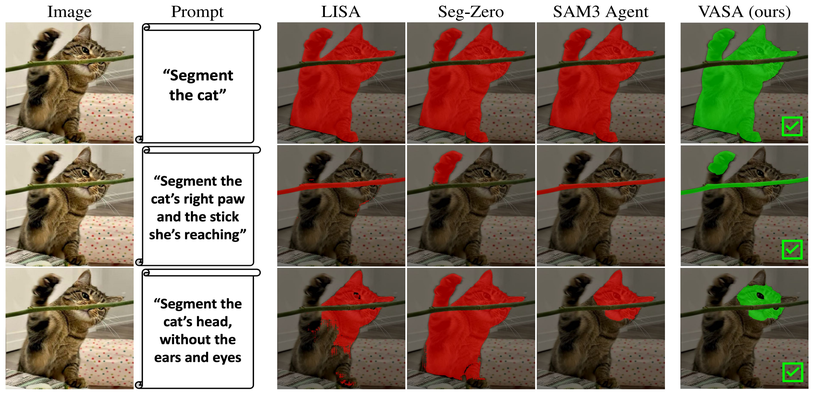
VASA: Vision Harnessing Agent for Open Ad-hoc Segmentation

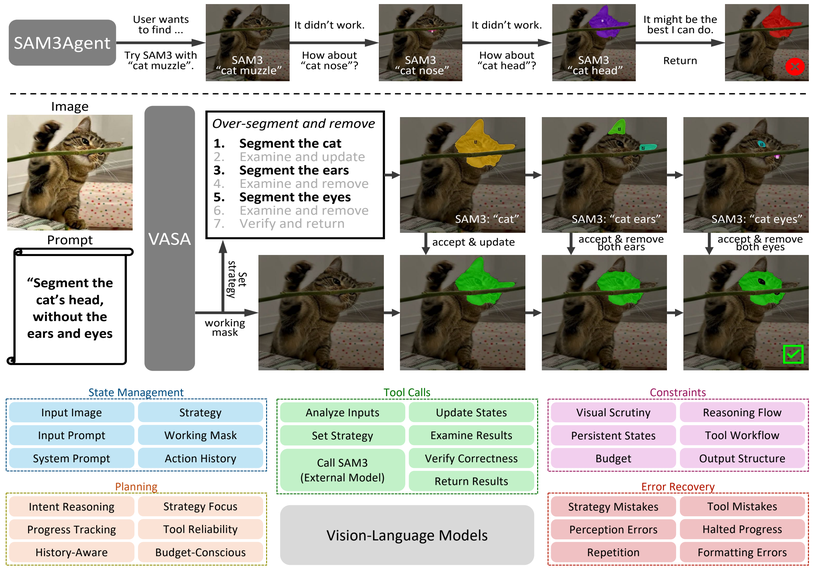
Zilin Wang Stella X. Yu

UNIVERSITY OF MICHIGAN

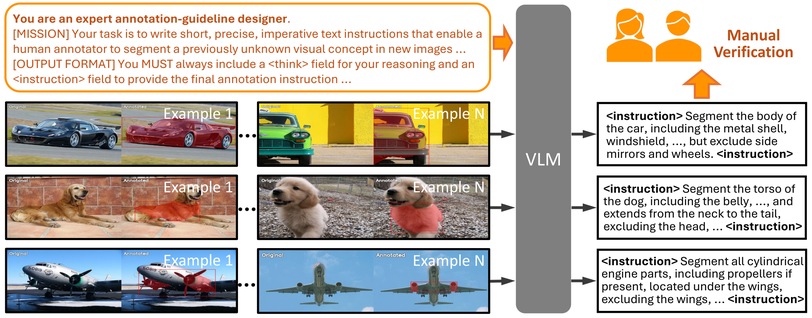
Open Ad-hoc Segmentation Builds Visual Concepts on the Fly



VASA Reasons, Constructs, Validates the Solution for OAS



PARS Benchmark: PartImageNet Ad-hoc Referring Segmentation



$$\text{gIoU} = \frac{1}{N}\sum_i \frac{|P_i \cap G_i|}{|P_i \cup G_i|}, \quad \text{cIoU} = \frac{\sum_i |P_i \cap G_i|}{\sum_i |P_i \cup G_i|}, \quad \text{xIoU} = \frac{1}{N}\sum_i \frac{|P_i \cap O_i|}{|P_i|}$$

where P_i is the prediction, G_i is the GT, O_i is the union of all GT except the target

SOTA on Ad-hoc Referring Segmentation PARS

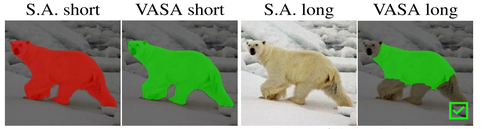
Method	VLM Version	Train. Free	Ad-hoc Concepts			Common Concepts			Total		
			gIoU	cIoU	xIoU↓	gIoU	cIoU	xIoU↓	gIoU	cIoU	xIoU↓
Trained on PartImageNet											
VLPart*	N/A	N/A	55.2	53.8	35.1	60.5	73.3	10.7	57.5	63.2	23.9
Open-Vocabulary Segmentation											
OVSeg	N/A	N/A	34.4	27.6	42.2	46.0	43.0	20.7	39.3	34.2	33.5
Semantic-SAM†	N/A	N/A	39.2	30.4	45.2	51.7	53.7	25.7	44.5	39.4	38.4
SAM3	N/A	N/A	34.1	36.1	59.8	41.8	57.8	30.6	37.3	45.0	48.9
VLM as Visual Reasoner											
LISA	Llama2 13B	✗	37.9	39.9	45.0	49.0	59.7	19.2	42.6	49.4	32.9
LISA++	LLaVA1.5 7B	✗	35.2	33.7	52.6	51.8	50.4	25.4	42.3	41.5	41.0
VisionReasoner	Qwen2.5-VL 7B	✗	27.2	28.8	47.9	29.4	41.9	16.5	28.1	35.0	34.5
Seg-Zero	Qwen2.5-VL 7B	✗	40.8	39.6	58.2	57.2	66.9	30.3	47.8	50.8	47.4
RESAnything	Qwen2.5-VL 7B	✓	33.1	34.0	49.7	45.2	51.0	22.8	38.2	41.8	38.1
Context Agent for Segmentation											
CoReS	LLaVA 7B	✗	33.0	34.1	36.2	42.2	53.4	19.7	36.9	44.0	27.5
SegAgent-SC	Qwen-VL 7B	✗	34.6	34.8	46.6	48.2	61.8	20.9	40.4	47.9	33.8
SegAgent-SAM	Qwen-VL 7B	✗	33.1	31.7	48.2	45.9	56.0	19.8	38.5	43.1	35.0
SAM3 Agent	Qwen3-VL 32B	✓	40.5	39.1	58.8	55.0	62.1	27.5	46.4	48.1	48.0
Vision Harnessing Agent for Segmentation											
VASA (ours)	Qwen3-VL 32B	✓	**54.0**	**56.9**	**33.5**	**60.8**	**77.0**	**10.0**	**56.9**	**66.5**	**22.9**
v.s. prior SOTA			+13.5	+17.8	+25.3	+ 5.8	+14.9	+17.5	+10.5	+18.4	+25.1

SOTA on Multi-Granularity Referring Segmentation RefCOCOm

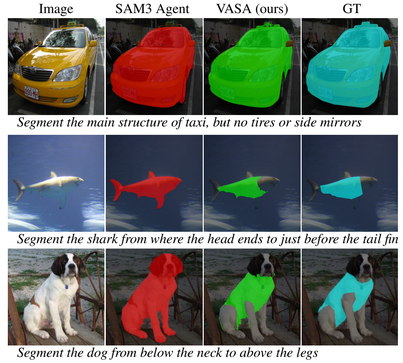
Method	val		testA		testB	
	Part	+Obj	Part	+Obj	Part	+Obj
X-Decoder	16.2	29.5	13.6	23.6	20.3	33.8
SEEM	16.1	29.4	13.6	23.4	20.4	33.9
UniRES	19.6	34.3	16.4	27.8	25.2	41.7
SAM3	12.3	17.3	9.0	13.0	20.7	26.1
LISA	21.3	34.3	18.5	28.6	25.7	40.1
GSVA	11.4	23.1	9.2	19.2	16.8	28.2
GLaMM	21.4	35.3	18.6	29.5	26.9	41.1
M^2SA	22.4	35.5	19.9	30.1	27.1	41.4
RESAnything	27.6	-	26.5	-	25.8	-
CoReS	21.2	29.5	17.3	24.0	20.6	30.6
SegAgent-SC	19.7	33.4	16.8	27.3	22.9	38.0
SegAgent-SAM	19.7	32.9	16.9	27.1	22.6	36.7
SAM3 Agent	40.2	48.1	34.4	41.5	41.3	50.4
VASA (ours)	**45.0**	**51.1**	**43.2**	**47.4**	**47.6**	**54.1**
v.s. prior SOTA	+4.8	+3.0	+8.8	+5.9	+6.3	+3.7

VASA Effectively Handles Detailed Long Queries

Method	Text	gIoU	cIoU	xIoU↓
SAM3 Agent	short	**46.4**	**48.1**	48.0
SAM3 Agent	long	24.3	34.5	**30.7**
VASA (ours)	short	51.0	51.3	41.7
VASA (ours)	long	**56.9**	**66.5**	**22.9**



PARS Visualizations



1. The open ad-hoc segmentation task.
2. VASA, the first vision harnessing agent with persistent visual state and long-horizon visual construction for OAS.
3. PARS, a dataset with detailed long-form queries for evaluating the OAS task.
4. SOTA on both PARS and RefCOCOm.
5. Points to a path for AI agents beyond wrapping foundation models as tools: programming them with task knowledge, VLM behaviors, visual routines, working memory, and failure-aware workflow.