Enhancing Prompt Generation with Adaptive Refinement for Camouflaged Object Detection

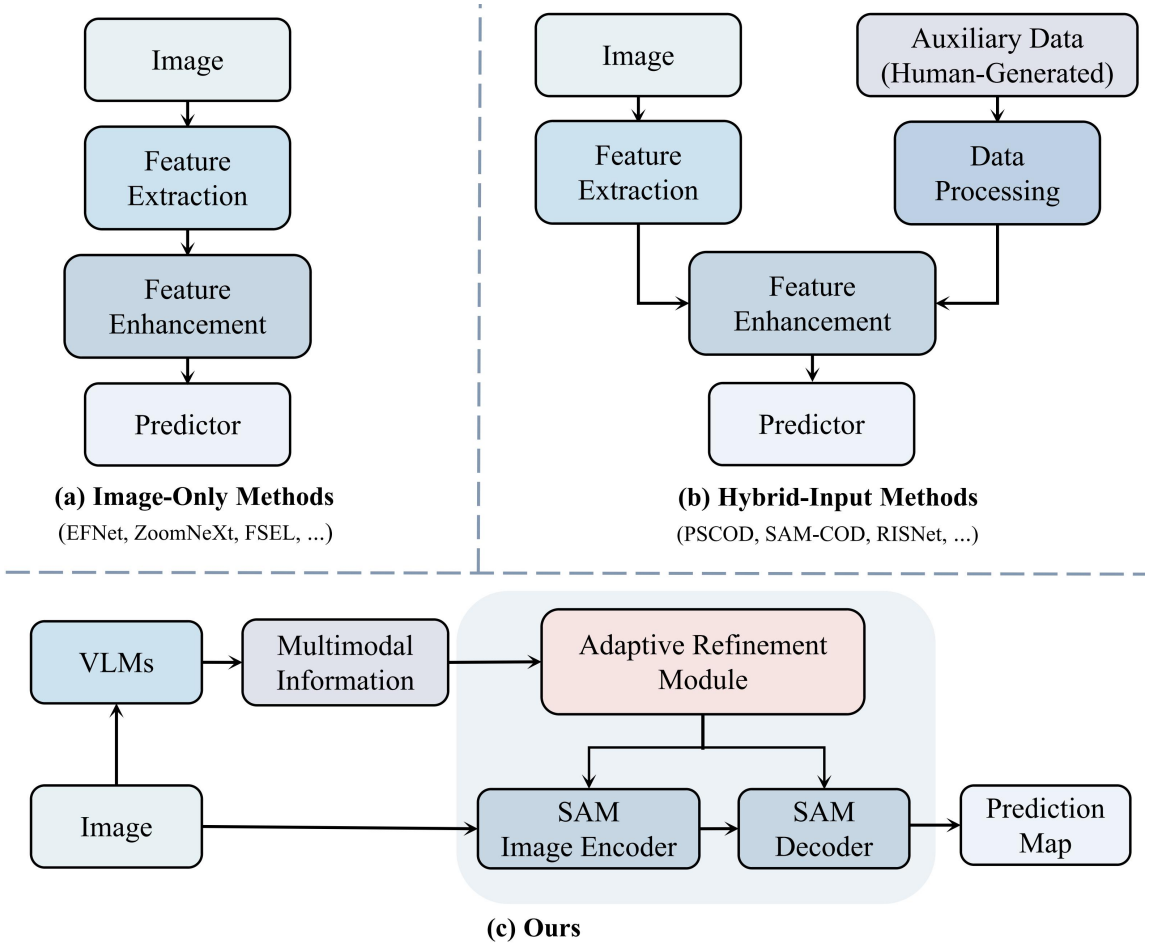
Xuehan Chen[1], Guangyu Ren[1*], Tianhong Dai[2], Tania Stathaki[2], Hengyan Liu[1*]
[1]Xi'an Jiaotong-Liverpool University, China
[2]Imperial College London, United Kingdom

IMPERIAL

Xi'an Jiaotong-Liverpool University
西交利物浦大學

ICCV
OCT 19-23, 2025
HONOLULU
HAWAII

Introduction

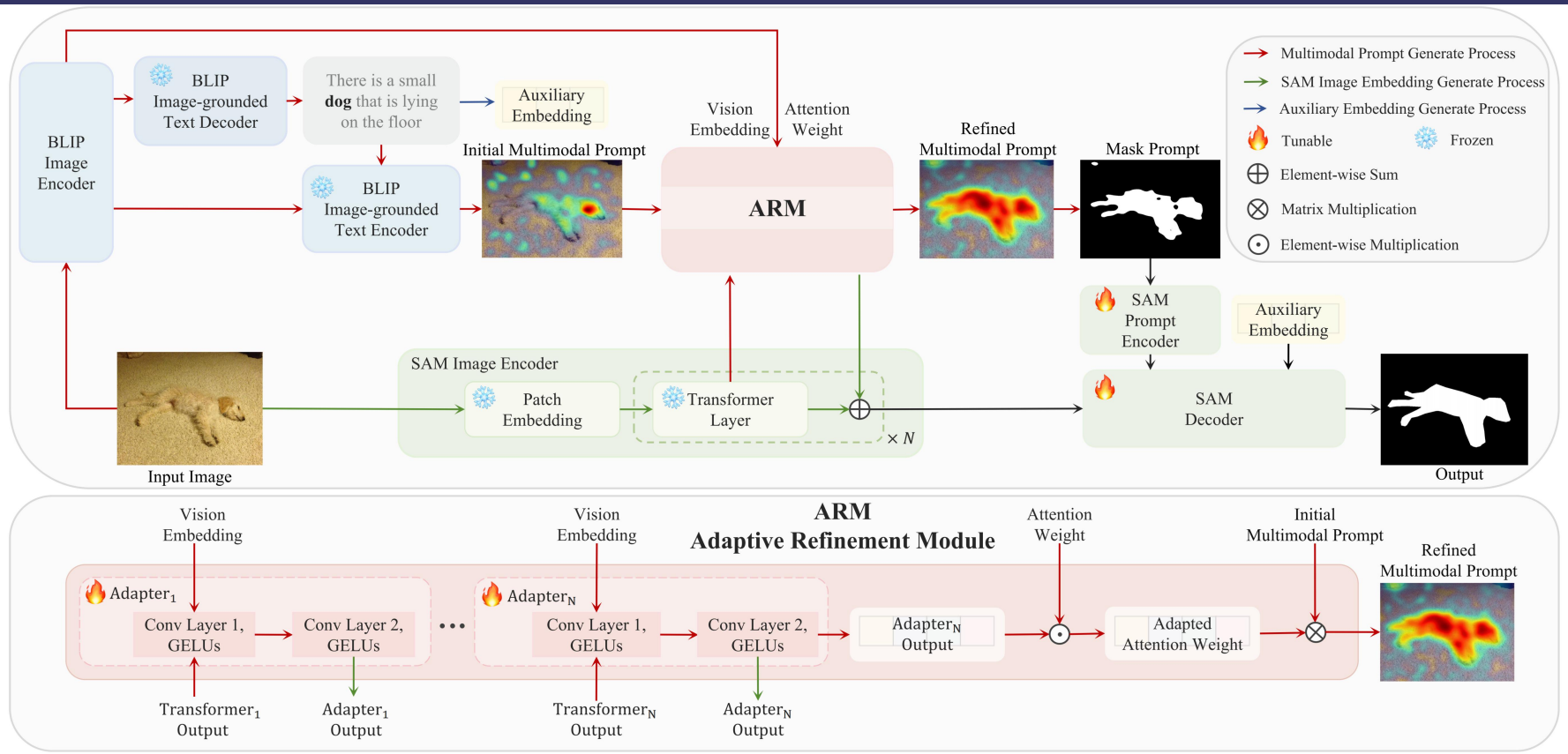


- Rethinking Camouflaged Object Detection (COD), we find that missing task-specific clues (e.g., category, size, location) and strong background noise lead to ambiguous boundaries and weak feature learning.
- Our approach leverages vision-language models (VLMs) to automatically generate multimodal data and introduces an Adaptive Refinement Module (ARM) to mitigate domain shift, providing high-quality prompts for SAM.

Contributions

- Propose a cost-effective method to automatically generate mask prompts from BLIP, providing effective guidance for SAM without extra training or fine-tuning.
- Design an Adaptive Refinement Module (ARM) to refine prompts and fine-tune SAM’s encoder, further enhancing SAM's segmentation accuracy and robustness.
- Introducing auxiliary embeddings from BLIP provides SAM with richer features.

Method



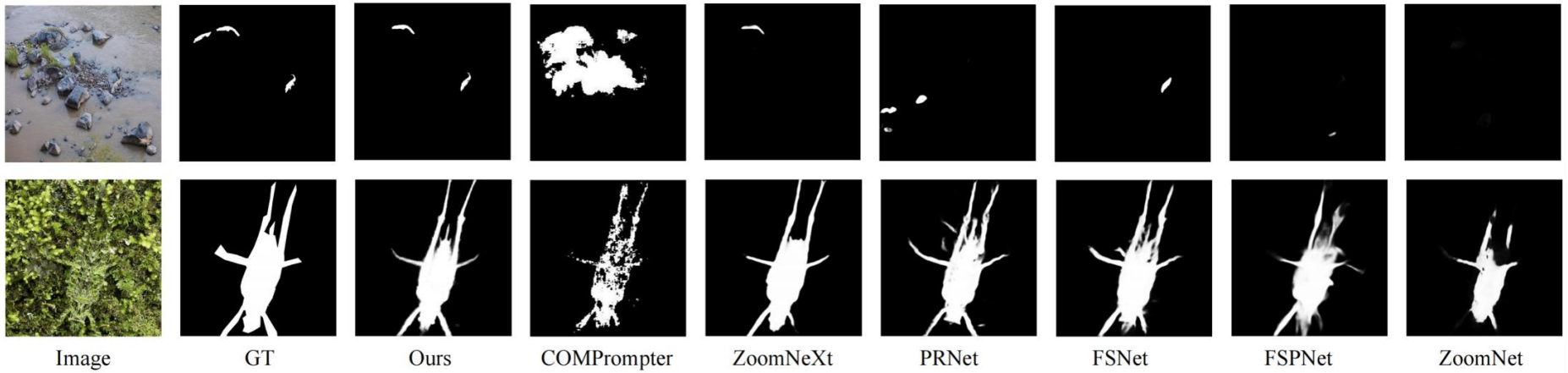
- BLIP automatically generates multimodal data from the input image and uses it to create the initial multimodal prompts and the auxiliary embedding.
- The Adaptive Refinement Module (ARM) addresses the locality issue of initial prompts by using multiple adapters to extract visual information from BLIP and SAM. It integrates attention weights to refine prompts into more accurate masks and simultaneously fine-tunes SAM's image encoder.
- Finally, the SAM decoder receives image embedding, dense mask prompt embedding, and auxiliary sparse embedding to tackle the COD task.

Experimental Results

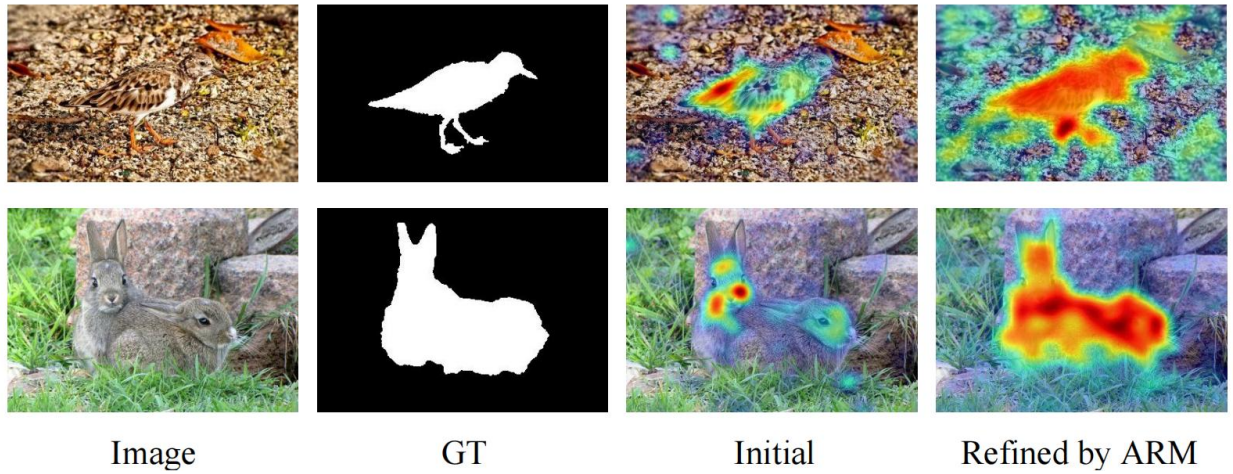
Method	Pub.Year	CHAMELEON				CAMO				COD10K				NC4K			
		Sm	F_β^{max}	E_φ^{max}	M	Sm	F_β^{max}	E_φ^{max}	M	Sm	F_β^{max}	E_φ^{max}	M	Sm	F_β^{max}	E_φ^{max}	M
ZoomNet [42]	$CVPR_{22}$	0.901	0.876	0.947	0.023	0.820	0.805	0.889	0.066	0.837	0.777	0.896	0.029	0.852	0.826	0.903	0.044
SegMaR [25]	$CVPR_{22}$	0.906	0.888	0.959	0.025	0.816	0.803	0.884	0.071	0.833	0.774	0.906	0.034	0.841	0.826	0.907	0.046
UEDG [39]	TMM_{23}	0.911	0.894	0.968	0.023	0.863	0.856	0.929	0.048	0.858	0.812	0.934	0.025	0.879	0.864	0.935	0.035
TinyCOD [58]	$ICASSP_{23}$	0.887	0.861	0.958	0.030	0.822	0.807	0.899	0.066	0.811	0.742	0.903	0.036	0.843	0.817	0.910	0.047
MSCAF-Net [36]	$TCSVT_{23}$	0.912	0.902	0.970	0.022	0.873	0.867	0.937	0.046	0.865	0.823	0.936	0.024	0.887	0.874	0.942	0.032
FSPNet [24]	$CVPR_{23}$	0.909	0.890	0.965	0.023	0.856	0.846	0.928	0.050	0.851	0.794	0.930	0.026	0.879	0.859	0.937	0.035
FEDER [16]	$CVPR_{23}$	0.887	0.868	0.954	0.030	0.802	0.789	0.873	0.071	0.822	0.768	0.905	0.032	0.847	0.833	0.915	0.044
FSNet [49]	TIP_{23}	0.905	0.891	0.975	0.022	0.880	0.878	0.941	**0.041**	0.870	0.833	0.948	0.023	0.891	0.880	0.948	0.031
CRNet [19]	$AAAI_{23}$	0.818	0.792	0.909	0.046	0.735	0.707	0.830	0.092	0.733	0.636	0.845	0.049	—	—	—	—
SDRNet [13]	KBS_{24}	0.914	0.901	0.961	0.024	0.872	0.867	0.932	0.049	0.871	0.828	0.936	0.023	0.889	0.876	0.940	0.032
PRNet [23]	$TCSVT_{24}$	0.915	0.902	0.973	0.020	0.872	0.867	0.930	0.050	0.873	0.839	0.943	0.022	0.891	0.881	0.942	0.031
CamoFormer-R [59]	$TPAMI_{24}$	0.898	0.880	0.949	0.026	0.817	0.801	0.883	0.068	0.838	0.786	0.928	0.029	0.854	0.829	0.908	0.042
CamoFormer-P [59]	$TPAMI_{24}$	0.910	0.898	0.966	0.022	0.872	0.868	0.938	0.046	0.869	0.829	0.939	0.023	0.892	0.880	0.946	0.030
$ZoomNeXt_{pvtv2b3}$ [43]	$TPAMI_{24}$	0.928	0.919	**0.977**	**0.017**	0.885	**0.886**	**0.942**	0.042	0.895	0.864	0.951	**0.018**	0.900	0.891	**0.948**	**0.028**
$ZoomNeXt_{pvtv2b2}$ [43]	$TPAMI_{24}$	0.922	0.908	0.969	0.017	0.874	0.873	0.931	0.047	0.887	0.856	0.945	0.019	0.892	0.884	0.941	0.030
DSAM [60]	$ACM\ MM_{24}$	—	—	—	—	0.832	0.834	0.920	0.061	0.845	0.805	0.930	0.033	0.872	0.864	0.942	0.040
MAMIFNet [55]	IF_{25}	0.914	0.899	0.959	0.021	0.872	0.870	0.935	0.045	0.869	0.826	0.940	0.023	0.890	0.878	0.943	0.031
COMPrompter [64]	$SCIS_{25}$	0.885	0.864	0.957	0.030	0.853	0.856	0.931	0.054	0.860	0.826	0.946	0.027	0.880	0.876	0.946	0.036
Ours	—	**0.932**	**0.922**	0.964	0.023	**0.887**	0.883	0.935	0.046	**0.909**	**0.885**	**0.955**	0.020	**0.906**	**0.897**	0.947	0.033

Our method achieves outstanding performance, particularly excelling in Sm and F_β^{max}. This demonstrates its superior target localization and structure capture of camouflaged objects.

Experimental Results



Our approach achieves a more significant overlap with the GT than other SOTA models, effectively segmenting completely camouflaged objects.



We also visualize both the initial and ARM-refined prompts. Initial prompts often cover only partial target regions due to domain shifts, while ARM-refined prompts achieve more complete coverage.