

# When SAM2 Meets Video Camouflaged Object Segmentation: A Comprehensive Evaluation and Adaptation

Yuli Zhou<sup>1,2</sup>  · Guolei Sun<sup>1\*</sup>  · Yawei Li<sup>1,3</sup>  · Guo-Sen Xie<sup>4</sup>  · Luca Benini<sup>3,5</sup>  · Ender Konukoglu<sup>1</sup> 

Received: date / Accepted: date

**Abstract** This study investigates the application and performance of the Segment Anything Model 2 (SAM2) in the challenging task of video camouflaged object segmentation (VCOS). VCOS involves detecting objects that blend seamlessly in the surroundings for videos due to similar colors and textures and poor light conditions. Compared to the objects in normal scenes, camouflaged objects are much more difficult to detect. SAM2, a video foundation model, has shown potential in various tasks. However, its effectiveness in dynamic camouflaged scenarios remains under-explored. This study presents a comprehensive study on SAM2’s ability in VCOS. First, we assess SAM2’s performance on camouflaged video datasets using different models and prompts (click, box, and mask). Second, we explore the integration of SAM2 with existing multimodal large language models (MLLMs) and VCOS methods. Third, we specifically adapt SAM2 by fine-tuning it on the video camouflaged dataset. Our comprehensive experiments demonstrate that SAM2 has the excellent zero-shot ability to detect camouflaged objects in videos. We also show that this ability could be further improved by specifically adjusting SAM2’s parameters for VCOS.

---

Yuli Zhou, Guolei Sun, Yawei Li, and Ender Konukoglu are with Computer Vision Laboratory, ETH Zürich, Zürich, Switzerland. (Email: yulzhou@vision.ee.ethz.ch, guolei.sun@vision.ee.ethz.ch, yawei.li@vision.ee.ethz.ch, ender.konukoglu@vision.ee.ethz.ch).

Yawei Li and Luca Benini are with Integrated System Laboratory, ETH Zürich, Zürich, Switzerland. (Email: yawei.li@vision.ee.ethz.ch, lbenini@iis.ee.ethz.ch).

Guo-Sen Xie is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. (Email: gsxiehm@gmail.com).

Corresponding author: Guolei Sun  
(guolei.sun@vision.ee.ethz.ch).

**Keywords** Multimodal large language model · Prompt engineering · SAM2 · Video camouflaged object segmentation

## 1 Introduction

Camouflaged object segmentation, aiming to identify objects that blend into their surroundings, is a fundamental task in computer vision. Unlike conventional segmentation tasks, where objects typically exhibit distinct boundaries, camouflaged objects often have similar colors or textures to the background, making them difficult to be perceived. This task becomes even more complex for video sequences, where both objects and background can change dynamically over time.

Traditional segmentation or detection approaches [1–4], such as fully convolutional networks (FCNs) [1] and mask R-CNN [2], have made significant contributions to the field of object detection and segmentation in common scenes where objects are generally distinguishable from the environment. However, these models often struggle with camouflaged scenes. More recent techniques such as SINet [5], SLT-Net [6], and ZoomNeXt [7] have been specifically designed to address the challenges of camouflaged object detection. Despite these advancements, existing video camouflaged object segmentation (VCOS) methods still face significant challenges in maintaining temporal consistency, as they often lack effective object-level memory, making it difficult to achieve stable segmentation across long video sequences. Some approaches, such as Flow-SAM [8] and SLT-Net [6], incorporate flow-guided or motion-guided networks to enhance temporal coherence. However, these methods struggle when camouflaged objects remain stationary or exhibit minimal movement, leading to segmentation

failures. Additionally, they are prone to errors in dynamic backgrounds, where optical flow may incorrectly associate background motion with the target object, further compromising segmentation accuracy.

Recently, the advanced large foundation model Segment Anything Model 2 (SAM2) [9] has shown promise in handling complex video-based segmentation tasks. It adopts prompt-based segmentation, where specific user-defined prompts such as bounding boxes, clicks, or masks are used to guide the segmentation process. This allows for great flexibility in segmenting target objects, particularly in scenarios like camouflaged detection, where object boundaries are usually invisible. Based on Segment Anything Model (SAM) [10], SAM2 is equipped with a memory module containing: a memory encoder, which encodes the predicted mask and image embedding of the current frame; a memory bank, which stores information about the object and previous interactions; and memory attention, which enhances the current frame's image embedding using past memories. This approach allows SAM2 to track camouflaged objects more reliably across video sequences.

Given SAM2's strong generalization ability and its flexible prompting mechanisms, we systematically investigate its effectiveness in addressing VCOS through three key aspects: 1) zero-shot capability, by directly adapting SAM2 on two camouflaged datasets, MoCA-Mask [6] and CAD [11]; 2) integration with state-of-the-art methods, by leveraging multimodal large language models (MLLMs) and existing VCOS techniques with SAM2, exploring hybrid approaches for VCOS; and 3) adaptation through task-specific fine-tuning on the largest-scale VCOS benchmark MoCA-Mask. Our study is organized into three parts, with each targeting a specific aspect of this complex problem:

**1) Evaluating SAM2's zero-shot ability for VCOS.** We thoroughly assess the performance of SAM2 in segmenting camouflaged objects, an inherently challenging task due to the high degree of similarity between the object and its background. SAM2 is evaluated in two modes: *automatic* and *semi-supervised* mode. In automatic mode, we leverage SAM2's built-in automatic mask generator to produce initial segmentation masks, which are then used as mask prompts for subsequent frames. In semi-supervised mode, we investigate how manual interactions (via click, box, and mask-based prompts) and prompting time can impact segmentation quality. This analysis provides the first detailed exploration of SAM2's behavior specifically for VCOS, a domain where traditional segmentation models typically struggle.

**2) Augmenting SAM2 with state-of-the-art MLLMs and VCOS methods.** We further explore the effective-

ness of SAM2 when combined with advanced MLLMs and VCOS techniques. For MLLMs, we design specific prompts to generate bounding boxes around potential camouflaged regions, which are then used by SAM2 as box prompts for segmentation. For VCOS, we enhance segmentation by refining the initial mask generated by VCOS methods with SAM2's powerful refinement capabilities. Our contribution here is the novel integration of SAM2 with MLLMs and VCOS methods, demonstrating how prompt-driven refinement can improve segmentation accuracy in highly complex visual scenes.

**3) Fine-tuning SAM2 on largest-scale VCOS benchmark.** We explore how task-specific fine-tuning can improve SAM2's segmentation performance in VCOS. By fine-tuning SAM2 on the MoCA-Mask dataset, we extend SAM2's utility beyond its initial design, showing how it can be adapted to specialized datasets for enhanced performance.

Our contributions can be summarized as follows:

- 1) We provide the first thorough evaluation and in-depth analysis of SAM2's performance in the challenging task of VCOS, experimenting with both automatic and semi-supervised modes.
- 2) We propose a novel hybrid approach that combines SAM2 with existing MLLMs and VCOS methods, demonstrating significant improvements in segmentation accuracy through prompt-based refinements.
- 3) We enhance SAM2's capabilities by adjusting its parameters on the well-known MoCA-Mask dataset to better fit VCOS, achieving state-of-the-art results.

## 2 Related Work

### 2.1 Camouflaged Object Detection

Camouflaged scene understanding (CSU) aims to perceive the scenes where objects are difficult to be distinguished with the background. These scenes commonly exist in natural environments, such as forests, oceans, and deserts. Among various CSU tasks [5, 12–15], camouflaged object detection (COD) attracts lots of research attention, which is to identify objects that blend seamlessly into their environment, posing a significant challenge for conventional perception techniques. Traditional approaches to object detection, which rely on strong edge features, color contrasts, and texture variations, often fail when applied to camouflaged objects. This led to the development of specialized COD models that incorporate unique feature extraction techniques, focusing on small differences in texture, edge irregularities, and subtle shifts in color or shading that indicate the presence of a hidden object [5, 16]. Several innovative

models have been proposed for COD. HGINet [17] uses a hierarchical graph interaction transformer with dynamic token clustering to capture multi-scale features for better object detection. AGLNet [18] enhances COD performance by employing adaptive guidance learning, which adjusts feature extraction based on the object’s appearance and context. PAD [19] leverages a multi-task learning approach, pre-training on multiple datasets and fine-tuning for COD tasks through task-specific adaptation. DQNet [20] focuses on cross-model detail querying, using multiple models to enhance the detection of subtle features in camouflaged objects. R2CNet [21] integrates linguistic cues with visual data to improve detection, particularly by using referring expressions to locate specific camouflaged objects. ZoomNeXt [7] presents a unified collaborative pyramid network to enhance feature extraction across multiple scales. WSSCOD [22] improves COD using weak supervision by learning from noisy pseudo labels, while CamoTeacher [23] applies dual-rotation consistency learning for semi-supervised detection, making use of limited labeled data. POCINet [24] integrates contrast information and part-object relational knowledge across search and identification stages, bridged by a search-to-identification guidance (SIG) module that enhances feature encoding through search results and semantic knowledge. STANet [25] introduces scale-feature attention and type-feature attention to enhance COD by integrating multi-scale information and contrast from CNNs and part-whole relationships from CapsNet. MCRNet [26] is a Mamba capsule routing network, for part-whole relational COD, where a Mamba capsule generation (MCG) module enables lightweight type-level capsule routing, and a capsules spatial details retrieval (CSDR) module refines spatial details for final detection. The development of these advanced COD models demonstrates significant progress in detecting camouflaged objects by integrating multi-scale feature extraction, adaptive learning, and novel supervision techniques, paving the way for more robust and accurate object detection in challenging environments.

## 2.2 Video Camouflaged Object Segmentation

Video camouflaged object segmentation (VCOS), an important task of CSU, aims to detect camouflaged objects in dynamic and video-based environments. To deal with VCOS, models are required to consider object motion, temporal continuity, and varying backgrounds, adding additional complexities to COD. Several advanced models have been introduced to tackle these challenges by integrating motion learning and temporal-spatial attention mechanisms. IMEX [27] introduces implicit-explicit motion learning, allowing for more robust detection of

camouflaged objects through an integration of both implicit and explicit motion features. Temporal-spatial attention is also crucial for many VCOS models, as seen in TSP-SAM [28], which enhances SAM with a focus on temporal-spatial prompt learning to identify subtle object movements. SAM-PM [29] further extends this approach by applying spatio-temporal attention to boost accuracy in video sequences by tracking subtle changes in movement and background. An earlier method [6] focuses on implicit motion handling to refine video object detection, particularly for scenes where motion is subtle or hard to detect. OCLR [30] introduces three scores to automatically evaluate the effectiveness of camouflage by measuring background-foreground similarity and boundary visibility. These scores are used to enhance camouflaged datasets and integrate them into a generative model. MG [31] introduces a self-supervised Transformer-based model for motion segmentation using optical flow demonstrating the effectiveness of motion cues over visual appearance in VCOS. These models illustrate the importance of combining motion analysis with traditional object detection techniques to enhance VCOS performance in real-world applications.

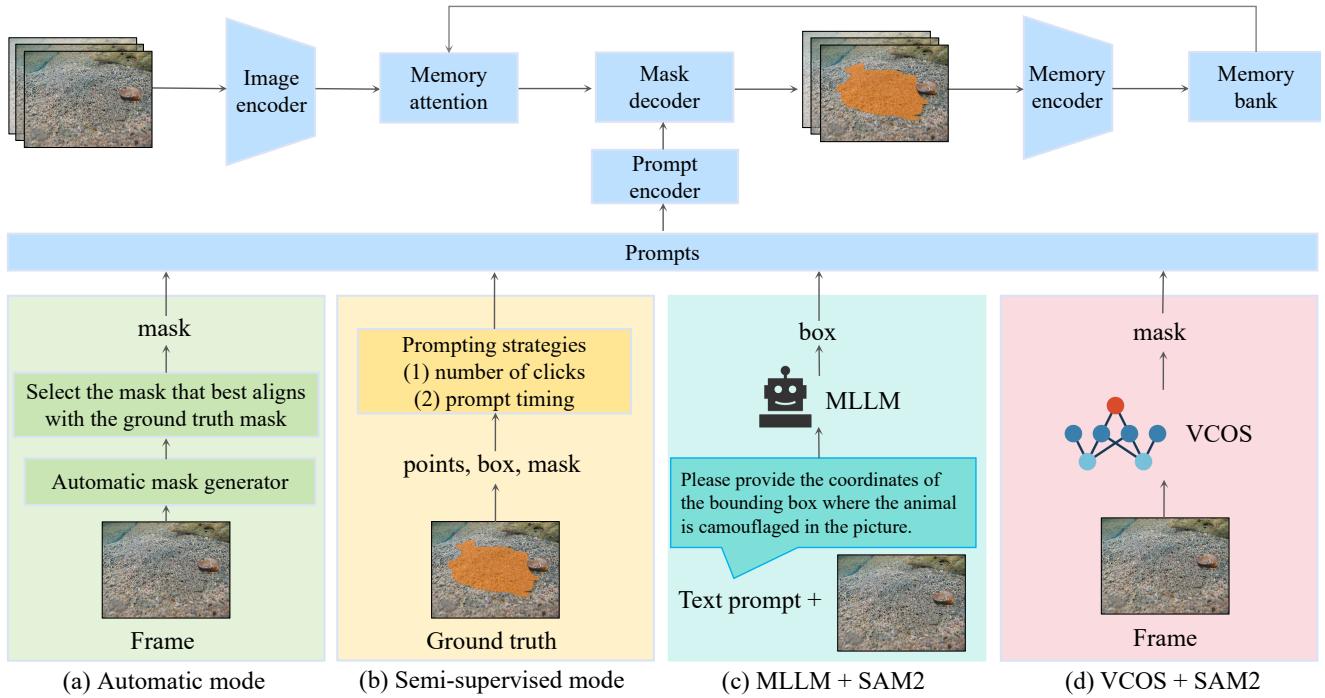
## 2.3 Segment Anything Model 2

Segment Anything Model 2 (SAM2) [9] is a vision foundation model for segmenting objects across images and videos. SAM2 has shown excellent performances in medical video and 3D segmentation, including tasks such as polyp detection, surgical video segmentation, and other medical image segmentation [32–38]. Furthermore, SAM2 has been applied in the segmentation of 3D meshes and point clouds [39], remote sensing [40], and image camouflaged object detection [37, 38, 41].

Despite these advances, to the best of our knowledge, there is no specific study focusing on VCOS using SAM2. This study fills this gap by systematically examining SAM2’s performance in VCOS and proposing strategies to further improve SAM2’s capability.

## 3 Methods

In this section, we present our evaluation framework, as summarized in Fig. 1. We also explain our strategy of fine-tuning SAM2. We first introduce the dataset in Sect. 3.1 and describe the evaluation metrics in Sect. 3.2. In Sect. 3.3, we outline the experimental setup for the automatic and semi-supervised modes of SAM2. Additionally, Sect. 3.4 details the refinement process for results generated by existing MLLM and VCOS



**Fig. 1 Overview of our evaluation framework.** The framework explores different prompts to SAM2 for VCOS. (a) Automatic mode selects the mask (generated by the built-in automatic mask generator) that best aligns with the ground truth mask to serve as the mask prompt. (b) The semi-supervised mode explores different prompt types and different prompt timing based on clicks, boxes, and masks. (c) MLLM + SAM2 utilizes an MLLM to generate bounding box coordinates as the box prompt. (d) VCOS + SAM2 employs a VCOS model to generate a coarse mask as the mask prompt. MLLM: multi-model large language model; SAM2: segment anything model 2; VCOS: video camouflaged object segmentation.

methods. Finally, we present the fine-tuning procedure of SAM2 on MoCA-Mask in Sect. 3.5.

### 3.1 Datasets

We use two video COD datasets: MoCA-Mask [6] and CAD [11]. MoCA-Mask is a densely annotated dataset derived from the moving camouflaged animals (MoCA) dataset [42]. It consists of 87 video sequences, with 71 sequences totaling 19,313 frames for training and 16 sequences comprising 3,626 frames for inference, extending the original bounding box annotations to dense segmentation masks on every 5th frames. The camouflaged animal dataset (CAD) includes 9 short video sequences in total that have 181 hand-labeled masks on every 5th frame. Due to some missing ground truth in the CAD test set, we only evaluate the valid video sequences.

### 3.2 Metrics

We use seven common metrics for evaluation including  $S$ -measure ( $S_m$ ) [43], weighted  $F$ -measure ( $F_\beta^\omega$ ) [44], mean absolute error (MAE) [45],  $F$ -measure ( $F_\beta$ ) [46],

$E$ -measure ( $E_m$ ) [47], mean Dice (mDice) and mean IoU (mIoU).

### 3.3 Prompting Strategies

We evaluate SAM2’s performance in both automatic and semi-supervised modes by utilizing the automatic mask generator and the interactive prompts, respectively. These two modes allow us to thoroughly assess SAM2’s flexibility and effectiveness.

#### 3.3.1 Automatic Mode

We evaluate SAM2 in automatic mode using its built-in automatic mask generator. In this setup, following the work [48], SAM2 automatically generates a number of segmentation masks on the first video frame and we select the mask with the highest IoU value compared to the corresponding ground truth. Let  $M = \{m_1, m_2, \dots, m_n\}$  represent the set of segmentation masks generated by SAM2 on the first video frame.  $G$  represents the ground truth, and  $\text{IoU}(m_i, G)$  calculates the IoU between the generated mask  $m_i$  and the ground truth  $G$ . The mask  $m^*$  with the highest IoU is selected

using Eq.(1):

$$m^* = \arg \max_{m_i \in M} \text{IoU}(m_i, G) \quad (1)$$

The selected mask  $m^*$  is used as the mask prompt, without any manual modification. This ensures that SAM2 operates in an unsupervised manner, relying on its automatic mask generation capabilities to track and segment camouflaged objects throughout the video.

This evaluation aims to assess SAM2’s effectiveness in segmenting the camouflaged animals in the video without any user guidance and explore SAM2’s potential in such scenarios.

### 3.3.2 Semi-supervised Mode

In the semi-supervised mode, we employ three distinct prompt strategies: click-based, box-based, and mask-based prompts. Each strategy is evaluated across different frames to investigate how prompt types and timing affect SAM2’s segmentation performance. For click-based prompts, 1, 3, and 5 foreground clicks (camouflaged animals) are randomly selected from the corresponding ground-truth mask. For box-based and mask-based prompts, we directly use the object’s bounding box or mask from the dataset as the prompt.

### 3.3.3 Prompt Timing and Frame Selection

We extend the analysis beyond the first frame by applying prompts at different times in the video sequence. In this experiment, the  $0^{th}$ ,  $5^{th}$ ,  $10^{th}$ ,  $-11^{th}$ ,  $-6^{th}$ ,  $-1^{st}$ , and middle frame are selected as the prompted frames. Here, the frame index follows indexing rules in the Python list, i.e.,  $-1^{st}$  represents the last frame. These frames are purposely chosen, from the beginning to the end, allowing us to analyze how prompt timing affects SAM2’s performance. Early frames like the  $0^{th}$ ,  $5^{th}$ , and  $10^{th}$  provide insight into how well the model tracks and segments the object from the start, while frames closer to the end,  $-11^{th}$ ,  $-6^{th}$ , and  $-1^{st}$  help evaluate SAM2’s ability to handle backward propagation over time. The middle frame is helpful to assess how well SAM2 performs when provided with information at a pivotal moment in the video sequence. This comprehensive selection of frames allows us to analyze how the prompt timing impacts SAM2’s segmentation accuracy and robustness when processing videos.

## 3.4 Refine MLLMs and VCOS Methods with SAM2

In this experiment, we explore using SAM2 to refine the results generated by the existing MLLMs and VCOS methods.

### 3.4.1 Refine MLLMs with SAM2

The use of MLLMs (Multimodal Large Language Models) with SAM2 is motivated by the limitation that most current MLLMs can only output bounding boxes, but not the segmentation masks. If the approach of using MLLMs-generated bounding boxes as prompts for SAM2 proves effective, it allows for automated identification and segmentation of objects, removing the need for manual prompts.

To be more specific, we employ two large Multimodal LLM models, LLaVA-1.5-7b [49] and Shikra-7b-delta-v1 [50], in a zero-shot setting to detect camouflaged objects in the first frame of the video. We experimented with several prompts to ask MLLMs to generate the bounding box inspired by [51], and we finally selected the prompt “Please provide the coordinates of the bounding box where the animal is camouflaged in the picture”. The input to the model consists of this question prompt alongside the visual input of the first video frame. The model processes this input and outputs the coordinates of a bounding box that is presumed to encapsulate the camouflaged object. The bounding box coordinates generated by the MLLMs serve as the box prompt for SAM2. SAM2 uses these coordinates to guide the segmentation process across the whole video. By leveraging the box prompt, SAM2 is expected to effectively segment the camouflaged object, even in the presence of complex backgrounds and low contrast.

### 3.4.2 Refine VCOS with SAM2

In this experiment, we focus on refining the camouflaged object masks generated by existing VCOS models using SAM2. Since VCOS models already output the object masks, we explore how SAM2 can enhance the quality of these masks through its advanced segmentation capabilities.

Specifically, we use the object masks produced by the VCOS models as prompts for SAM2. These masks serve as initial rough segmentation, which SAM2 uses to further refine the details. By leveraging SAM2’s powerful mask tracking capabilities, we aim to improve the results generated by existing VCOS methods. We use the predicted mask of the middle frame as the prompt.

### 3.5 Fine-tune SAM2 on MoCA-Mask

SAM2 was trained on the large-scale SA-V [9] dataset which generally contains videos from common scenes. We propose a fine-tuning strategy to adjust SAM2’s parameter on the popular MoCA-Mask [6] dataset to improve its ability on challenging camouflaged scenes. Specifically, we follow the work of MedSAM2 [52], fine-tune the image encoder, the mask decoder, and both, with freezing other parameters to maintain generalization since SAM2 was initially trained on a much larger dataset. We followed the approach of MedSAM2 and added a shift of the bounding box used during training. This selective fine-tuning focuses on adapting SAM2 to the specific challenges of camouflaged object detection, where precise feature encoding and decoding are critical. We employ the AdamW [53] optimizer with the combination of Dice loss and Binary Cross-Entropy (BCE) loss to achieve accurate segmentation of camouflaged objects. We use the learning rate ( $lr = 1e - 8$ ) and weight decay ( $wd = 0.01$ ). The overall loss function is computed as a sum of the segmentation loss and the BCE loss:

$$\mathcal{L} = \mathcal{L}_{seg} + \mathcal{L}_{ce} \quad (2)$$

## 4 Results

In this section, we present the results of our evaluation and adaptation of SAM2. We compare SAM2’s performance with state-of-the-art methods and analyze the effects of different prompting strategies, including variations in the number of clicks and prompt timing. We also examine the effectiveness of integrating MLLMs and existing VCOS methods with SAM2 and fine-tuning SAM2 on the MoCA-Mask dataset. We conduct all experiments on a single NVIDIA RTX A6000 with the `torch.float32` precision.

### 4.1 Comparison with State-of-the-art VCOS Methods

We compare the results obtained by prompting SAM2 using 1-click, box, and mask at the first video frame, with existing state-of-the-art methods, in Tab. 1 and Tab. 2. It is observed that SAM2 with semi-supervised mode outperforms current SOTA models in the VCOS task, where prompts guide the segmentation process. This naturally gives SAM2 an advantage, as the use of interactive prompts allows for better adaptability compared to fully automated models. These results highlight

the effectiveness of SAM2 for camouflaged video segmentation.

Notably, SAM2 demonstrates a strong balance between model size, inference speed (FPS), and segmentation performance, making it highly adaptable for VCOS. Compared to models such as EGNet (111.7M), BASNet (87.1M), and PNS-Net (142.9M), SAM2 achieves superior performance with significantly fewer parameters, particularly in its Hiera-T (38.9M) and Hiera-S (46.0M) versions. Inference speed is another critical factor, with Hiera-T achieving 47.2 FPS and Hiera-S running at 43.3 FPS, outperforming many state-of-the-art models, including SINet-V2 (38.3 FPS) and SLT-Net-LT (35.6 FPS). While SAM2 of larger size versions like Hiera-B+ (80.8M) further enhances segmentation accuracy, they come at the cost of increased computational complexity and slower FPS. Nevertheless, SAM2 achieves higher segmentation accuracy than transformer-based models such as ZoomNeXt (26.5 FPS) despite being faster and more lightweight, demonstrating its efficiency. These results highlight that Hiera-T and Hiera-S offer the best trade-off between speed and model size, making SAM2 a strong candidate for real-time and resource-constrained VCOS scenarios.

### 4.2 Effect of Prompting Strategies

We compare results obtained from various prompting strategies as mentioned in Sect. 3.3. From the results, we have three observations: 1) Mask-based prompt results in the best segmentation result, compared with click and box prompts; 2) Increasing the number of clicks significantly improves the segmentation result; 3) Prompting at the middle frame of the video generally gives better performance than prompting at other frames and results vary a lot when prompting at different times.

#### 4.2.1 Comparisons among Different Prompting Strategies

Tab. 1 and Tab. 2 presents a comprehensive comparison of SAM2’s performance using different prompting strategies (1-click, box, and mask) across various model sizes on the CAD and MoCA-Mask datasets. The results clearly demonstrate the superiority of mask-based prompts over both 1-click and box-based strategies. This result is consistent with our intuition: the more detailed the prompt, the better the segmentation. For instance, on the CAD dataset, SAM2 with mask prompting achieves the highest mIoU of 0.775 for the large model (Hiera-L), outperforming both 1-click (mIoU 0.497) and box-based prompting (mIoU 0.735). Similar trends are observed on MoCA-Mask with Hiera-L, where

**Table 1** State-of-the-art comparisons on the CAD dataset [11]. FPS: inference speed; Param.: parameters; MAE: mean absolute error. The best results are shown in bold.

Model	Backbone	FPS	Param.	CAD						
				$S_m \uparrow$	$F_\beta^\omega \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	$E_m \uparrow$	mDice $\uparrow$	mIoU $\uparrow$
EGNet [54]	ResNet-50	5.7	111.7M	0.619	0.298	0.044	0.350	0.666	0.324	0.243
BASNet [55]	ResNet-50	37.4	87.1M	0.639	0.349	0.054	0.394	0.773	0.393	0.293
CPD [56]	ResNet-50	43.7	47.9M	0.622	0.289	0.049	0.357	0.667	0.330	0.239
PraNet [57]	ResNet-50	41.9	32.6M	0.629	0.352	0.042	0.397	0.763	0.378	0.290
SINet [5]	ResNet-50	56.5	48.9M	0.636	0.346	0.041	0.395	0.775	0.381	0.283
SINet-V2 [58]	Res2Net-50	38.3	27.0M	0.653	0.382	0.039	0.432	0.762	0.413	0.318
PNS-Net [59]	ResNet-50	31.7	142.9M	0.655	0.325	0.048	0.417	0.673	0.384	0.290
RCRNet [60]	ResNet-50	58.1	53.8M	0.627	0.287	0.048	0.328	0.666	0.309	0.229
MG [31]	VGG	272.9	4.8M	0.594	0.336	0.059	0.375	0.692	0.368	0.268
SLT-Net-LT [6]	PVTv2-B5	35.6	82.3M	0.696	0.481	0.030	0.524	0.845	0.493	0.402
ZoomNeXt [7]	PVTv2-B5	26.5	84.8M	0.757	0.593	0.020	0.631	0.865	0.599	0.510
SAM2 [9] (1-click)	Hiera-T	47.2	38.9M	0.754	0.612	0.033	0.652	0.813	0.622	0.515
SAM2 [9] (box)	Hiera-T	47.2	38.9M	0.852	0.803	0.016	0.826	0.951	0.810	0.694
SAM2 [9] (mask)	Hiera-T	47.2	38.9M	0.887	0.845	0.010	0.856	0.971	0.852	0.756
SAM2 [9] (1-click)	Hiera-S	43.3	46.0M	0.742	0.587	0.031	0.641	0.755	0.586	0.486
SAM2 [9] (box)	Hiera-S	43.3	46.0M	0.879	0.840	0.011	0.854	0.969	0.844	0.743
SAM2 [9] (mask)	Hiera-S	43.3	46.0M	0.888	0.854	0.010	0.865	0.975	0.859	0.766
SAM2 [9] (1-click)	Hiera-B+	34.8	80.8M	0.775	0.669	0.027	0.736	0.823	0.671	0.546
SAM2 [9] (box)	Hiera-B+	34.8	80.8M	0.866	0.816	0.012	0.835	0.960	0.821	0.714
SAM2 [9] (mask)	Hiera-B+	34.8	80.8M	0.883	0.842	0.009	0.856	0.966	0.847	0.752
SAM2 [9] (1-click)	Hiera-L	24.2	224.4M	0.749	0.592	0.028	0.647	0.738	0.592	0.497
SAM2 [9] (box)	Hiera-L	24.2	224.4M	0.872	0.836	0.012	0.853	0.969	0.839	0.735
SAM2 [9] (mask)	Hiera-L	24.2	224.4M	<b>0.893</b>	<b>0.863</b>	<b>0.008</b>	<b>0.873</b>	<b>0.976</b>	<b>0.864</b>	<b>0.775</b>

mask prompting achieves 0.709 in mIoU, compared to 0.575 for 1-click prompting and 0.653 for box-based prompting.

#### 4.2.2 Impact of the Number of Clicks for Point Prompting

We examine the impact of different numbers of clicks (1, 3, and 5) used as prompts at the middle frame on segmentation performance, across various model sizes. The results on the MoCA-Mask dataset are shown in Tab. 3. For the SAM2-L model, increasing the number of clicks leads to a consistent performance improvement; the 5-click prompt achieves the highest scores (i.e.,  $S_m=0.831$ , mDice=0.748, and mIoU=0.652) compared to both the 1-click and 3-click settings. In contrast, for the other three models (SAM2-B+, SAM2-S, and SAM2-T), the 3-click outperforms the 5-click settings, although the performance differences are minimal. For instance, the SAM2-B+ model achieves a mIoU of 0.624 with 3-click prompts against 0.606 with 5-click prompts. Although the trends between the 3-click and 5-click settings vary across different model sizes, both settings consistently outperform the 1-click settings. This indicates that incorporating additional point prompts generally provides more informative guidance, thereby enhancing the seg-

mentation performance. In conclusion, employing multiple clicks leads to improved results compared to using a single click, highlighting the benefit of human guidance in the VCOS task.

#### 4.2.3 Impact of Prompt Timing

We evaluate the effect of prompt timing on the performance using click-based, box-based, and mask-based prompts on the small version of the SAM2 using the MoCA-Mask dataset. Tab. 4 presents the results for different prompt timings across all three prompt strategies. In general, the results show that applying the prompt on the middle frame yields the best segmentation performance across all strategies. For instance, using the click-based prompts in the middle frame gives a mIoU of 0.538, which is higher than prompting at other times. Similarly, for the box-based and mask-based prompts, the middle frame provides the best results in most evaluation metrics. The trend is particularly obvious for the mask-based prompts where the middle frame achieves a mIoU of 0.719, the highest among all experiments. These results suggest that prompt timing is a critical factor in achieving optimal segmentation performance.

**Table 2** State-of-the-art comparisons on the MoCA-Mask dataset [6].

Model	Backbone	FPS	Param.	MoCA-Mask						
				$S_m \uparrow$	$F_\beta^\omega \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	$E_m \uparrow$	mDice $\uparrow$	mIoU $\uparrow$
EGNet [54]	ResNet-50	5.7	111.7M	0.547	0.110	0.035	0.136	0.574	0.143	0.096
BASNet [55]	ResNet-50	37.4	87.1M	0.561	0.154	0.042	0.173	0.598	0.190	0.137
CPD [56]	ResNet-50	43.7	47.9M	0.561	0.121	0.041	0.152	0.613	0.162	0.113
PraNet [57]	ResNet-50	41.9	32.6M	0.614	0.266	0.030	0.296	0.674	0.311	0.234
SINet [5]	ResNet-50	56.5	48.9M	0.598	0.231	0.028	0.256	0.699	0.277	0.202
SINet-V2 [58]	Res2Net-50	38.3	27.0M	0.588	0.204	0.031	0.229	0.642	0.245	0.180
PNS-Net [59]	ResNet-50	31.7	142.9M	0.526	0.059	0.035	0.084	0.530	0.084	0.054
RCRNet [60]	ResNet-50	58.1	53.8M	0.555	0.138	0.033	0.159	0.527	0.171	0.116
MG [31]	VGG	272.9	4.8M	0.530	0.168	0.067	0.195	0.561	0.181	0.127
SLT-Net-LT [6]	PVTv2-B5	35.6	82.3M	0.631	0.311	0.027	0.331	0.759	0.360	0.272
ZoomNeXt [7]	PVTv2-B5	26.5	84.8M	0.734	0.476	0.010	0.497	0.736	0.497	0.422
SAM2 [9] (1-click)	Hiera-T	47.2	38.9M	0.680	0.523	0.074	0.556	0.775	0.532	0.444
SAM2 [9] (box)	Hiera-T	47.2	38.9M	0.813	0.697	0.006	0.707	0.895	0.726	0.621
SAM2 [9] (mask)	Hiera-T	47.2	38.9M	0.846	0.760	<b>0.004</b>	0.770	0.918	0.771	0.679
SAM2 [9] (1-click)	Hiera-S	43.3	46.0M	0.719	0.534	0.009	0.567	0.757	0.546	0.456
SAM2 [9] (box)	Hiera-S	43.3	46.0M	0.826	0.715	0.006	0.724	0.887	0.738	0.637
SAM2 [9] (mask)	Hiera-S	43.3	46.0M	<b>0.862</b>	0.791	<b>0.004</b>	0.801	<b>0.946</b>	0.802	0.706
SAM2 [9] (1-click)	Hiera-B+	34.8	80.8M	0.752	0.622	0.008	0.666	0.804	0.639	0.523
SAM2 [9] (box)	Hiera-B+	34.8	80.8M	0.819	0.707	0.007	0.717	0.909	0.730	0.631
SAM2 [9] (mask)	Hiera-B+	34.8	80.8M	0.846	0.768	<b>0.004</b>	0.780	0.915	0.776	0.684
SAM2 [9] (1-click)	Hiera-L	24.2	224.4M	0.787	0.658	0.006	0.679	0.844	0.670	0.575
SAM2 [9] (box)	Hiera-L	24.2	224.4M	0.832	0.729	0.006	0.733	0.926	0.748	0.653
SAM2 [9] (mask)	Hiera-L	24.2	224.4M	<b>0.862</b>	<b>0.796</b>	<b>0.004</b>	<b>0.806</b>	0.943	<b>0.806</b>	<b>0.709</b>

**Table 3** Click-based prompt performance on the middle frame with different click counts on the MoCA-Mask dataset [6].

Model	Click Count	MoCA-Mask						
		$S_m \uparrow$	$F_\beta^\omega \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	$E_m \uparrow$	mDice $\uparrow$	mIoU $\uparrow$
SAM2-L	1	0.746	0.619	0.064	0.635	0.810	0.634	0.548
	3	0.785	0.702	0.063	0.713	0.873	0.718	0.624
	5	<b>0.831</b>	<b>0.729</b>	<b>0.008</b>	<b>0.736</b>	<b>0.909</b>	<b>0.748</b>	<b>0.652</b>
SAM2-B+	1	0.738	0.604	0.065	0.617	0.821	0.617	0.533
	3	<b>0.812</b>	<b>0.701</b>	<b>0.007</b>	<b>0.712</b>	<b>0.886</b>	<b>0.722</b>	<b>0.624</b>
	5	0.808	0.677	0.009	0.686	0.870	0.703	0.606
SAM2-S	1	0.748	0.611	0.058	0.626	0.808	0.621	0.538
	3	0.783	<b>0.682</b>	0.053	<b>0.693</b>	<b>0.873</b>	<b>0.703</b>	<b>0.606</b>
	5	<b>0.788</b>	0.671	<b>0.047</b>	0.680	0.866	0.699	0.599
SAM2-T	1	0.768	0.614	<b>0.007</b>	0.638	0.820	0.629	0.536
	3	0.787	<b>0.649</b>	0.011	<b>0.661</b>	0.872	<b>0.676</b>	<b>0.576</b>
	5	<b>0.793</b>	0.644	0.012	0.654	<b>0.881</b>	0.675	0.572

### 4.3 Automatic Mode Results

The results of using SAM2 in automatic mode on the MoCA-Mask dataset are shown in Tab. 5. It can be observed that the performance is notably lower compared to semi-supervised mode with guided prompts. Fig. 2 presents some visualizations of automatically generated masks for the first frame in each video, using Hiera-L as the backbone. From left to right, SAM2 successfully segments the camouflaged objects in the first two examples. However, in the middle example, it only partially

segments the camouflaged object, and in the last two examples, SAM2 fails to segment the camouflaged object entirely. SAM2 is optimized for natural scenes rather than camouflaged environments, so it is more difficult for SAM2 to determine camouflaged objects without any guidance (prompts), especially when the camouflaged objects are not clearly distinguishable from the background.

**Table 4** Performance of the SAM2-S with different prompt timings on the MoCA-Mask dataset [6].

Prompt Type	Frame	MoCA-Mask						
		$S_m \uparrow$	$F_\beta^\omega \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	$E_m \uparrow$	mDice $\uparrow$	
1-click	0	0.719	0.534	0.009	0.567	0.757	0.546	0.456
	5	<b>0.766</b>	0.597	<b>0.007</b>	0.605	0.795	0.611	0.532
	10	0.732	0.548	0.009	0.568	0.763	0.561	0.479
	-11	0.732	0.566	0.054	0.585	0.758	0.578	0.497
	-6	0.746	0.603	0.054	<b>0.632</b>	0.779	0.618	0.524
	-1	0.742	0.601	0.067	0.623	0.773	0.616	0.531
Box	middle	0.748	<b>0.611</b>	0.058	0.626	<b>0.808</b>	<b>0.621</b>	<b>0.538</b>
	0	0.826	0.715	0.006	0.724	0.887	0.738	0.637
	5	0.816	0.698	0.006	0.705	0.871	0.718	0.623
	10	0.810	0.699	<b>0.005</b>	0.715	0.878	0.717	0.614
	-11	0.832	0.719	<b>0.005</b>	0.723	0.899	0.737	0.643
	-6	<b>0.834</b>	0.731	<b>0.005</b>	0.735	0.898	0.747	0.653
Mask	-1	0.831	0.718	<b>0.005</b>	0.722	0.891	0.739	0.644
	middle	<b>0.834</b>	<b>0.738</b>	<b>0.005</b>	<b>0.744</b>	<b>0.906</b>	<b>0.753</b>	<b>0.657</b>
	0	0.862	0.791	<b>0.004</b>	0.801	0.946	0.802	0.706
	5	0.850	0.770	<b>0.004</b>	0.780	0.930	0.781	0.688
	10	0.854	0.777	<b>0.004</b>	0.787	0.932	0.787	0.693
	-11	0.863	0.784	<b>0.004</b>	0.790	0.936	0.798	0.704
SAM2	-6	0.862	0.787	<b>0.004</b>	0.797	0.939	0.799	0.703
	-1	0.860	0.781	<b>0.004</b>	0.790	0.934	0.797	0.699
	middle	<b>0.873</b>	0.803	0.004	0.810	0.954	0.813	0.719

**Fig. 2** Visualization of masks generated by Automatic mode of SAM2 on MoCA-Mask. From top to bottom: the input frames, masks generated in automatic mode, and the ground truths. SAM2 can generate multiple masks (shown in different colors) for this mode. Best viewed in color.

#### 4.4 Refine MLLMs and VCOS with SAM2

##### 4.4.1 Refine MLLMs with SAM2

In our experiment, we utilize two large Multimodal LLM models, LLaVA-1.5-7b [49] and Shikra-7b-delta-v1 [50], in combination with SAM2 for video segmentation. The results are presented in Tab. 6. It shows that the performance is unsatisfactory. For instance, the mIoU and mDice scores for both LLaVA+SAM2 and Shikra+SAM2 models are relatively low, with the large model sizes achieving mIoU values of 0.291 and 0.124, respectively. The poor performance can be attributed to

the heavy reliance of SAM2 on accurate bounding box detection by the MLLMs in the first frame. When the MLLMs fail to generate the correct bounding box, it severely impacts segmentation in the subsequent frames, leading to poor segmentation masks. This highlights the crucial role of accurate object detection in the initial frame. Although the current MLLM+SAM2 framework does not achieve strong performance on the VCOS problem, it presents a promising heuristic approach for fully automating SAM2-based segmentation without requiring manual prompts. For example, LLaVA+SAM2-L, while performing worse than the semi-supervised mode

**Table 5** Automatic mode performance of SAM2 on the MoCA-Mask dataset [6].

Size	$S_m \uparrow$	$F_\beta^\omega \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	$E_m \uparrow$	mDice $\uparrow$	mIoU $\uparrow$
SAM2-L	0.466	0.157	0.190	0.165	0.584	0.157	0.138
SAM2-B+	0.468	0.137	<b>0.141</b>	0.150	0.561	0.138	0.113
SAM2-S	<b>0.497</b>	<b>0.201</b>	0.148	<b>0.214</b>	0.608	<b>0.202</b>	<b>0.174</b>
SAM2-T	0.495	0.166	0.154	0.167	<b>0.634</b>	0.167	0.151

**Table 6** Performance comparison of LLaVA+SAM2 and Shikra+SAM2 models on VCOS on the MoCA-Mask dataset [6].

Model	$S_m \uparrow$	$F_\beta^\omega \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	$E_m \uparrow$	mDice $\uparrow$	mIoU $\uparrow$
LLaVA+SAM2-L	<b>0.624</b>	<b>0.315</b>	<b>0.046</b>	<b>0.317</b>	<b>0.688</b>	<b>0.334</b>	<b>0.291</b>
LLaVA+SAM2-B+	0.552	0.213	0.080	0.227	0.657	0.249	0.190
LLaVA+SAM2-S	0.518	0.159	0.090	0.173	0.558	0.179	0.137
LLaVA+SAM2-T	0.605	0.289	0.070	0.298	0.687	0.329	0.266
Shikra+SAM2-L	<b>0.502</b>	0.146	<b>0.107</b>	0.155	<b>0.590</b>	0.157	0.124
Shikra+SAM2-B+	0.490	<b>0.193</b>	0.170	<b>0.204</b>	0.537	<b>0.208</b>	<b>0.171</b>
Shikra+SAM2-S	0.444	0.107	0.190	0.115	0.530	0.114	0.090
Shikra+SAM2-T	0.473	0.119	0.139	0.130	0.532	0.133	0.101

of SAM2, still surpasses many existing methods, as we listed in Tab. 2.

The integration of MLLMs inevitably introduces additional computational overhead; however, in our approach, this impact is minimal. Specifically, we utilize MLLMs only for the first frame of the video sequence to generate the initial bounding box prompt rather than processing each frame individually. Once the prompt is obtained, SAM2 operates independently throughout the remaining frames, ensuring efficient segmentation and tracking without further reliance on MLLM inference. As a result, the FPS degradation and latency introduced by MLLM processing are negligible over the course of a long video sequence. Since the MLLM inference occurs only once at the beginning, its computational cost is amortized across the entire sequence, making the overhead practically insignificant for real-world applications.

However, due to the inherent difficulty in detecting camouflaged animals, MLLMs often generate imprecise bounding box coordinates, leading to incorrect prompts for SAM2. These inaccurate prompts, in turn, cause segmentation failures and disrupt object tracking across frames, reducing the effectiveness of the framework.

#### 4.4.2 Refine VCOS with SAM2

In this experiment, we focus on refining the segmentation masks generated by the TSP-SAM [28] model for VCOS using SAM2. The initial masks produced by TSP-SAM are used as prompts to SAM2 for refining details. The TSP-SAM model segments each frame based on the preceding frames, progressively improving segmentation with each step. Therefore, we assume that the last frame in the video contains the most refined segmentation result, making it the best candidate for further refinement

by SAM2, so we prompt the mask of the last frame of the video sequence for SAM2 refinement.

Tab. 7 presents the results of this refinement process across different model sizes (large, base plus, small, and tiny). Compared to the baseline (TSP-SAM), SAM2 clearly shows improvements on most metrics. For instance, using the large model for the refinement gives an  $S_m$  of 0.696 and mIoU of 0.408, marking improvements in structure and segmentation quality. Similar trends are observed across other models, with the tiny model also showing an increase in mIoU improving from 0.388 to 0.392. These results highlight that SAM2’s advanced capabilities can clearly enhance the performance of VCOS models.

#### 4.5 Fine-tune SAM2 on MoCA-Mask

We fine-tune the SAM2-Tiny model on the MoCA-Mask dataset for 50 epochs. When evaluating, we use the box-based prompt of the first frame as input, the results are shown in Tab. 8. We observe notable performance improvements. The mIoU increases by approximately 0.029, from 0.623 to 0.652. Similarly, the mDice score sees a significant improvement of 0.028, from 0.728 to 0.756. The table further breaks down the improvements for different fine-tuning configurations. Fine-tuning both the image encoder and mask decoder simultaneously yields the best result. These results demonstrate the effectiveness of fine-tuning the image encoder and the mask decoder of SAM2-T on the MoCA-Mask dataset for improving segmentation accuracy in VCOS. We also evaluated our fine-tuned SAM2-T on the CAD with the box prompt, as shown in Tab. 9. The CAD is evaluated in the zero-shot setting since the test cases are not seen

**Table 7** Refinement of TSP-SAM with SAM2 on camouflaged object segmentation.

Model	$S_m \uparrow$	$F_\beta^\omega \uparrow$	MAE $\downarrow$	$E_m \uparrow$	mDice $\uparrow$	mIoU $\uparrow$
TSP-SAM	0.689	0.444	0.008	0.808	<b>0.458</b>	0.388
SAM2-L	<b>0.696</b>	<b>0.448</b>	0.008	<b>0.825</b>	0.457	<b>0.408</b>
SAM2-B+	0.695	0.444	<b>0.007</b>	0.821	0.451	0.403
SAM2-S	0.693	0.441	0.008	0.824	0.450	0.399
SAM2-T	0.689	0.433	0.008	0.812	0.442	0.392

**Table 8** Performance of fine-tuning SAM2-T on the MoCA-Mask dataset [6]. Note that  $\clubsuit$  represents that the weights are updated during fine-tuning.

Method	$S_m \uparrow$	$F_\beta^\omega \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	$E_m \uparrow$	mDice $\uparrow$	mIoU $\uparrow$
SAM2-T(baseline)	0.815	0.699	0.006	0.710	0.895	0.728	0.623
SAM2-T(Image Encoder $\clubsuit$ )	0.831	0.724	<b>0.005</b>	0.730	0.907	0.754	0.651
SAM2-T(Mask Decoder $\clubsuit$ )	0.818	0.706	0.006	0.716	0.899	0.733	0.629
SAM2-T(Image Encoder $\clubsuit$ +Mask Decoder $\clubsuit$ )	<b>0.832</b>	<b>0.726</b>	<b>0.005</b>	<b>0.733</b>	<b>0.908</b>	<b>0.756</b>	<b>0.652</b>

**Table 9** Zero-shot performance of fine-tuned SAM2-T on the CAD dataset [11]. The image encoder and the mask decoder are fine-tuned. We evaluate with box prompt.

Method	$S_m \uparrow$	$F_\beta^\omega \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	$E_m \uparrow$	mDice $\uparrow$	mIoU $\uparrow$
SAM2-T [9]	0.852	0.803	0.016	<b>0.826</b>	0.951	0.810	0.694
SAM2-T(fine-tuned)	<b>0.854</b>	<b>0.807</b>	<b>0.015</b>	0.825	<b>0.958</b>	<b>0.817</b>	<b>0.704</b>

during fine-tuning. The mIoU increases by 0.01, from 0.694 to 0.704, showing the model’s domain adaptability.

#### 4.6 Failure Cases Analysis

In our experiments, we identify several failure cases in SAM2’s segmentation of camouflaged animals, using a 1-click prompt at the first frame, as shown in Fig. 3. These examples, including the copperhead snake, hedgehog, and stick insect, compared with the qualitative cases (in Fig. 4), reveal challenges when dealing with complex camouflaged scenarios. Here, we analyze the key factors that contribute to the failure:

1) **Low Contrast with Background:** The copperhead snake blends almost seamlessly with the surroundings, leading to an incomplete segmentation mask as SAM2 struggles to differentiate it from the background.

2) **Occlusions and Distracting Elements:** The hedgehog case highlights SAM2’s difficulty with occlusions caused by the cluttered environment, including branches and leaves, resulting in inaccurate object segmentation.

3) **Thin and Complex Structures:** The stick insect’s fine details are not well captured, especially its thin legs, showing limitations in segmenting intricate structures.

These failure cases highlight the need for improved handling of low-contrast, cluttered environments, and complex structures in future adaptations of SAM2 for VCOS.

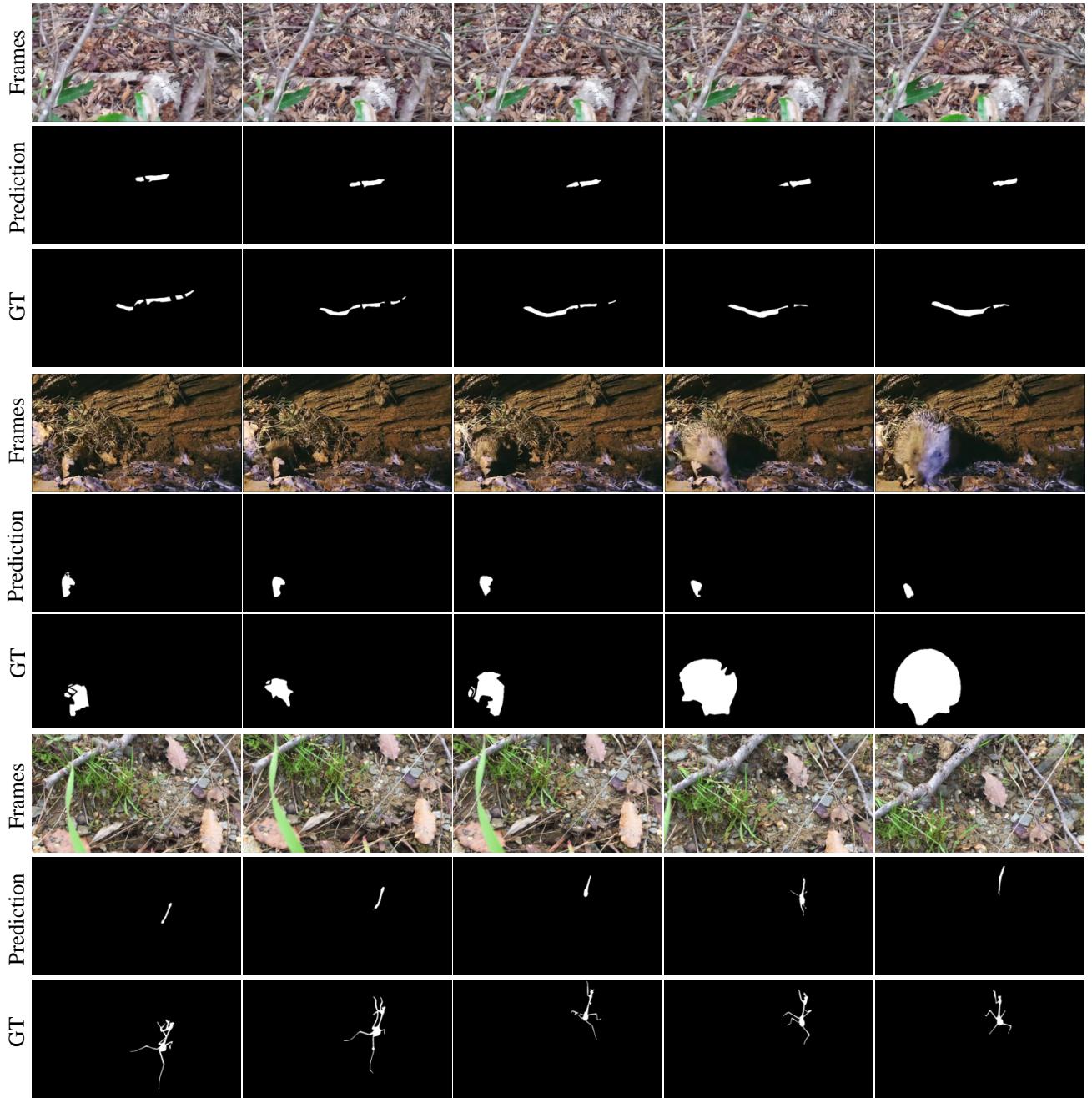
## 5 Conclusion

This study provides a systematic evaluation of the Segment Anything Model 2 (SAM2) in video camouflaged object segmentation across two popular datasets MoCA-Mask and CAD. Our experiments highlight several key findings:

1) **Prompt-Based Segmentation:** SAM2 demonstrates notable performance improvements with different prompting strategies, such as clicks, boxes, and masks. Box-based and mask-based prompts consistently outperformed click-based prompts across both datasets, with prompting at the middle frame often yielding the best results. This indicates the importance of spatial details (box and mask) in guiding SAM2 for accurate segmentation of camouflaged objects.

2) **Automatic Mode Performance:** In automatic mode, SAM2 struggles with fully unsupervised segmentation in camouflaged scenes, achieving unsatisfactory mIoU and mDice scores. The results show that user guidance or prompts are crucial for SAM2 to effectively segment camouflaged objects in dynamic environments.

3) **Refinement with SAM2:** Using SAM2 to refine outputs from VCOS models significantly improves segmentation performance. SAM2’s advanced mask refinement capabilities result in higher mIoU and mDice scores. However, when combining SAM2 with Multimodal Large Language Models (MLLMs), the results were suboptimal, highlighting the importance of accurate initial object detection by MLLMs. Nevertheless, despite its



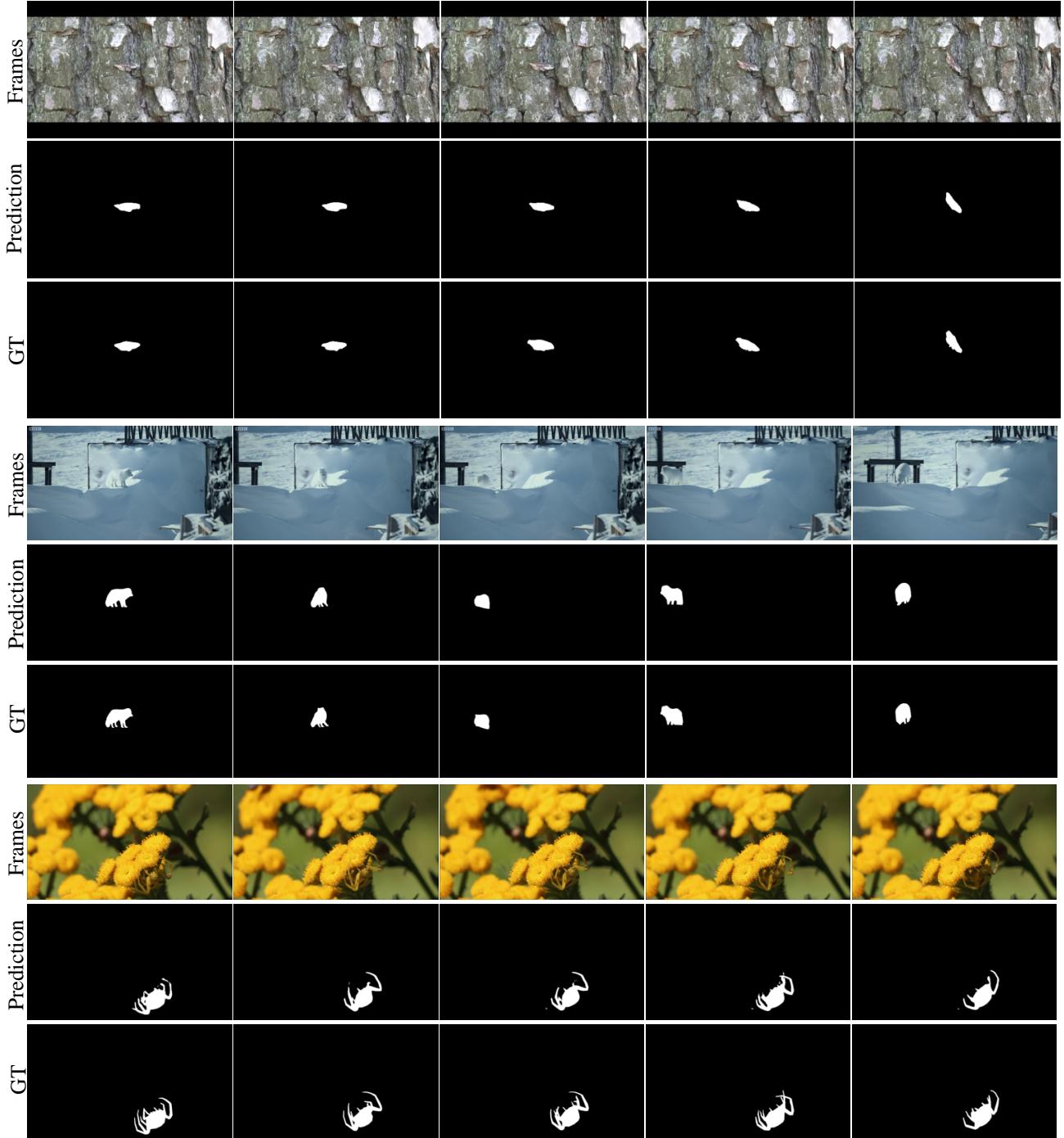
**Fig. 3** Failure cases of SAM2 on MoCA-Mask.

limitations, this approach still outperforms some existing VCOS methods, demonstrating its potential for further refinement.

**4) Fine-tuning SAM2:** Fine-tuning SAM2 on the MoCA-Mask dataset leads to clear improvements in mIoU and mDice scores, underscoring SAM2’s adaptability to camouflaged segmentation when provided with specific training data. Fine-tuning SAM2 can significantly enhance its performance for tasks involving com-

plex object-background blending, such as camouflaged segmentation.

Overall, SAM2 demonstrates strong capabilities in video camouflaged object segmentation through effective prompting strategies, model refinement, and dataset-specific fine-tuning. These findings suggest that SAM2 is a promising segmentation model for challenging camouflaged scenarios and has great potential for further improvement.



**Fig. 4** Qualitative examples of SAM2 on MoCA-Mask.

## Abbreviations

- Abbreviations. COD, camouflaged object detection; CSU, camouflaged scene understanding; MLLM, multimodal large language model; SAM2, segment anything model 2; VCOS, video camouflaged object segmentation.

## Declarations

- Availability of data and materials. The models and source code are released at <https://github.com/zhoustan/SAM2-VCOS>. The datasets used in this study are publicly available: MoCA-Mask (<https://xueliancheng.github.io/SLT-Net-project>),

- CAD (<https://vis-www.cs.umass.edu/motionSegmentation>).
- Competing interests. The authors declare that they have no conflict of interest or competing interests.
  - Funding. No funding was received to assist with the preparation of this manuscript.
  - Authors’ contributions. GS designed the concept and method of the work. All authors jointly helped experiments and analyze the results. YZ and GS both contributed to the implementation. YZ and GS wrote the paper while other authors carefully refined it. GX, YL, LB and EK provided valuable suggestions on how to improve the manuscript.
  - Author details. 1. Computer Vision Laboratory, ETH Zürich, Sternwartstrasse 7, 8092, Zürich, Switzerland. 2. Department of Informatics, University of Zürich, Binzmühlestrasse 14, 8050, Zürich, Switzerland. 3. Integrated System Laboratory, ETH Zürich, Gloriastrasse 35, 8092, Zürich, Switzerland. 4. School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. 5. Department of Electrical, Electronic, and Information Engineering, University of Bologna, Via Zamponi 33, Bologna, 40126, Bologna, Italy.
  - Acknowledgments. Not Applicable.

## References

1. Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440. Piscataway: IEEE, 2015.
2. Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969. Piscataway: IEEE, 2017.
3. Peiliang Huang, Dingwen Zhang, De Cheng, Longfei Han, Pengfei Zhu, and Junwei Han. M-RRFS: A memory-based robust region feature synthesizer for zero-shot object detection. *International Journal of Computer Vision*, 132(10):4651–4672, 2024.
4. Guolei Sun, Yun Liu, Henghui Ding, Min Wu, and Luc Van Gool. Learning local and global temporal contexts for video semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6919–6934, 2024.
5. Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2774–2784. Piscataway: IEEE, 2020.
6. Xuelian Cheng, Huan Xiong, Deng-Ping Fan, Yiran Zhong, Mehrtash Harandi, Tom Drummond, and Zongyuan Ge. Implicit motion handling for video camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13864–13873. Piscataway: IEEE, 2022.
7. Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoomnext: A unified collaborative pyramid network for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9205–9220, 2024.
8. Junyu Xie, Charig Yang, Weidi Xie, and Andrew Zisserman. Moving object segmentation: All you need is SAM (and Flow). In *Proceedings of the Asian Conference on Computer Vision*, pages 162–178. Cham: Springer, 2024.
9. Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.
10. Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026. Piscataway: IEEE, 2023.
11. Pia Bideau and Erik Learned-Miller. It’s moving! a probabilistic model for causal motion segmentation in moving camera videos. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Proceedings of the 14th European Conference on Computer Vision*, pages 433–449. Cham: Springer, 2016.
12. Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11591–11601. Piscataway: IEEE, 2021.
13. Guolei Sun, Zhaochong An, Yun Liu, Ce Liu, Christos Sakaridis, Deng-Ping Fan, and Luc Van Gool. Indiscernible object counting in underwater scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13791–13801. Piscataway: IEEE, 2023.
14. Jialun Pei, Tianyang Cheng, Deng-Ping Fan, He Tang, Chuhanbo Chen, and Luc Van Gool. Osformer: One-stage camouflaged instance segmentation with transformers. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Proceedings of the 17th European Conference on Computer Vision*, pages 19–37. Cham: Springer, 2022.
15. Ziyang Luo, Nian Liu, Wangbo Zhao, Xuguang Yang, Dingwen Zhang, Deng-Ping Fan, Fahad Khan, and Junwei Han. Vscode: General visual salient and camouflaged object detection with 2D prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17169–17180. Piscataway: IEEE, 2024.
16. Xuebin Qin, Deng-Ping Fan, Chenyang Huang, Cyril Diagne, Zichen Zhang, Adrià Cabeza Sant’Anna, Albert Suarez, Martin Jagersand, and Ling Shao. Boundary-aware segmentation network for mobile and web applications. *arXiv preprint arXiv:2101.04704*, 2021.
17. Siyuan Yao, Hao Sun, Tian-Zhu Xiang, Xiao Wang, and Xiaochun Cao. Hierarchical graph interaction transformer with dynamic token clustering for camouflaged object detection. *IEEE Transactions on Image Processing*, 33:5936–5948, 2024.
18. Zhennan Chen, Xuying Zhang, Tian-Zhu Xiang, and Ying Tai. Adaptive guidance learning for camouflaged object detection. *arXiv preprint arXiv:2405.02824*, 2024.
19. Yinghui Xing, Dexuan Kong, Shizhou Zhang, Geng Chen, Lingyan Ran, Peng Wang, and Yanning Zhang. Pre-train,

- adapt and detect: Multi-task adapter tuning for camouflaged object detection. *arXiv preprint arXiv:2307.10685*, 2023.
20. Wei Sun, Chengao Liu, Linyan Zhang, Yu Li, Pengxu Wei, Chang Liu, Jialing Zou, Jianbin Jiao, and Qixiang Ye. DQnet: Cross-model detail querying for camouflaged object detection. *arXiv preprint arXiv:2212.08296*, 2022.
  21. Xuying Zhang, Bowen Yin, Zheng Lin, Qibin Hou, Deng-Ping Fan, and Ming-Ming Cheng. Referring camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3597–3610, 2025.
  22. Jin Zhang, Ruiheng Zhang, Yanjiao Shi, Zhe Cao, Nian Liu, and Fahad Shahbaz Khan. Learning camouflaged object detection from noisy pseudo label. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gü̈l Varol, editors, *Proceedings of the 18th European Conference on Computer Vision*, pages 158–174. Cham: Springer, 2024.
  23. Kunfa Lai, Zhiyu Yang, Jie Hu, Shengchuan Zhang, Liujuan Cao, Guannan Jiang, Zhiyu Wang, Songan Zhang, and Rongrong Ji. Camoteacher: Dual-rotation consistency learning for semi-supervised camouflaged object detection. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gü̈l Varol, editors, *Proceedings of the 18th European Conference on Computer Vision*, pages 438–455. Cham: Springer, 2024.
  24. Yi Liu, Dingwen Zhang, Qiang Zhang, and Jungong Han. Integrating part-object relationship and contrast for camouflaged object detection. *IEEE Transactions on Information Forensics and Security*, 16:5154–5166, 2021.
  25. Yi Liu and Hui Meng. Camouflaged object detection via scale-feature attention and type-feature attention. In Zhouchen Lin, Ming-Ming Cheng, Ran He, Kurban Ulub, Wushouer Silamu, Hongbin Zha, Jie Zhou, and Cheng-Lin Liu, editors, *Proceedings of the 7th Chinese Conference on Pattern Recognition and Computer Vision*, pages 201–213. Cham: Springer, 2024.
  26. Dingwen Zhang, Liangbo Cheng, Yi Liu, Xinggang Wang, and Junwei Han. Mamba capsule routing towards part-whole relational camouflaged object detection. *arXiv preprint arXiv:2410.03987*, 2024.
  27. Wenjun Hui, Zhenfeng Zhu, Guanghua Gu, Meiqin Liu, and Yao Zhao. Implicit-explicit motion learning for video camouflaged object detection. *IEEE Transactions on Multimedia*, 26:7188–7196, 2024.
  28. Wenjun Hui, Zhenfeng Zhu, Shuai Zheng, and Yao Zhao. Endow SAM with keen eyes: Temporal-spatial prompt learning for video camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19058–19067. Piscataway: IEEE, 2024.
  29. Muhammad Nawfal Meeran, Gokul Adethya T, and Bhanu Pratyush Mantha. SAM-PM: Enhancing video camouflaged object detection using spatio-temporal attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1857–1866, June 2024.
  30. Hala Lamdouar, Weidi Xie, and Andrew Zisserman. The making and breaking of camouflage. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 832–842. Piscataway: IEEE, 2023.
  31. Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7177–7188. Piscataway: IEEE, 2021.
  32. Mobina Mansoori, Sajjad Shahabodini, Jamshid Abouei, Konstantinos N. Plataniotis, and Arash Mohammadi. Polyp SAM 2: Advancing zero shot polyp segmentation in colorectal cancer detection. *arXiv preprint arXiv:2408.05892*, 2024.
  33. Yiqing Shen, Hao Ding, Xinyuan Shao, and Mathias Unterath. Performance and non-adversarial robustness of the segment anything model 2 in surgical video segmentation. *arXiv preprint arXiv:2408.04098*, 2024.
  34. Jieming Yu, An Wang, Wenzhen Dong, Mengya Xu, Mobarakol Islam, Jie Wang, Long Bai, and Hongliang Ren. SAM 2 in robotic surgery: An empirical evaluation for robustness and generalization in surgical video segmentation. *arXiv preprint arXiv:2408.04593*, 2024.
  35. Haofeng Liu, Erli Zhang, Junde Wu, Mingxuan Hong, and Yueming Jin. Surgical SAM 2: Real-time segment anything in surgical video by efficient frame pruning. *arXiv preprint arXiv:2408.07931*, 2024.
  36. Yufan He, Pengfei Guo, Yucheng Tang, Andriy Myronenko, Vishwesh Nath, Ziyue Xu, Dong Yang, Can Zhao, Daguang Xu, and Wenqi Li. A short review and evaluation of sam2’s performance in 3D CT image segmentation. *arXiv preprint arXiv:2408.11210*, 2024.
  37. Tianrun Chen, Ankang Lu, Lanyun Zhu, Chaotao Ding, Chunyan Yu, Deyi Ji, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. SAM2-Adapter: Evaluating & adapting segment anything 2 in downstream tasks: Camouflage, shadow, medical image segmentation, and more. *arXiv preprint arXiv:2408.04579*, 2024.
  38. Xinyu Xiong, Zihuang Wu, Shuangyi Tan, Wenxue Li, Feilong Tang, Ying Chen, Siying Li, Jie Ma, and Guanbin Li. SAM2-UNet: Segment anything 2 makes strong encoder for natural and medical image segmentation. *arXiv preprint arXiv:2408.08870*, 2024.
  39. George Tang, William Zhao, Logan Ford, David Benhaim, and Paul Zhang. Segment any mesh: Zero-shot mesh part segmentation via lifting segment anything 2 to 3D. *arXiv preprint arXiv:2408.13679*, 2024.
  40. Osher Rafaeli, Tal Svoray, Roni Blushtein-Livnon, and Ariel Nahlieli. Prompt-based segmentation at multiple resolutions and lighting conditions using segment anything model 2. *arXiv preprint arXiv:2408.06970*, 2024.
  41. Lv Tang and Bo Li. Evaluating SAM2’s role in camouflaged object detection: From SAM to SAM2. *arXiv preprint arXiv:2408.21596*, 2024.
  42. Hala Lamdouar, Charig Yang, Weidi Xie, and Andrew Zisserman. Betrayed by motion: Camouflaged object discovery via motion segmentation. In *Proceedings of the Asian conference on computer vision*, pages 488–503. Cham: Springer, 2020.
  43. Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4548–4557. Piscataway: IEEE, 2017.
  44. Ran Margolin, Lihi Zelnik-Manor, and Ayallet Tal. How to evaluate foreground maps? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255. Piscataway: IEEE, 2014.
  45. Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 733–740. Piscataway: IEEE, 2012.
  46. Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Süsstrunk. Frequency-tuned salient

- region detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1597–1604. Piscataway: IEEE, 2009.
47. Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 698–704. Cham: Springer, 2018.
  48. Tiantian Zhang, Zhangjun Zhou, and Jialun Pei. Evaluation study on SAM 2 for class-agnostic instance-level segmentation. *arXiv preprint arXiv:2409.02567*, 2024.
  49. Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the 37th Advances in neural information processing systems*, pages 34892–34916. Red Hook: Curran Associates, 2024.
  50. Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multi-modal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.
  51. Lv Tang, Peng-Tao Jiang, Zhi-Hao Shen, Hao Zhang, Jin-Wei Chen, and Bo Li. Chain of visual perception: Harnessing multimodal large language models for zero-shot camouflaged object detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8805–8814. New York: ACM, 2024.
  52. Jun Ma, Sumin Kim, Feifei Li, Mohammed Baharoon, Reza Asakereh, Hongwei Lyu, and Bo Wang. Segment anything in medical images and videos: Benchmark and deployment. *arXiv preprint arXiv:2408.03322*, 2024.
  53. Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations*. Retrieved May 5, 2025, from <https://openreview.net/forum?id=Bkg6RiCqY7>, 2019.
  54. Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8779–8788. Piscataway: IEEE, 2019.
  55. Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. BASNet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7479–7489. Piscataway: IEEE, 2019.
  56. Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3907–3916. Piscataway: IEEE, 2019.
  57. Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. PraNet: Parallel reverse attention network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 263–273. Cham: Springer, 2020.
  58. Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6024–6042, 2021.
  59. Ge-Peng Ji, Yu-Cheng Chou, Deng-Ping Fan, Geng Chen, Huazhu Fu, Devesh Jha, and Ling Shao. Progressively normalized self-attention network for video polyp segmentation. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Proceedings of the 24th International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 142–152. Cham: Springer, 2021.
  60. Pengxiang Yan, Guanbin Li, Yuan Xie, Zhen Li, Chuan Wang, Tianshui Chen, and Liang Lin. Semi-supervised video salient object detection using pseudo-labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7284–7293. Piscataway: IEEE, 2019.