NEURAL INFORMATION PROCESSING SYSTEMS

Leveraging Hallucinations to Reduce Manual Prompt Dependency in Promptable Segmentation

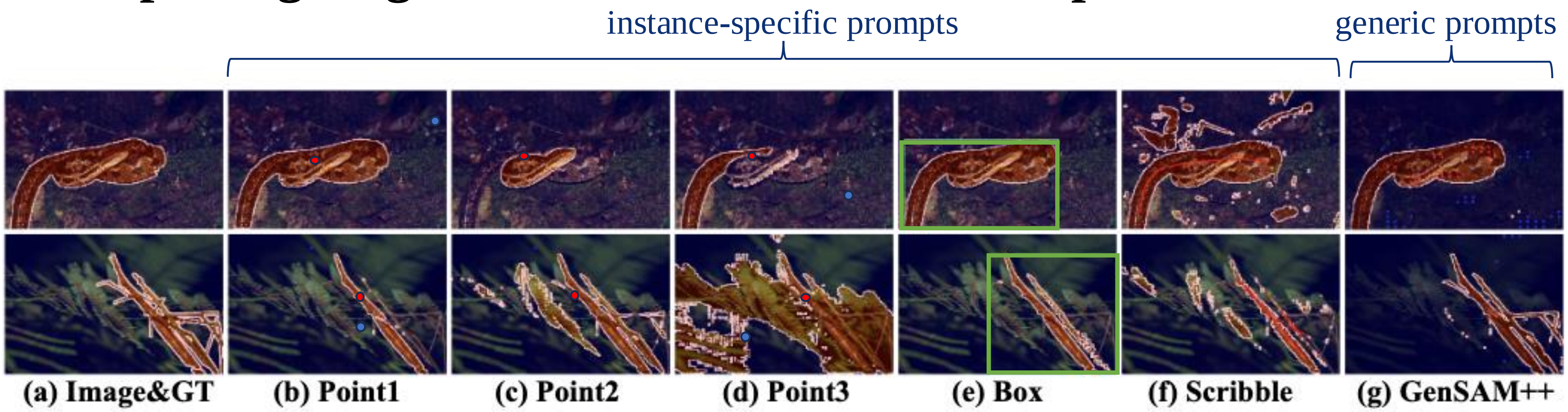
Jian Hu, Jiayi Lin, Junchi Yan, Shaogang Gong

Queen Mary University of London, Shanghai Jiao Tong University

Project page: https://lwpyh.github.io/ProMaC/

Background

- In SAM, manual point and scribble prompts suffer from **ambiguity in interpreting targets and is sensitive to minor spatial variations**.

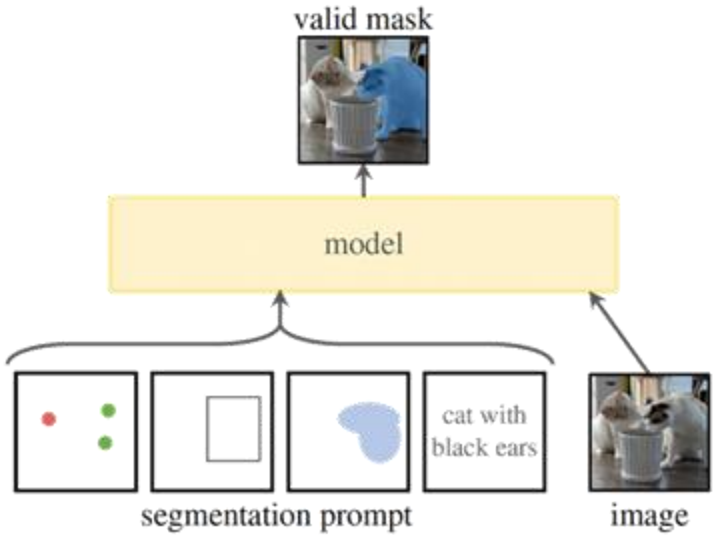


Segmentation results using different prompts in SAM with various approaches.

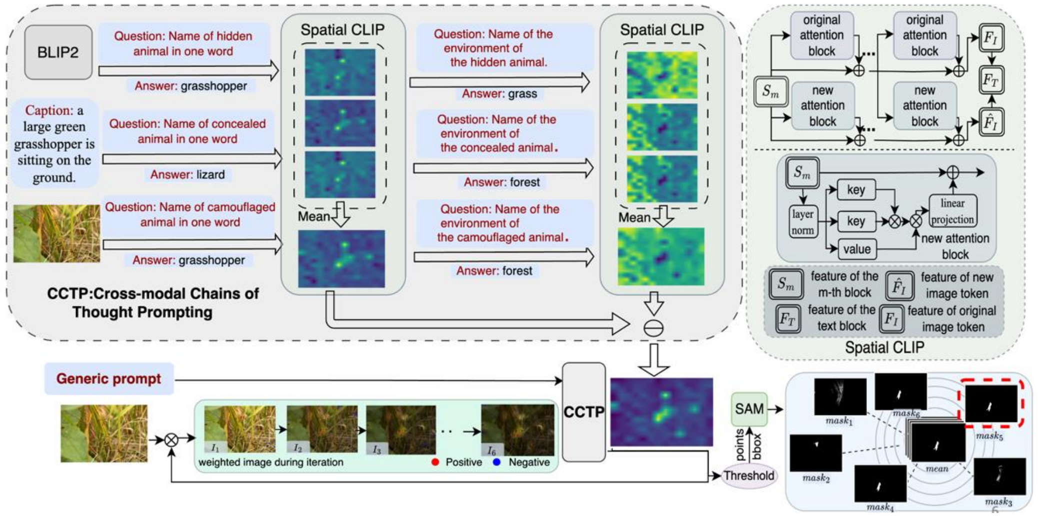
[1] Hu J, Lin J, Gong S, et al. Relax Image-Specific Prompt Requirement in SAM: A Single Generic Prompt for Segmenting Camouflaged Objects[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(11): 12511-12518.

Limitation of current promptable segmentation

- Segment Anything Model (SAM)
  - Learning from 1.1 billion prompt-mask pairs
  - Better generalization ability
  - Relies on Manual Instance-Specific Prompt



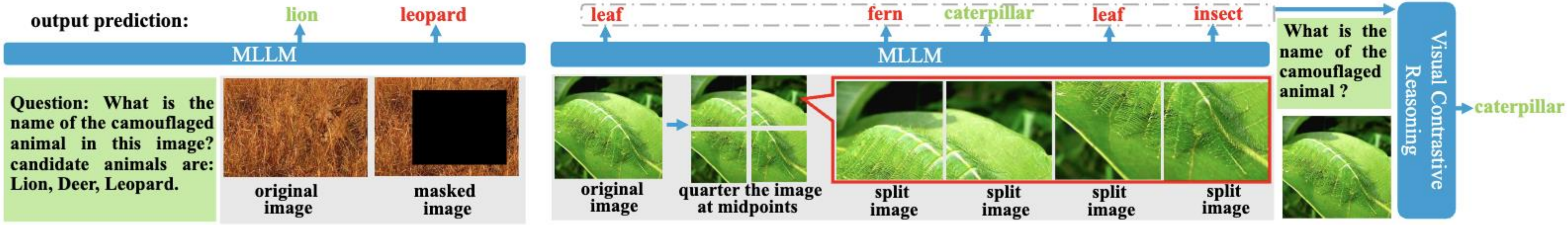
- GenSAM
  - Only need one task-generic prompt for each task
  - Can be generalized to more tasks
  - Generated instance-specific prompts are inaccurate



[1] Kirillov, Alexander, et al. Segment anything. arXiv:2304.02643 (2023).

[2] Hu J, Lin J, Gong S, et al. Relax Image-Specific Prompt Requirement in SAM: A Single Generic Prompt for Segmenting Camouflaged Objects[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(11): 12511-12518.

Motivation



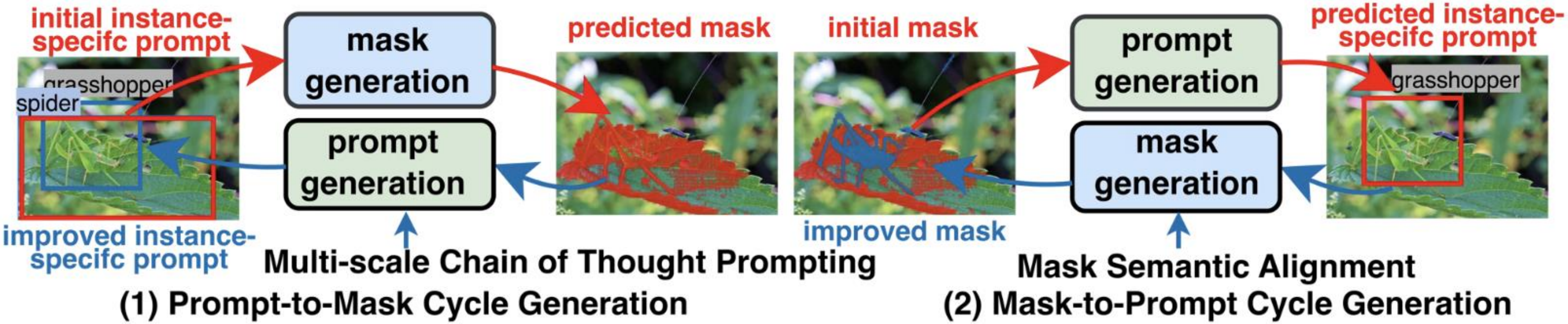
(a) Hallucination by co-occurrence prior. (b) Using hallucinations can benefit accuracte prompt generation.

(a) During MLLM pretraining, leopards often co-occur with grass. If the lion is masked, the model incorrectly identifies it as a leopard based on the grass.

(b) Directly inputting the image into MLLM causes the hidden caterpillar being incorrectly predicted as a leaf. Splitting the image results in interested objects being incomplete or absent, prompting MLLM to induce hallucinations and utilize prior knowledge to predict potential task-related objects within the image.

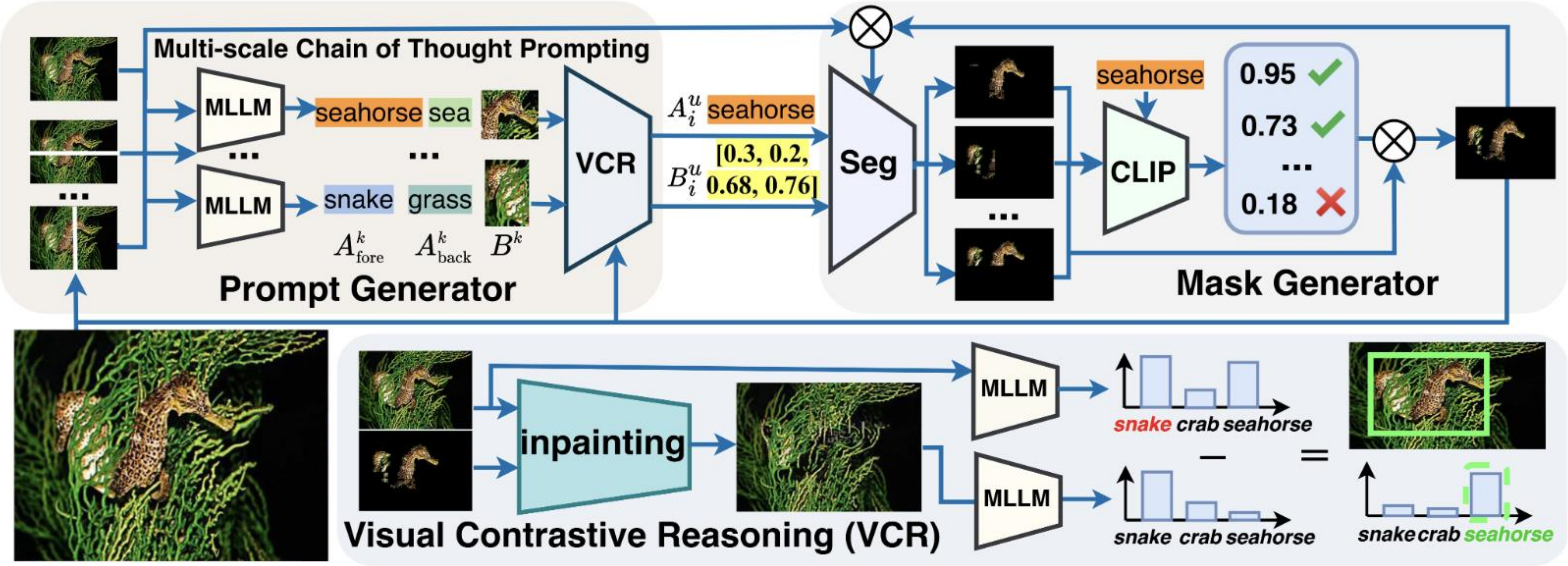
Our visual contrastive reasoning eliminates the hallucinations and validates the gathered predictions, aiding in the accurate identification of the caterpillar.

Key idea



An overview of ProMac: Masks created iteratively by the mask generator guide the prompt generator to jointly improve instance-specific prompts and visual masking in segmentation.

Framework of ProMaC



Experiments

Table 1: Results on Camouflaged Object Detection (COD) under different settings. Best are in **bold**.

Methods	Venue	Camouflaged Object Detection											
		CHAMELEON [50]				CAMO [30]				COD10K [14]			
		$M \downarrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
Scribble Supervision Setting													
WSSA[62]	CVPR20	0.067	0.692	0.860	0.782	0.118	0.615	0.786	0.696	0.071	0.536	0.770	0.684
SCWS[60]	AAAI21	0.053	0.758	0.881	0.792	0.102	0.658	0.795	0.713	0.055	0.602	0.805	0.710
TEL[62]	CVPR22	0.073	0.708	0.827	0.785	0.104	0.681	0.797	0.717	0.057	0.633	0.826	0.724
SCOD[17]	AAAI23	**0.046**	**0.791**	**0.897**	0.818	**0.092**	0.709	0.815	0.735	0.049	0.637	0.832	0.733
SAM-S[29]	ICCV23	0.076	0.729	0.820	0.650	0.105	0.682	0.774	0.731	0.046	0.695	0.828	0.772
WS-SAM[16]	NeurIPS23	**0.046**	0.777	**0.897**	**0.824**	**0.092**	**0.742**	**0.818**	**0.759**	**0.038**	**0.719**	**0.878**	**0.803**
Point Supervision Setting													
WSSA[62]	CVPR20	0.105	0.660	0.712	0.711	0.148	0.607	0.652	0.649	0.087	0.509	0.733	0.642
SCWS[60]	AAAI21	0.097	0.684	0.739	0.714	0.142	0.624	0.672	0.687	0.082	0.593	0.777	0.738
TEL[62]	CVPR22	0.094	0.712	0.751	0.746	0.133	0.662	0.674	0.645	0.063	0.623	0.803	0.727
SCOD[17]	AAAI23	0.092	0.688	0.746	0.725	0.137	0.629	0.688	0.663	0.060	0.607	0.802	0.711
SAM[29]	ICCV23	0.207	0.595	0.647	0.635	0.160	0.597	0.639	0.643	0.093	0.673	0.737	0.730
SAM-P[29]	ICCV23	0.101	0.696	0.745	0.697	0.123	0.649	0.693	0.677	0.069	0.694	0.796	0.765
WS-SAM[16]	NeurIPS23	**0.056**	**0.767**	**0.868**	**0.805**	**0.102**	**0.703**	**0.757**	**0.718**	**0.039**	**0.698**	**0.856**	**0.790**
Task-Generic Prompt Setting													
CLIP_Surgey+SAM	Arxiv23	0.147	0.606	0.741	0.689	0.189	0.520	0.692	0.612	0.173	0.488	0.698	0.629
GPT4V+SAM [43, 29]	Arxiv23	0.180	0.557	0.710	0.637	0.206	0.466	0.666	0.573	0.187	0.448	0.672	0.601
LLaVA1.5+SAM [37, 29]	NeurIPS23	0.168	0.561	0.718	0.666	0.314	0.401	0.585	0.501	0.170	0.530	0.728	0.662
X-Decoder [69]	CVPR23	0.124	0.654	0.748	0.716	0.104	0.628	0.745	0.709	0.171	0.556	0.705	0.652
SEEM [71]	NeurIPS23	0.094	0.011	0.307	0.454	0.192	0.023	0.315	0.404	0.143	0.001	0.280	0.425
GroundingSAM [29, 38]	ICCV23	0.122	0.662	0.776	0.744	0.157	0.656	0.753	0.707	0.085	0.670	0.813	0.764
GenSAM [21]	AAAI24	0.073	0.696	0.806	0.774	0.106	0.669	0.798	0.729	0.058	0.695	0.843	0.783
ProMaC	Ours	**0.044**	**0.790**	**0.899**	**0.833**	**0.090**	**0.725**	**0.846**	**0.767**	**0.042**	**0.716**	**0.876**	**0.805**

Experiments

Table 2: Results for Medical Image Segmentation (MIS) under task-generic prompt setting.

Methods	Venue	Polyp Image Segmentation								Skin Lesion Segmentation			
		CVC-ColonDB [51]				Kvasir [25]				ISIC [10]			
		$M \downarrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
GPT4V+SAM [43, 29]	Arxiv23	0.578	0.051	0.246	0.242	0.614	0.128	0.236	0.253	0.514	0.387	0.366	0.334
LLaVA1.5+SAM [37, 29]	NeruIPS23	0.491	0.194	0.355	0.357	0.479	0.293	0.400	0.403	0.369	0.473	0.497	0.477
X-Decoder [69]	CVPR23	0.462	0.095	0.327	0.331	0.449	0.202	0.371	0.384	0.338	0.315	0.127	0.407
SEEM [71]	NeruIPS23	0.570	0.085	0.280	0.284	0.520	0.215	0.339	0.367	0.362	0.250	0.002	0.280
GroundingSAM [29, 38]	ICCV23	0.711	0.071	0.195	0.206	0.387	0.353	0.521	0.468	0.301	0.348	0.247	0.533
GenSAM [21]	AAAI24	0.244	0.059	0.494	0.379	0.172	0.210	0.619	0.487	0.171	0.699	0.744	0.678
ProMaC	Ours	**0.176**	**0.243**	**0.583**	**0.530**	**0.166**	**0.394**	**0.726**	**0.573**	**0.168**	**0.717**	**0.755**	**0.689**

Table 3: Result on Transparent Object Segmentation and Open-Vocabulary Segmentation Tasks.

(a) Transparent Object Segmentation.

Methods	GSD [34]				Trans10K-hard [56]			
	$M \downarrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
GPT4V+SAM [43, 29]	0.312	0.104	0.392	0.363	0.288	0.199	0.607	0.512
LLaVA1.5+SAM [37, 29]	0.197	0.202	0.545	0.433	0.272	0.167	0.621	0.555
X-Decoder [69]	0.191	0.240	0.643	0.480	0.568	**0.611**	0.218	0.280
SEEM [71]	0.184	0.224	0.573	0.479	0.557	0.501	0.013	0.256
GroundingSAM [29, 38]	0.168	0.230	0.572	0.483	0.436	0.415	0.047	0.424
GenSAM [21]	0.155	0.394	0.700	0.559	0.263	0.489	0.612	0.536
ProMaC	**0.147**	**0.409**	**0.723**	**0.569**	**0.251**	0.509	**0.654**	**0.557**

(b) Open-vocabulary Segmentation.

Methods	Venue	Seg. Anno.	Image-Text pairs	VOC	Context	Object
				mIoU ↑	mIoU ↑	mIoU ↑
MaskCLIP[67]	ECCV22	-	-	38.8	23.6	20.6
TCL [6]	CVPR23	-	CC3M [48], CC12M [7]	51.2	24.3	**30.4**
GroupViT [57]	CVPR22	-	CC12M [7], YFCC14M [53]	52.3	22.4	-
ViewCo [46]	ICLR23	-	CC12M [7], YFCC14M [53]	52.4	23.0	23.5
SegCLIP [39]	ICML23	COCO [35]	CC [48]	52.6	24.7	26.5
OVSegmentor [58]	CVPR23	-	CC12M [7]	53.8	20.4	25.1
ProMaC	Ours	-	-	**59.3**	**30.7**	25.2

Visualization

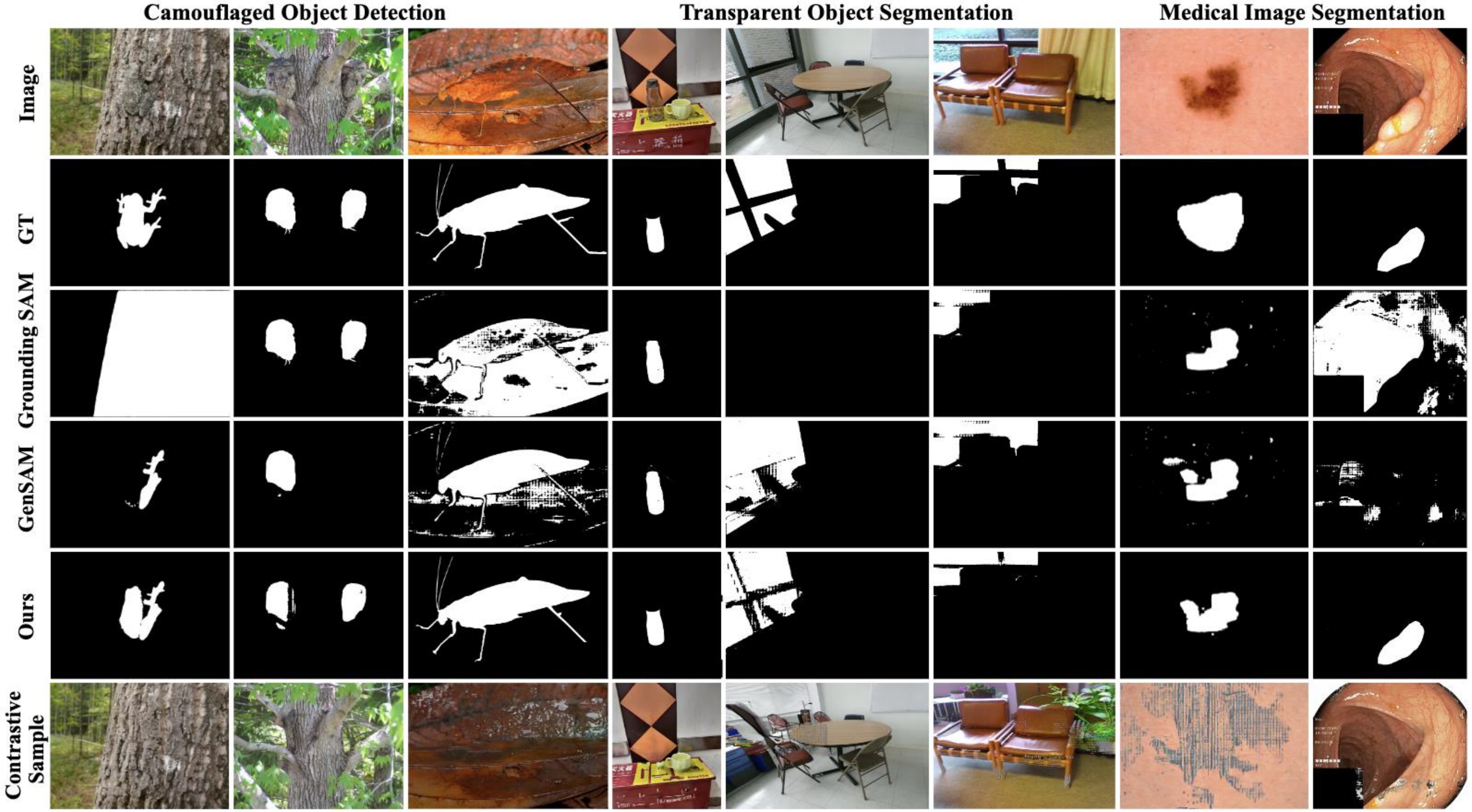


Figure 4: Visualization of various segmentation methods among various segmentation tasks.

NEURAL INFORMATION PROCESSING SYSTEMS

Ablation studies

Table 4: Ablation Study on COD and MIS Tasks

Method's Variants					CHAMELEON [50]				CVC-ColobNB [51]			
MCoT	IVP	ITP	VCR	MSA	$M \downarrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
	✓	✓	✓	✓	0.052	0.764	0.885	0.816	0.187	0.214	0.570	0.513
✓		✓	✓	✓	0.080	0.720	0.833	0.757	0.260	0.123	0.466	0.425
✓	✓		✓	✓	0.089	0.685	0.823	0.756	0.177	0.233	0.556	0.524
✓	✓	✓		✓	0.061	0.769	0.893	0.815	0.311	0.152	0.460	0.424
✓	✓	✓	✓		0.054	0.740	0.884	0.798	**0.156**	0.220	0.565	0.517
✓	✓	✓	✓	✓	**0.044**	**0.790**	**0.899**	**0.833**	0.176	**0.243**	**0.583**	**0.530**

Table 6: Parameter ablation study on COD10K [14].

(a) Number of iteration **I**.

I	cos↑	IoU↑	$M \downarrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
1	0.864	0.563	0.080	0.626	0.818	0.765
2	0.876	0.589	0.050	0.683	0.859	0.796
3	0.879	0.593	0.045	0.702	0.869	0.802
4	**0.882**	0.601	0.042	0.714	0.875	**0.804**
5	0.881	**0.602**	**0.041**	0.718	0.875	**0.804**
6	**0.882**	0.599	**0.041**	**0.721**	**0.876**	0.803

(b) Image preprocess strategy.

Scale	$M \downarrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
Original	0.075	0.535	0.750	0.662
Havel	0.069	0.579	0.775	0.689
Quarters	0.087	0.423	0.673	0.586
Original+Havel	**0.042**	**0.714**	**0.875**	**0.804**
Original +Havel+Quarters	0.049	0.702	0.867	0.796

(c) Visual marker strategy.

strategy	$M \downarrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$
None	0.058	0.690	0.855	0.789
Bbox	0.065	0.682	0.836	0.766
VCD	0.047	0.705	0.863	0.793
Ours	**0.042**	**0.714**	**0.875**	**0.804**

Summary

- We explore how to utilize hallucinations as prior knowledge to assist task-generic promptable segmentation.

- We applied our approach in camouflaged animal detection, medical image segmentation, transparent object detection tasks, achieving promising results.

Code link: https://github.com/lwpyh/ProMaC_code

NEURAL INFORMATION PROCESSING SYSTEMS

Thank you!