# **CamSAM2**: Segment Anything Accurately in Camouflaged Videos

Yuli Zhou[1,3] Yawei Li[1] Yuqian Fu[4] Luca Benini[1,5] Ender Konukoglu[1] Guolei Sun[2*]

[1]ETH Zurich [2]Nankai University [3]University of Zurich

[4]INSAIT, Sofia University “St. Kliment Ohridski” [5]University of Bologna

# Task definition:

## Video Camouflaged Object Segmentation (VCOS)

- **Challenge**: **SAM2** achieves strong video segmentation, but struggle in camouflage cases due to:
  - SAM2 is optimized for natural scenes rather than **camouflaged** environments.
  - The architecture does not account for the complexities of segmenting and tracking camouflaged objects across time.
- **Motivation**:
  - Can we adapt SAM2 for accurate segmentation in camouflaged videos without breaking its general zero-shot capability?

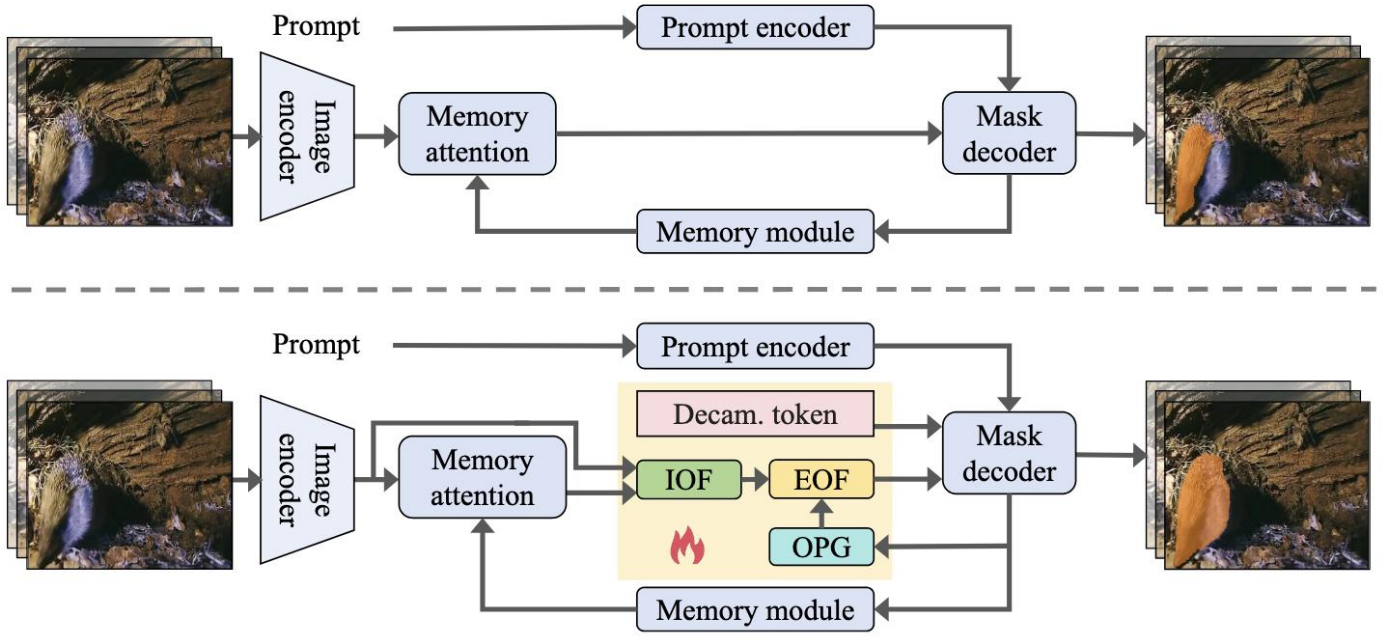



Figure 1: **Illustration of SAM2 and CamSAM2**. *Top:* SAM2's segmentation of the camouflaged object is suboptimal, primarily because its feature optimization is biased toward natural videos, and its design does not account for the unique challenges inherent to VCOS. *Bottom:* CamSAM2 improves SAM2's ability to segment and track camouflaged objects by introducing a *decamouflaged token*, *IOF* to enhance features with high-resolution features, and *EOF* and *OPG* to further enhance features by exploiting informative object details across time. CamSAM2 only adds a limited number of parameters to SAM2 while keeping all SAM2's parameters fixed and fully inheriting SAM2's zero-shot ability. The segmentation result is overlaid in orange on the frame.

# Method Overview: **CamSAM2**

- **Core ideas**:
  - Decamouflaged token
  - Implicit and Explicit Object-aware Fusion (IOF & EOF)
  - Object Prototype Generation (OPG)
- **Without** modifying SAM2's main parameters
- Only ~0.5M parameters added

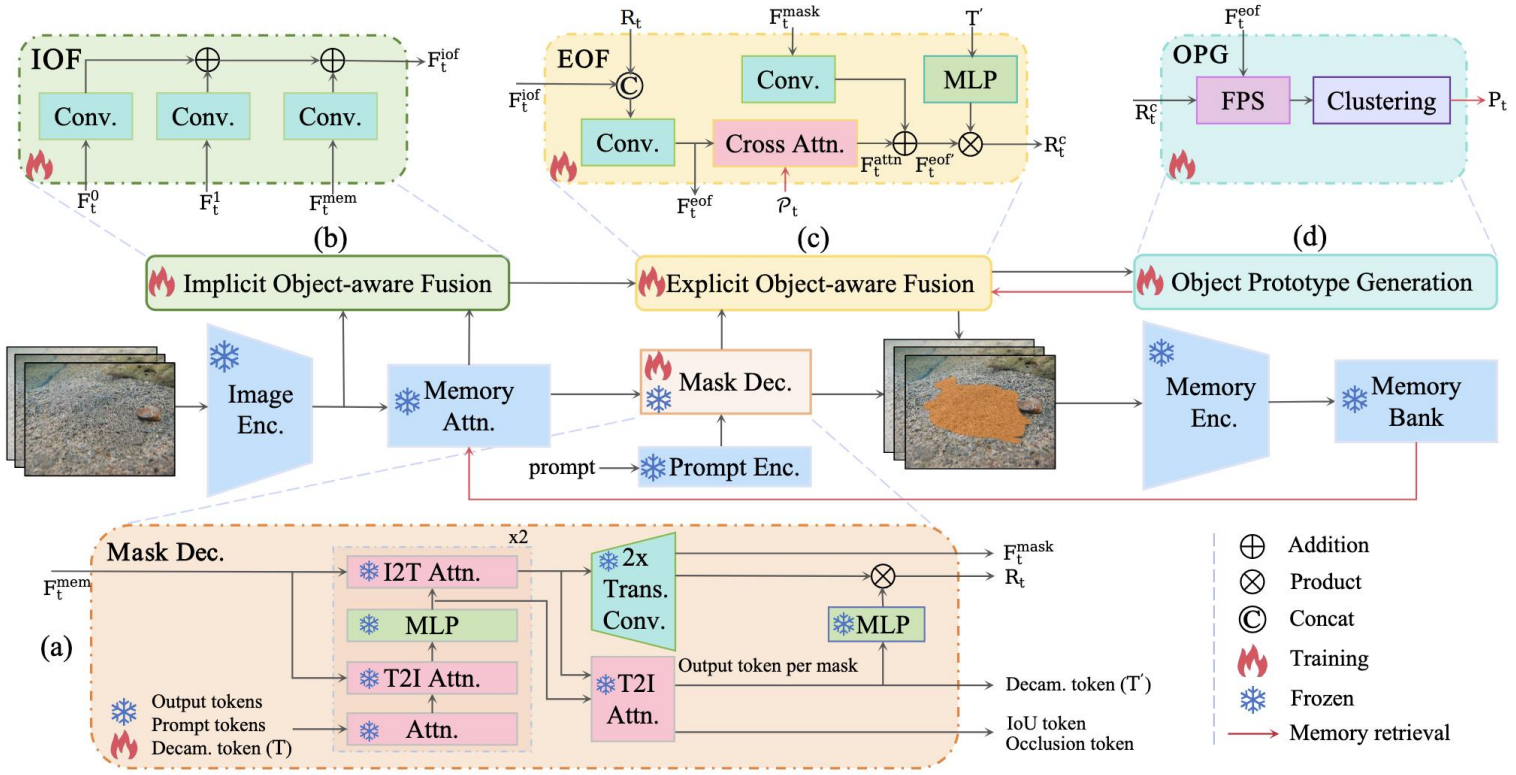



Figure 2: **Overall architecture of CamSAM2**. CamSAM2 effectively captures and segments camouflaged objects by leveraging implicit and explicit object-aware information from the current or previous frames. It includes the following key components: *(a)* the *decamouflaged token*, which extends SAM2's token structure to learn features suitable for camouflaged objects; *(b)* an *IOF* module to enrich memory-conditioned features with implicitly object-aware high-resolution features; *(c)* an *EOF* module to aggregate explicit object-aware features; and *(d)* an *OPG* module, generating informative object prototypes, which guides cross-attention in EOF. These components work together to preserve fine details, enhance segmentation quality, and track camouflaged objects across time.

# Experiments & Key Results:

- Datasets:
  - Wildlife animal camouflage: MoCA-Mask, CAD
  - Medical polyp camouflage: SUN-SEG
- Results:
  - MoCA-Mask

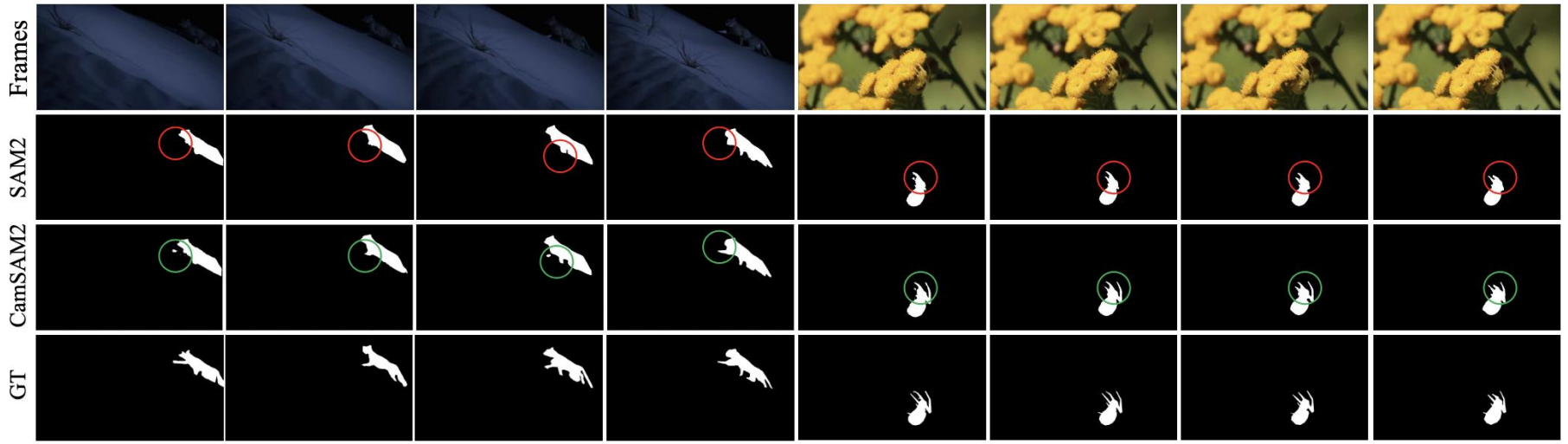
| Model | Prompt | Hiera-T | | Hiera-S | |
|---|---|---|---|---|---|
| | | mDice ↑ | mIoU ↑ | mDice ↑ | mIoU ↑ |
| **SAM2** | 1-click | 52.1 | 44.8 | 54.9 | 46.7 |
| **CamSAM2** | | **64.3** (+12.2) | **54.6** (+9.8) | **68.0** (+13.1) | **58.8** (+12.1) |
| **SAM2** | box | 72.7 | 62.3 | 73.7 | 63.8 |
| **CamSAM2** | | **75.5** (+2.8) | **64.8** (+2.5) | **76.4** (+2.7) | **66.1** (+2.3) |
| **SAM2** | mask | 77.1 | 67.9 | 80.3 | 70.7 |
| **CamSAM2** | | **80.2** (+3.1) | **70.5** (+2.6) | **81.4** (+1.1) | **71.7** (+1.0) |

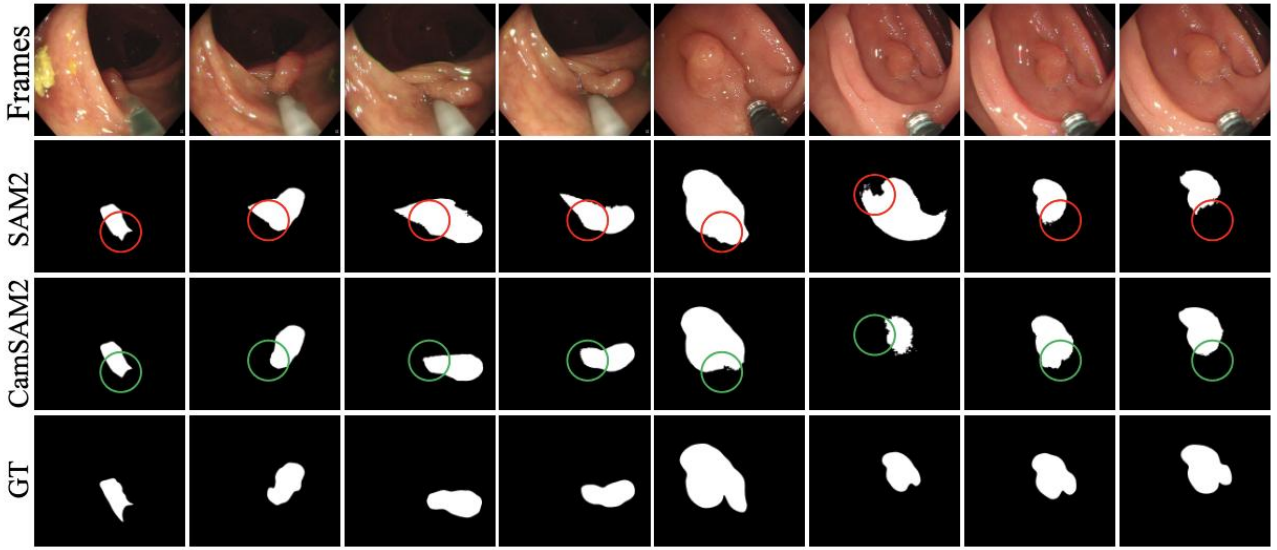
SUN-SEG

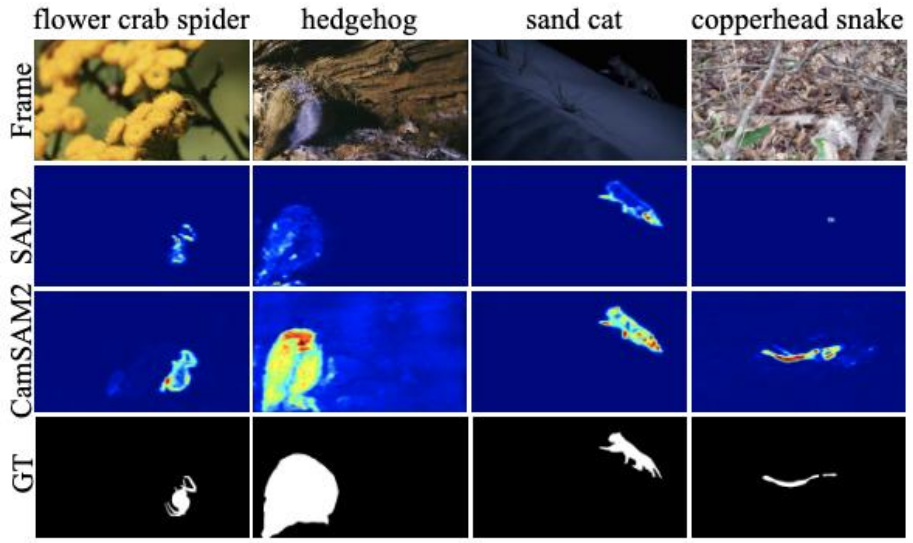
| Model | $S_m$ ↑ | $F_\beta^\omega$ ↑ | $E_m$ ↑ | mDice ↑ |
|---|---|---|---|---|
| SUN-SEG-Easy | | | | |
| SAM2 [15] | 83.4 | 71.6 | 83.0 | 73.6 |
| CamSAM2 | **88.3** | **82.6** | **93.4** | **84.3** |
| SUN-SEG-Hard | | | | |
| SAM2 [15] | 75.5 | 58.4 | 73.4 | 61.0 |
| CamSAM2 | **86.4** | **78.2** | **91.2** | **80.6** |

# Visualizations

- Top-right:
  - MoCA-Mask
- Bottom-left:
  - SUN-SEG
- Bottom-right:
  - Attention maps of SAM2 token and the decamouflaged token



# Takeaways

- CamSAM2 extends SAM2 to handle camouflaged video segmentation effectively.
- The **Decamouflaged Token**, **Implicit** and **Explicit Object-aware Fusion**, and **Object Prototype Generation together** improve both spatial accuracy and temporal consistency.
- Achieves SOTA performance while keeping SAM2's parameters unchanged and general-purpose.


- Contact:
  - Yuli Zhou: zhoustan98@gmail.com
  - Guolei Sun: guolei.sun@nankai.edu.cn