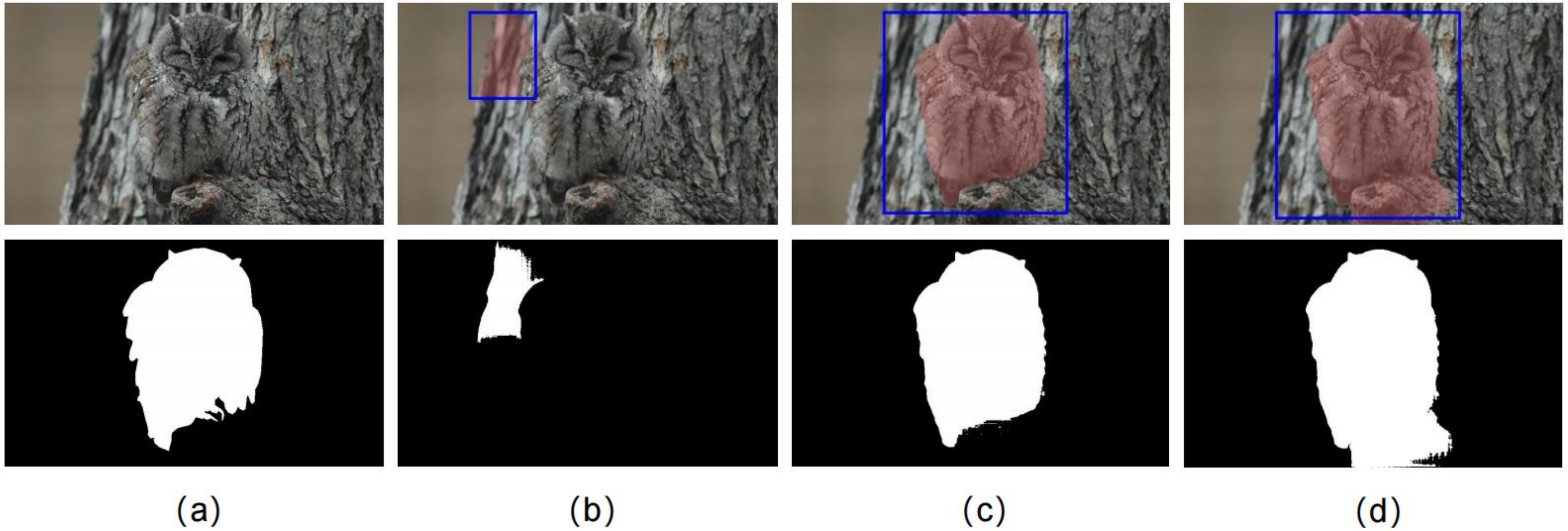
# Multi-modal Segment Anything Model for Camouflaged Scene Segmentation

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

IMPERIAL

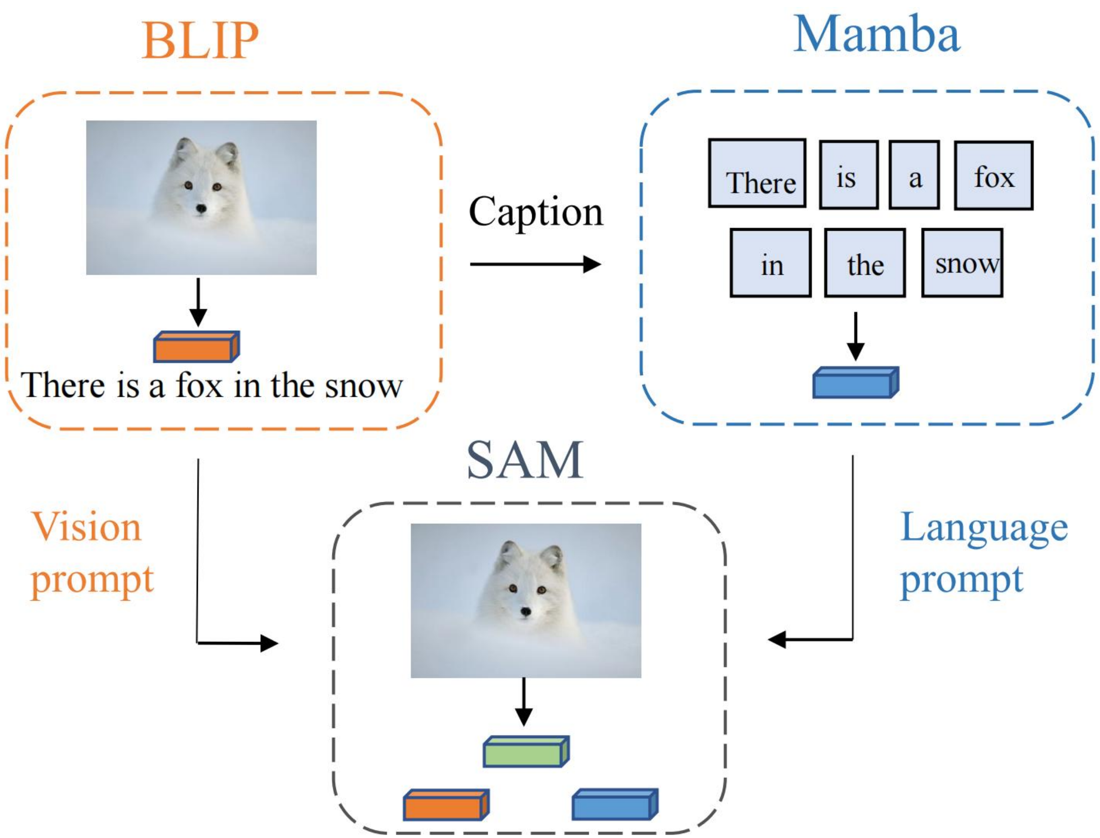
Xi'an Jiaotong-Liverpool University
西交利物浦大學

ICCV
OCT 19-23, 2025
HONOLULU
HAWAII



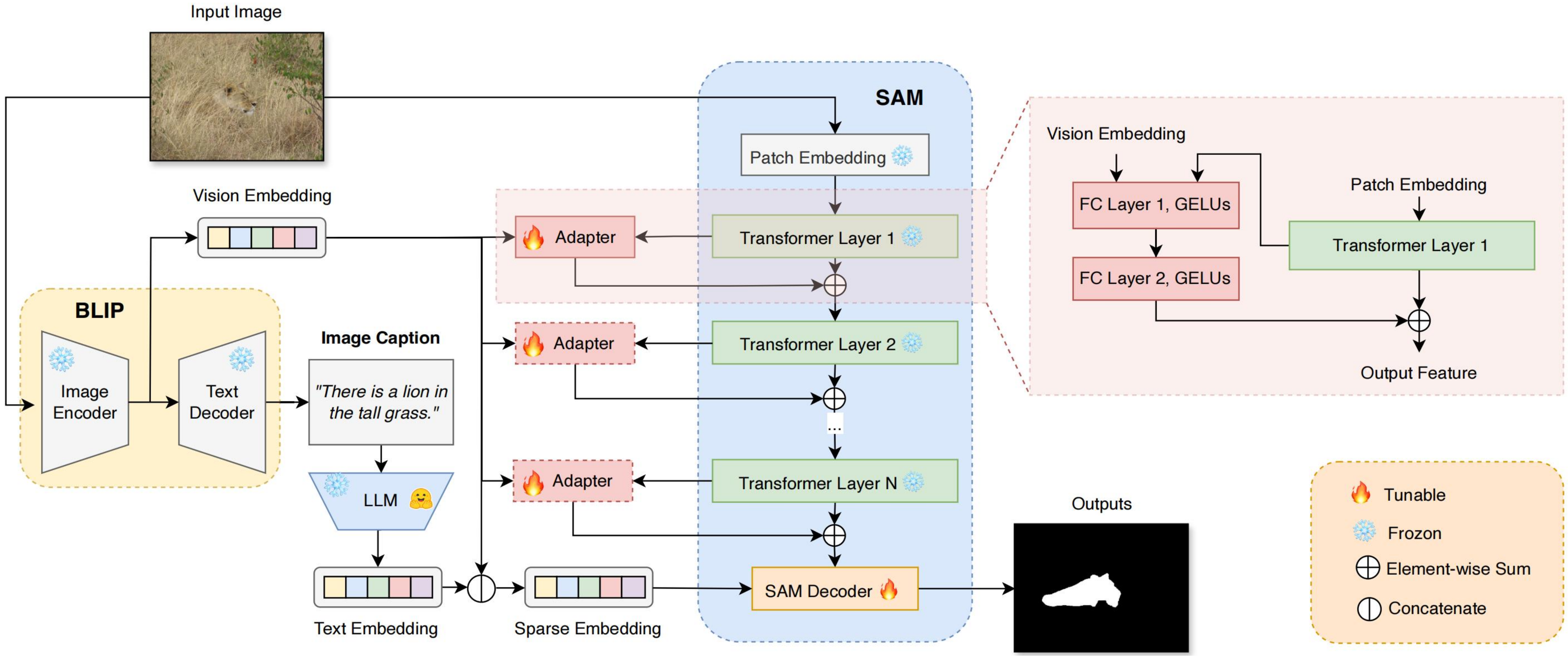

Visualization of segmentation results generated by SAM under different bounding-box prompts.

- Despite their success in many vision tasks, deep learning-based methods struggle with COD due to the limited size of available datasets, since insufficient data hinders feature learning and expanding datasets requires costly human annotation.
- Even SAM struggle with COD due to the similarity between background and foreground, as well as their dependence on manual sparse prompts, which are highly error-prone.




- We use BLIP to generate captions and exploit text embeddings for COD.
- To the best of our knowledge, we are the first to combine vision and language information to enhance SAM in COD.
- We integrate a multi-level adapter and a dense embedding based on SAM's image embedding.
- Our method achieves state-of-the-art results on 11/12 metrics across three COD benchmarks.

# Methodology



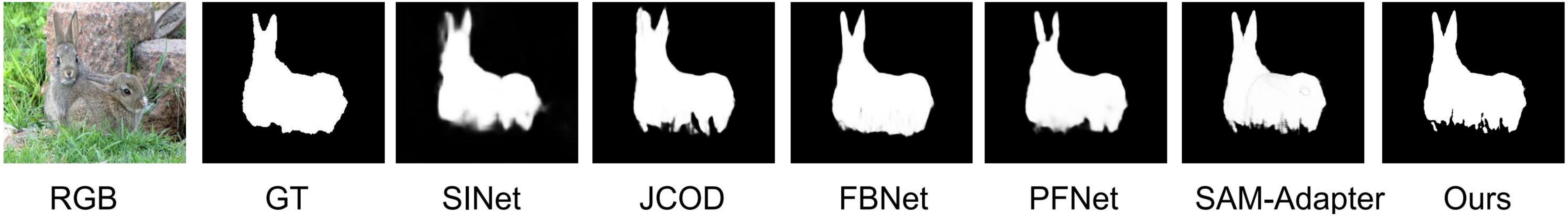

- BLIP Image encoder is used to obtain the vision embedding.
- BLIP's text decoder and the Mamba text encoder are used to obtain the text embedding.
- The Vision embedding is incorporated into SAM using a multi-level adapter.
- The text embedding and the vision embedding are concatenated to form sparse embedding.
- The sparse embedding is concatenated with SAM's image embedding to SAM's decoder.
- SAM's decoder outputs the predicted segmentation mask.

# Main Experimental Results

| Method | Venue | CVC-ColonDB | | | | Kvasir | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $S_\alpha \uparrow$ | $E_\phi \uparrow$ | $F_\beta^w \uparrow$ | $Mae \downarrow$ | $S_\alpha \uparrow$ | $E_\phi \uparrow$ | $F_\beta^w \uparrow$ | $Mae \downarrow$ |
| GPT4V+SAM [17] | Arxiv23 | 0.242 | 0.246 | 0.051 | 0.578 | 0.253 | 0.236 | 0.128 | 0.614 |
| LLaVA1.5+SAM [17, 24] | NeruIPS23 | 0.357 | 0.355 | 0.194 | 0.491 | 0.403 | 0.400 | 0.293 | 0.479 |
| X-Decoder [51] | CVPR23 | 0.331 | 0.327 | 0.095 | 0.462 | 0.384 | 0.371 | 0.202 | 0.449 |
| SEEM [52] | NeruIPS23 | 0.284 | 0.280 | 0.085 | 0.570 | 0.367 | 0.337 | 0.215 | 0.520 |
| GroundingSAM [17, 27] | ICCV23 | 0.206 | 0.195 | 0.071 | 0.711 | 0.468 | 0.521 | 0.353 | 0.387 |
| GenSAM [10] | AAAI24 | 0.379 | 0.494 | 0.059 | 0.244 | 0.487 | 0.619 | 0.210 | 0.172 |
| ProMaC [11] | NeruIPS24 | 0.530 | **0.583** | **0.243** | **0.176** | 0.573 | 0.726 | 0.394 | 0.166 |
| **MM-SAM**(Training-Free) | Ours | 0.451 | 0.527 | 0.107 | 0.317 | 0.475 | 0.637 | 0.159 | 0.235 |
| **MM-SAM**(Zero-Shot) | Ours | **0.565** | 0.520 | 0.220 | 0.185 | **0.740** | **0.756** | **0.535** | **0.134** |

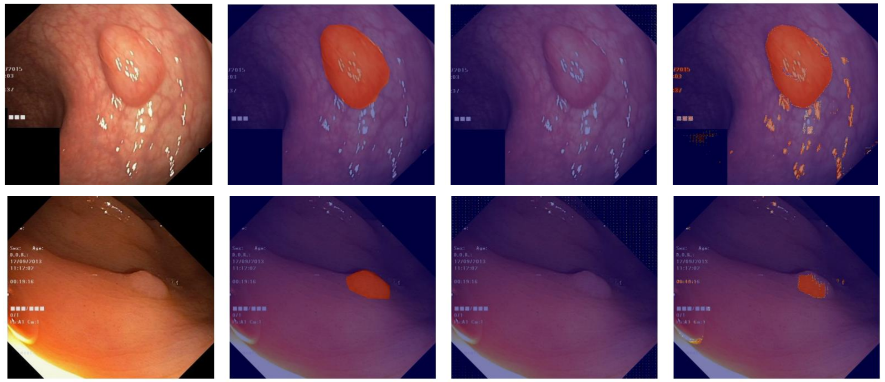
The proposed method outperforms existing algorithms on all datasets and metrics, demonstrating superior robustness, precision, and error reduction in camouflage detection.

# Main Experimental Results




Our approach successfully segments challenging samples where other methods struggle, and unlike the ground truth, it also captures fine details such as overlapping grass and rabbits.

| Method | Venue | CVC-ColonDB | | | | Kvasir | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $S_\alpha \uparrow$ | $E_\phi \uparrow$ | $F_\beta^w \uparrow$ | $Mae \downarrow$ | $S_\alpha \uparrow$ | $E_\phi \uparrow$ | $F_\beta^w \uparrow$ | $Mae \downarrow$ |
| GPT4V+SAM [17] | Arxiv23 | 0.242 | 0.246 | 0.051 | 0.578 | 0.253 | 0.236 | 0.128 | 0.614 |
| LLaVA1.5+SAM [17, 24] | NeruIPS23 | 0.357 | 0.355 | 0.194 | 0.491 | 0.403 | 0.400 | 0.293 | 0.479 |
| X-Decoder [51] | CVPR23 | 0.331 | 0.327 | 0.095 | 0.462 | 0.384 | 0.371 | 0.202 | 0.449 |
| SEEM [52] | NeruIPS23 | 0.284 | 0.280 | 0.085 | 0.570 | 0.367 | 0.337 | 0.215 | 0.520 |
| GroundingSAM [17, 27] | ICCV23 | 0.206 | 0.195 | 0.071 | 0.711 | 0.468 | 0.521 | 0.353 | 0.387 |
| GenSAM [10] | AAAI24 | 0.379 | 0.494 | 0.059 | 0.244 | 0.487 | 0.619 | 0.210 | 0.172 |
| ProMaC [11] | NeruIPS24 | 0.530 | **0.583** | **0.243** | **0.176** | 0.573 | 0.726 | 0.394 | 0.166 |
| **MM-SAM**(Training-Free) | Ours | 0.451 | 0.527 | 0.107 | 0.317 | 0.475 | 0.637 | 0.159 | 0.235 |
| **MM-SAM**(Zero-Shot) | Ours | **0.565** | 0.520 | 0.220 | 0.185 | **0.740** | **0.756** | **0.535** | **0.134** |



RGB GT SAM Training-free

- Training-free: Achieves competitive results using foundation model embeddings without fine-tuning.
- Prompting: Foundation models directly provide effective multi-modal prompts for SAM.
- Generalization: Fine-tuned on COD, the model performs strongly on polyp segmentation.

# Zero-shot Transfer

| Method | DUTS | | CUHK | |
|---|---|---|---|---|
| | $E_\phi \uparrow$ | $Mae \downarrow$ | $mIoU \uparrow$ | $F_\beta \uparrow$ |
| SAMed [48] | 0.764 | 0.104 | 0.554 | 0.717 |
| SEEM [52] | 0.599 | 0.326 | 0.608 | 0.675 |
| Painter [45] | 0.811 | 0.113 | 0.186 | 0.273 |
| PerSAM [49] | 0.641 | 0.257 | 0.558 | 0.716 |
| AlignSAM [13] | 0.782 | 0.082 | 0.685 | **0.769** |
| **MM-SAM** | **0.877** | **0.058** | **0.698** | 0.746 |

Without fine-tuning, it achieves superior performance on saliency and blur detection task, demonstrating strong generalization across tasks.

IMPERIAL

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

ICCV OCT 19-23, 2025 HONOLULU HAWAII

# Text Encoder Analysis

| Method | CAMO | | | | COD10K | | | |
|---|---|---|---|---|---|---|---|---|
| | $S_\alpha \uparrow$ | $E_\phi \uparrow$ | $F_\beta \uparrow$ | $Mae \downarrow$ | $S_\alpha \uparrow$ | $E_\phi \uparrow$ | $F_\beta \uparrow$ | $Mae \downarrow$ |
| SAM | 0.790 | 0.839 | 0.619 | 0.107 | 0.807 | 0.801 | 0.606 | 0.054 |
| SAM+CLIP | 0.779 | 0.848 | 0.629 | 0.101 | 0.812 | 0.816 | 0.631 | 0.047 |
| SAM+Bert | 0.782 | 0.854 | 0.634 | 0.099 | 0.812 | 0.823 | 0.631 | 0.045 |
| SAM+Llama | 0.774 | 0.852 | 0.620 | 0.102 | 0.806 | 0.826 | 0.626 | 0.047 |
| SAM+Mamba | 0.784 | 0.848 | 0.630 | 0.101 | 0.818 | 0.833 | 0.651 | 0.046 |

- The SAM baseline includes only the image encoder and mask decoder.
- Compared to CLIP, Mamba achieves more robust COD fine-tuning, while other LLMs provide similar improvements, likely due to BLIP's simple captions.