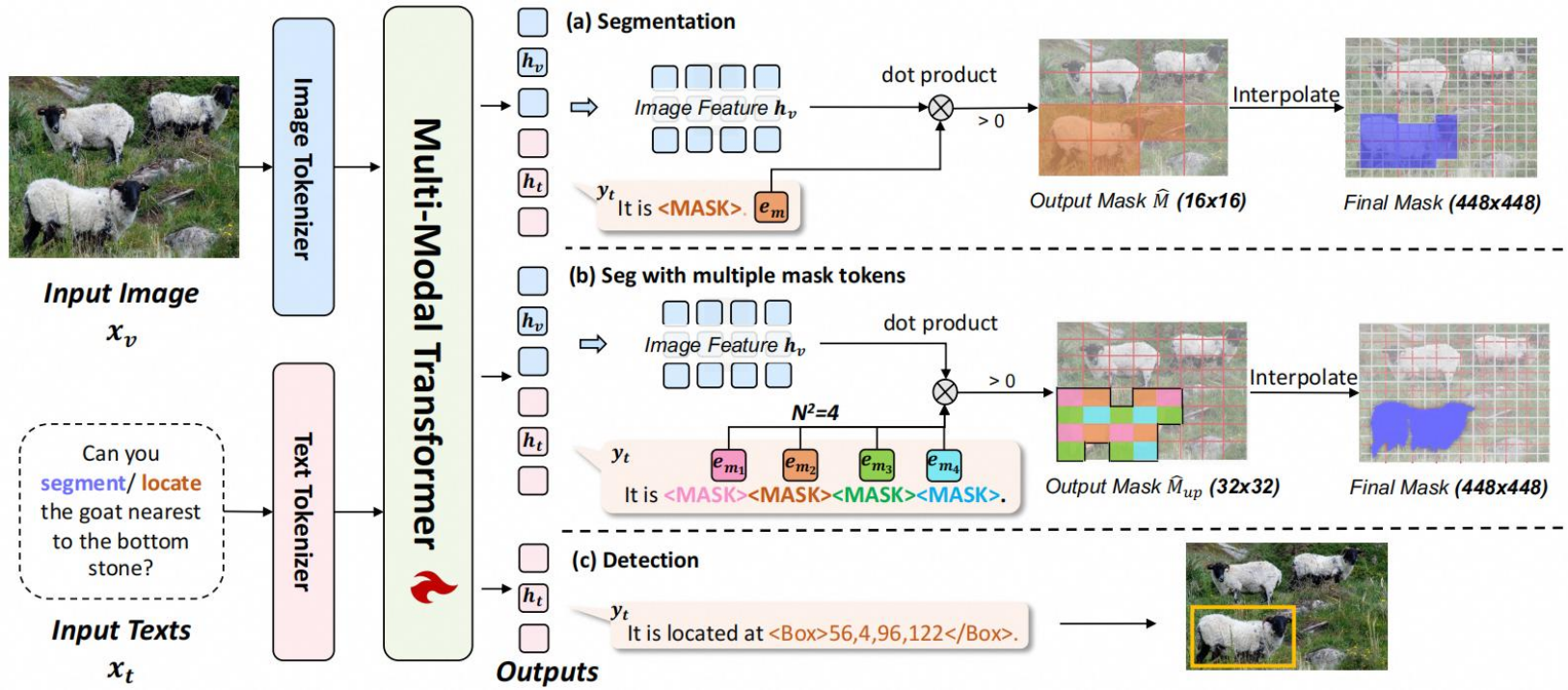


Methods	Instance Seg AP	AP_{50}	AP_{75}
Specialist Models			
Faster R-CNN-FPN [73]	-	-	-
DETR-DC5 [13]	-	-	-
Deformable-DETR [106]	-	-	-
Pix2Seq [21]	-	-	-
Mask R-CNN [36]	37.1	58.4	40.1
Polar Mask [93]	30.5	52.0	31.1
Mask2Former [25]	43.7	-	-

Method

- MLLM can answer “Where” and “What”
  - The category and mask information is in image features
- Modeling by similarity: Embedding Retrieval

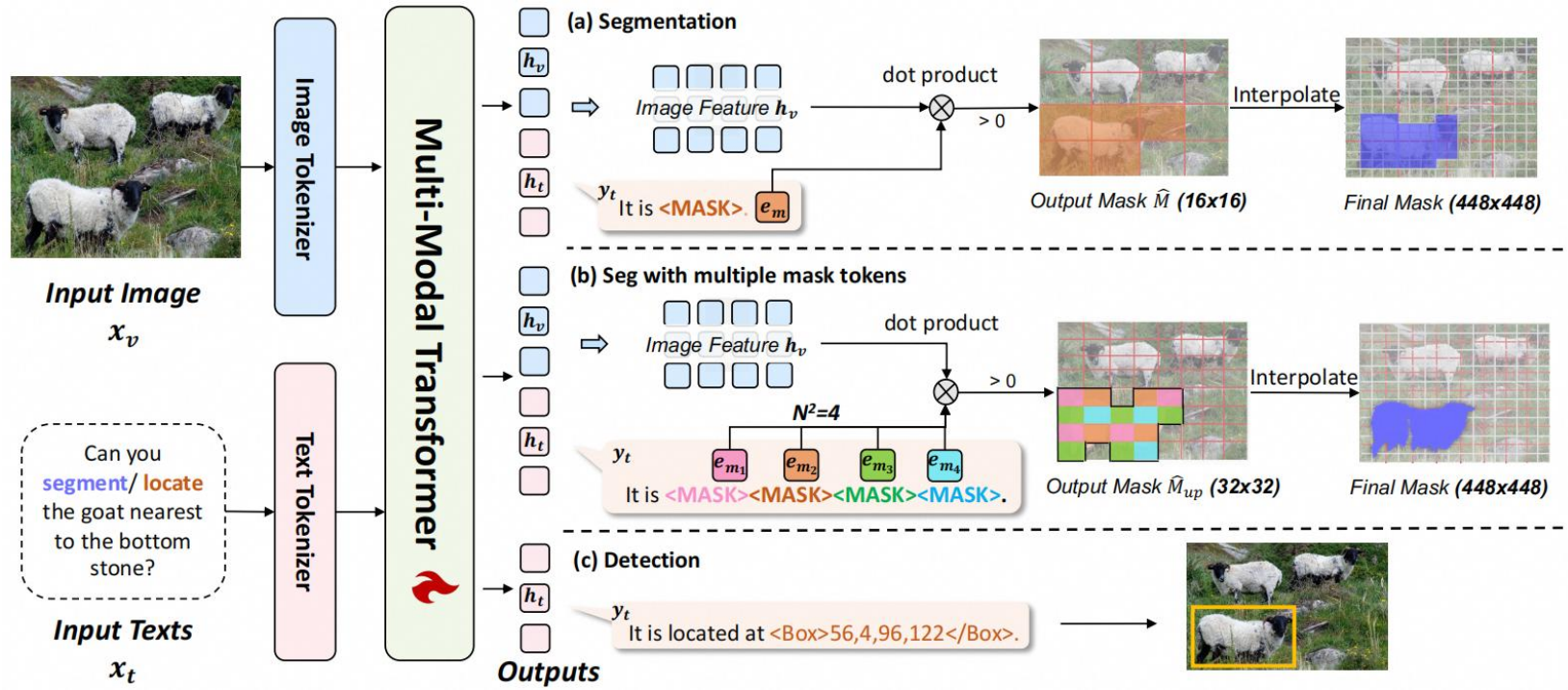
$$\mathbf{h_v}, \mathbf{y_t}, \mathbf{h_t} = \mathcal{F}(\mathbf{x_v}, \mathbf{x_t}). \qquad \boldsymbol{s} = \frac{\mathbf{e_m}\,\mathbf{h_v}^{\top}}{\sqrt{d}}, \quad \hat{\mathbf{M}} = \mathbb{I}(\boldsymbol{s} > 0),$$



Method

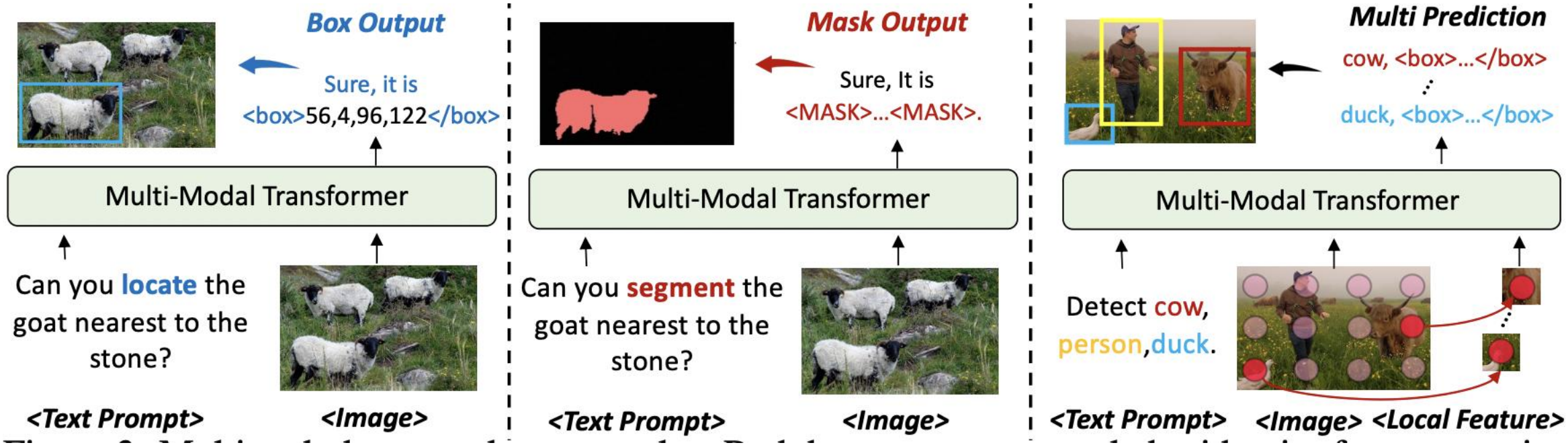
- Image features are typically downsampled
  - e.g. 28x in InternVL
- Upsampling by mutliple mask tokens

$$s_i = \frac{\mathbf{e_{m_i}}\,\mathbf{h_v}^\top}{\sqrt{d}}, \qquad s_{\text{concat}} = \text{concat}(\{s_i\}_{i=1}^{N^2}), \quad s_{\text{concat}} \in \mathbb{R}^{N^2 \times H_p \times W_p},$$
$$s_{\text{up}} = \text{reshape}(s_{\text{concat}}), \quad s_{\text{up}} \in \mathbb{R}^{(H_p N) \times (W_p N)}.$$



Method

- Multi prediction tasks: Object detection, instance segmentation
  - Long sequence, inefficient, difficult to learn
- Split multiple predictions to independent subtasks
- Parallel decoding



Training

北京大学 PEKING UNIVERSITY

- Multi-task training
  - Five tasks: COCO Det/InsSeg/Caption, ADE20K, RefCOCO
  - Evaluation: Multi-task benchmark in GiT
  - Architecture: UFO-ViT, UFO-InternVL2.5-8B

- Instruction Tuning
  - 26 datasets with diverse tasks
  - Evaluation: RefCOCO (REC and RES), ReasonSeg
  - Architecture: UFO-InternVL2.5-8B, UFO-LLaVA-1.5-7B

Experiments

北京大学 PEKING UNIVERSITY

- Multi-task benchmark

Methods	Specific Modules Examples	Num	#Params	Object Detection AP	AP_{50}	AP_{75}	Instance Seg AP	AP_{50}	AP_{75}	Semantic Seg mIoU(SS)	Captioning BLEU-4	CIDEr	REC Acc@0.5
Specialist Models													
Deformable-DETR [94]	RegressionHead	5	40M	45.4	64.7	49.0	-	-	-	-	-	-	-
Mask R-CNN [26]	FPN,RPN	6	46M	41.0	61.7	44.9	37.1	58.4	40.1	-	-	-	-
Polar Mask [79]	CenternessHead	5	55M	-	-	-	30.5	52.0	31.1	-	-	-	-
Mask2Former [13]	PixelDecoder	5	44M	-	-	-	43.7	-	-	47.2	-	-	-
VL-T5 [15]	Faster R-CNN	3	440M	-	-	-	-	-	-	-	34.5	116.5	-
MDETR [30]	RoBERTa,DETR	6	188M	-	-	-	-	-	-	-	-	-	86.8
Generalist Models (MultiTask-Training)													
Uni-Perceiver [95]	None	1	124M	-	-	-	-	-	-	-	32.0	⋆	⋆
Uni-Perceiver-MoE [93]	None	1	167M	-	-	-	-	-	-	-	33.2	⋆	⋆
VisionLLM-R50 [74]	Deform-DETR	6	7B	44.6	64.0	48.1	25.1	50.0	22.4	-	31.0	112.5	80.6
GiT-$B_{\text{single-task}}$ [69]	None	1	131M	45.1	62.7	49.1	31.4	54.8	31.2	47.7	33.7	107.9	83.3
GiT-$B_{\text{multi-task}}$ [69]	None	1	131M	46.7	64.2	50.7	31.9	56.4	31.4	47.8	35.4	112.6	85.8
GiT-$L_{\text{multi-task}}$ [69]	None	1	387M	51.3	69.2	55.9	35.1	61.4	34.7	50.6	35.7	116.0	88.4
GiT-$H_{\text{multi-task}}$ [69]	None	1	756M	52.9	71.0	57.8	35.8	62.6	35.6	52.4	36.2	118.2	89.2
UFO-ViT-$B_{\text{single-task}}$	None	1	131M	47.8	65.7	52.0	42.6	65.8	46.1	49.5	34.2	111.1	83.6
UFO-ViT-$B_{\text{multi-task}}$	None	1	131M	48.3	66.6	52.6	43.5	66.2	47.0	50.2	35.3	114.2	85.8
Improvement (single→multi)				***+0.5***	***+0.9***	***+0.6***	***+0.9***	***+0.4***	***+0.9***	***+0.7***	***+1.1***	***+3.1***	***+2.2***
UFO-ViT-$L_{\text{multi-task}}$	None	1	387M	52.9	71.3	57.9	47.3	70.9	51.6	54.0	35.9	118.6	88.5
UFO-ViT-$H_{\text{multi-task}}$	None	1	756M	**54.1**	**72.4**	**58.9**	**48.1**	**71.6**	**53.0**	**55.7**	37.6	123.6	89.2
UFO-InternVL2.5-8$B_{\text{multi-task}}$	None	1	8B	52.3	71.7	56.5	45.8	69.5	49.7	54.6	**39.6**	**131.6**	90.4

Experiments

北京大学 PEKING UNIVERSITY

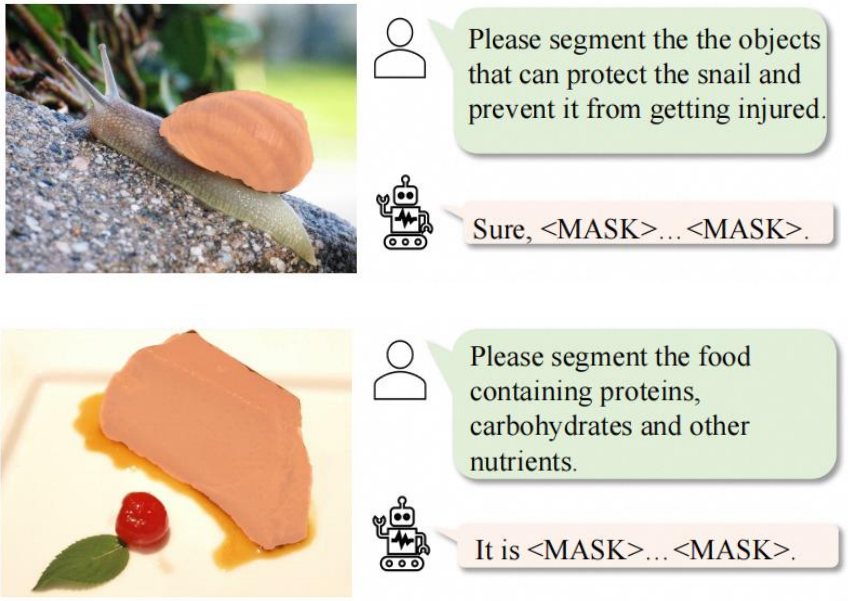
- REC and RES in RefCOCO

Methods	Referring Expression Comprehension (REC)									Referring Expression Segmentation (RES)								
	RefCOCO			RefCOCO+			RefCOCOg		Avg	RefCOCO			RefCOCO+			RefCOCOg		Avg
	val	testA	testB	val	testA	testB	val	test		val	testA	testB	val	testA	testB	val	test	
MLLMs with Task Decoders																		
GLaMM-7B [57]	-	-	-	-	-	-	-	-	-	79.5	**83.2**	76.9	72.6	78.7	64.6	74.2	74.9	75.6
SAM4MLLM-8B [11]	-	-	-	-	-	-	-	-	-	79.8	82.7	74.7	74.6	80.0	67.2	75.5	76.4	76.4
HiMTok-8B [73]	-	-	-	-	-	-	-	-	-	**81.1**	81.2	**79.2**	**77.1**	78.8	71.5	75.8	76.7	77.7
PerceptionGPT-7B [53]	88.6	92.5	84.6	82.1	88.6	74.2	84.1	85.2	85.0	75.1	78.6	71.7	68.5	73.9	61.3	70.3	71.7	71.4
VisionLLM v2 [77]	90.0	93.1	87.1	81.1	87.3	74.5	85.0	86.4	85.6	79.2	82.3	77.0	68.9	75.8	61.8	73.3	74.8	74.1
MLLMs w/o Task Decoders																		
Shirka-7B [7]	87.0	90.6	80.2	81.6	87.4	72.1	82.3	82.2	82.9	-	-	-	-	-	-	-	-	-
MiniGPT-v2-7B [6]	88.1	91.3	84.3	79.6	85.5	73.3	84.2	84.3	83.8	-	-	-	-	-	-	-	-	-
Ferret-v2-7B [85]	92.8	94.7	88.7	87.4	**92.8**	79.3	**89.4**	**89.3**	89.3	-	-	-	-	-	-	-	-	-
VistaLLM-7B [54]	88.1	91.5	83.0	82.9	89.8	74.8	83.6	84.4	84.8	74.5	76.0	72.7	69.1	73.7	64.0	69.0	70.9	71.2
UFO-LLaVA-1.5-7B	90.2	93.5	87.3	84.4	90.3	78.7	86.4	86.8	87.2	77.2	80.1	76.4	71.8	77.9	70.2	74.1	73.5	75.2
UFO-LLaVA-1.5-7B*	91.1	93.7	88.6	85.5	90.5	79.9	87.3	87.2	88.0	77.9	81.1	77.0	72.5	78.5	71.4	75.6	74.1	76.0
UFO-InternVL2.5-8B	91.8	94.3	87.5	86.9	91.3	80.6	87.9	88.6	88.6	80.0	81.6	78.1	76.7	79.9	72.3	75.5	76.3	77.6
UFO-InternVL2.5-8B*	**93.1**	**94.8**	**89.2**	**87.7**	92.1	**82.3**	88.2	89.2	**89.6**	81.0	82.6	78.6	**77.1**	**80.4**	**72.6**	**76.7**	**77.3**	**78.3**

Experiments

- Best performance on ReasonSeg
  - Outperfoms by 6.2 gIoU

Methods	ReasonSeg overall	ReasonSeg short query	ReasonSeg long query
X-Decoder [91]	21.7	20.4	22.2
SEEM [92]	24.3	20.1	25.6
LISA-7B [34]	36.8	37.6	36.6
LISA-7B [34]*	47.3	40.6	49.4
Cores-7B [2]	48.7	41.0	50.9
Cores-7B [2]*	52.4	44.2	55.0
HiMTok-8B [68]*	60.8	-	-
UFO-LLaVA-1.5-7B	54.4	41.2	58.5
UFO-LLaVA-1.5-7B*	58.8	46.5	62.7
UFO-InternVL2.5-8B	60.0	48.7	63.6
UFO-InternVL2.5-8B*	**67.0**	**56.2**	**70.4**



- Extend to depth estimation and surface normal prediction

Methods	RMSE↓	$\delta 1 \uparrow$	REL↓	log10↓
Painter [70]	0.327	0.930	0.090	-
Unified-IO 2 [44]	0.423	-	-	-
UFO-InternVL2.5-8B	0.305	0.936	0.087	0.035

Method	Mean	Median	11.25°	22.5°	30°
GeoNet++ [55]	18.5	11.2	9.502	0.732	0.907
Marigold [32]	18.8	-	0.559	-	-
GeoWizard [22]	17.0	-	0.565	-	-
UFO-InternVL2.5-8B	17.8	10.4	0.543	0.733	0.800

Conclusion

北京大学 PEKING UNIVERSITY

- UFO reformulate segmentation as embedding retrieval
  - Remove the need for task decoders
  - Fully aligned with open-ended Language interface

- UFO explore the image representation capabilities of MLLMs
  - A general way to extract information in image features

- UFO unifies both single-prediction and mutli-prediction tasks
  - Parallel decoding improves efficiency and performance