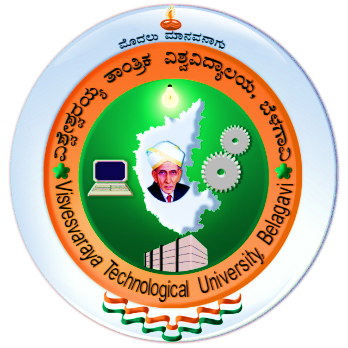
**VISVESVARAYA TECHNOLOGICAL UNIVERSITY**

“Jnana Sangama”, Belagavi -590018, Karnataka

****

**A Technical Seminar Report on**

“**SEGMENT ANYTHING MODEL2**”

Submitted in the partial fulfilment of the requirement for the award of the Degree of

**BACHELOR OF ENGINEERING**

**in**

**Artificial Intelligence and Data Science**

Submitted by

**Abhishek K N 4MN21AD002**

Under the guidance of

**Prof. Gagana M S**

**Assistant professor, Dept. of AI & DS**

**MIT Thandavapura**

****

Maharaja Institute of Technology Thandavapura

Department of Artificial Intelligence and Data Science

NH 766, Nanjangud taluk, Mysore district-571302

**2024-2025**

**MAHARAJA INSTITUTE OF TECHNOLOGY**

**THANDAVAPURA**

NH 766, Nanjangud Taluk, Mysuru-571 302

(An ISO 9001:2015 and ISO 21001:2018 Certified Institution)

(Affiliated to VTU, Belagavi and approved by AICTE, New Delhi)

**DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE**

# CERTIFICATE

Certified that the seminar work entitled **“****SEGMENT ANYTHING MODEL2”** is a bonafide work carried out by **ABHISHEK K N** (4MN21AD002) for the TECHNICAL SEMINAR with the course code 21AD81 of Eighth Semester in Artificial Intelligence and Data Science under Visvesvaraya Technological University, Belagavi during the academic year 2024-25. It is certified that all corrections/suggestions indicated for Internal Assignment have been incorporated in the report. The report has been approved as it satisfies the course requirements.

|  |  |  |
| --- | --- | --- |
| **---------------------------------**  **Signature of the Guide**  **Prof. Gagana M S**  Assistant Professor  Dept. of AI & DS | **---------------------------------**  **Signature of the HOD**  **Dr. Swarnalatha K**  Assistant Professor & HOD  Dept. of AI & DS | **---------------------------------**  **Signature of the Seminar Coordinator**  **Prof. Madhan Kumar G S**  Associate Professor  Dept. of AI & DS |

# ACKNOWLEDGMENTS

The success of any endeavour depends on clear goals and consistent guidance. I sincerely thank everyone who directly or indirectly contributed to the successful completion of my project.

I am grateful to the **Maharaja Education Trust®,** **Mysuru**, for providing me the opportunity to pursue my Bachelor’s degree in this college.

I express my sincere thanks to **Dr. Chethan H K**, Joint Secretary of Maharaja Institute of Technology, Thandavapura, for his constant support and appreciation during my project work.

I extend my heartfelt thanks to **Dr. Y T Krishne Gowda**, Principal, Maharaja Institute of Technology Thandavapura, for his unwavering cooperation throughout my academic journey. His support greatly contributed to the successful completion of my project.

I express thanks to **Dr. Swarnalatha K**, Associate Professor and Head of the Department of Artificial Intelligence and Data Science, for providing invaluable support throughout the period of the project work.

I express my deepest sense of gratitude to my guide **Prof. Gagana M S**, Assistant Professor, and my coordinator **Prof. Madhan Kumar G S**, Assistant Professor, Department of Artificial Intelligence and Data Science, for their valuable guidance and suggestions during the project work.

I extend my gratitude to all the teaching and non-teaching staff members of the department and am deeply indebted to MITT for their direct and indirect support throughout my project work.

Finally, I thank the Almighty and express my gratitude to my parents for their faith, love, and kindness in believing in me.

Abhishek K N (4MN21AD002)

# DECLARATION

I, **ABHISHEK K N** [4MN21AD002], student of 8th semester Artificial Intelligence and Data Science, Maharaja Institute of Technology Thandavapura, here by declare that the seminar entitled “**SEGMENT ANYTHING MODEL2**” submitted to the Visvesvaraya Technological University, Belgum during the academic year 2024-25, is a record of an original work done by me under the guidance of **Prof. GAGANA M S**, Assistant Professor, Department of Artificial Intelligence and Data Science, Maharaja Institute of Technology, Thandavapura. This seminar dissertation report is submitted in partial fulfilment for the award of Artificial Intelligence and Data Science. The results embodied in this report have not been submitted to any other University or Institute for the award of any degree.

ABHISHEK K N

4MN21AD002

# Abstract

In the rapidly evolving field of computer vision, the ability to segment visual content accurately and efficiently across diverse contexts is a crucial task. The Segment Anything Model (SAM) introduced a new paradigm of promptable segmentation, allowing a model to respond to user inputs like points or boxes to generate object masks. However, despite its versatility and strong zero-shot performance on static images, SAM had limitations in understanding dynamic scenes and handling video content effectively. To address these challenges, Meta AI introduced Segment Anything Model 2 (SAM2) a more advanced and capable successor. SAM2 introduces several key innovations, including a redesigned architecture for better generalization, improved efficiency, and the introduction of a new dataset called SA-V (Segment Anything in Video). This dataset, with over 26 million masks across 1,100 hours of video, enables SAM2 to perform high-quality video segmentation in addition to static image tasks. This report explores the architecture, working principles, applications, and benefits of SAM2, highlighting its capabilities and improvements over the original SAM. It also discusses the future potential of promptable segmentation models and their transformative role in fields such as medical imaging, robotics, autonomous vehicles, and content creation.

# Contents

[CERTIFICATE i](#_Toc196039991)

[ACKNOWLEDGMENTS ii](#_Toc196039992)

[DECLARATION iii](#_Toc196039993)

[Abstract iv](#_Toc196039994)

[List of Figures vi](#_Toc196039995)

[Chapter 1 1](#_Toc196039996)

[Introduction 1](#_Toc196039997)

[1.1 Image Segmentation Overview 1](#_Toc196039998)

[1.2 Evolution of Segmentation Models 1](#_Toc196039999)

[1.3 Rise of Foundation Models 2](#_Toc196040000)

[1.4 Introduction to SAM (Segment Anything Model) 2](#_Toc196040001)

[1.5 Limitations of SAM 3](#_Toc196040002)

[1.6 Motivation for SAM2 3](#_Toc196040003)

[Chapter 2 4](#_Toc196040004)

[Architecture and Working 4](#_Toc196040005)

[2.1 Overview of SAM2 Architecture 4](#_Toc196040006)

[2.2 Working of SAM2 6](#_Toc196040007)

[2.3 Task Tokens and Promptable Design 7](#_Toc196040008)

[2.4 Temporal Integration and Video Segmentation 7](#_Toc196040009)

[2.5 Training with SA-V Dataset 8](#_Toc196040010)

[Chapter 3 9](#_Toc196040011)

[Applications 9](#_Toc196040012)

[3.1 Overview of Applications 9](#_Toc196040013)

[3.2 Real-Time Video Object Segmentation 9](#_Toc196040014)

[3.3 Medical Imaging 10](#_Toc196040015)

[3.4 Content Creation and Editing 11](#_Toc196040016)

[3.5 Web and App Integration 11](#_Toc196040017)

[Chapter 4 12](#_Toc196040018)

[Advantages and Disadvantages 12](#_Toc196040019)

[4.1 Advantages 12](#_Toc196040020)

[4.2 Disadvantages 13](#_Toc196040021)

[Conclusion and Future Enhancement 14](#_Toc196040022)

[References 15](#_Toc196040023)

# List of Figures

| **Figure No.** | **Title** | **Page No.** |
| --- | --- | --- |
| Figure 1.1 | Types of Image Segmentation | 1 |
| Figure 1.2 | SAM performing prompt-based segmentation | 2 |
| Figure 2.1 | Architecture of SAM2 | 4 |
| Figure 2.2 | Example videos from the SA-V dataset with masklets overlaid | 8 |
| Figure 3.1 | Examples from SAM 2 zero-shot video benchmark suite | 9 |
| Figure 3.2 | Real-Time Video Object Segmentation | 10 |
| Figure 3.3 | Medical Imaging Segmentation | 10 |
| Figure 3.4 | Content Creation and Editing | 11 |

# Chapter 1

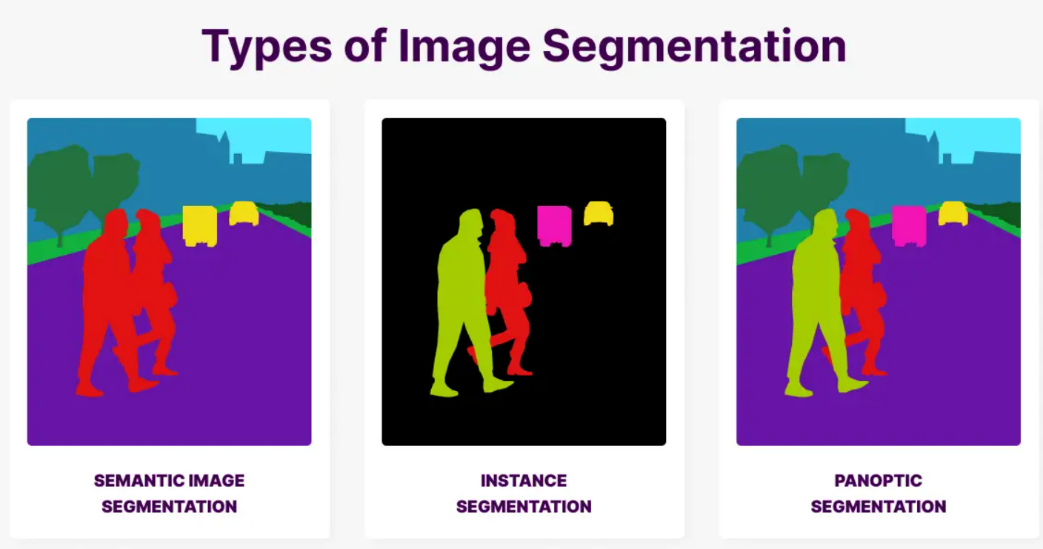
# Introduction

## 1.1 Image Segmentation Overview

Image segmentation is a fundamental task in computer vision that involves partitioning an image into multiple segments or regions, each corresponding to specific objects or areas of interest. Unlike classification or object detection, segmentation offers pixel-level understanding, enabling machines to precisely locate and identify boundaries of objects.

There are three primary types of image segmentation:

* **Semantic Segmentation** – classifies each pixel into a category (e.g., road, car, person).
* **Instance Segmentation** – identifies individual instances of objects (e.g., two different cars).
* **Panoptic Segmentation** – combines semantic and instance segmentation into a unified output.



**Figure 1.1: Types of Image Segmentation**

## 1.2 Evolution of Segmentation Models

The development of segmentation models has evolved through various stages, each improving on generalization and accuracy:

* **Classical methods** like edge detection, thresholding, and region growing.
* **CNN-based models** such as U-Net, DeepLab, and FCN that introduced deep learning into segmentation.
* **Instance segmentation models** like Mask R-CNN, combining object detection with pixel-wise masks.
* **Vision Transformers (ViTs)** and large-scale models brought improvements in long-range context understanding.

However, most of these models are trained for **specific datasets and tasks**, and often fail in generalizing across domains without fine-tuning.

## 1.3 Rise of Foundation Models

Foundation models in vision, like CLIP, DINO, and SAM, are trained on large-scale datasets and designed to be general-purpose. They bring:

* Scalability across diverse tasks
* Zero-shot capabilities
* Prompt-based input mechanisms

SAM (Segment Anything Model), developed by Meta AI, was a milestone in building such a universal segmentation model.

## 1.4 Introduction to SAM (Segment Anything Model)



**Figure 1.2: SAM performing prompt-based segmentation**

SAM introduced the concept of **promptable segmentation**, where the model could generate masks based on:

* **Point prompts** (foreground/background)
* **Bounding boxes**
* **Free-form masks**

SAM was trained on the **SA-1B dataset**, which consists of 1 billion masks across 11 million images. Its ability to generalize to unseen images made it revolutionary.

## 1.5 Limitations of SAM

Despite its impressive performance, SAM has several limitations:

* Trained primarily on static images (not optimized for video data).
* Unable to track or understand objects in motion over time.
* High computational requirements for inference.
* Dependent on the type and accuracy of the prompts provided.
* Lacks temporal consistency in outputs when applied frame-by-frame in videos.

These limitations highlighted the need for a new approach that extends SAM’s capabilities to handle videos and more complex real-world tasks.

## 1.6 Motivation for SAM2

To overcome these limitations, **Meta introduced SAM2**, designed to:

* Incorporate **temporal understanding** via video training.
* Work with both image and video data seamlessly.
* Improve **prompt efficiency and flexibility**.
* Leverage a new dataset: **SA-V (Segment Anything in Video)** – over 1,100 hours of video and 26M masks.

SAM2 pushes the boundary of general-purpose segmentation from still images to dynamic video scenes, making it more practical and powerful in real-world applications like medical diagnostics, autonomous driving, and video editing.

# Chapter 2

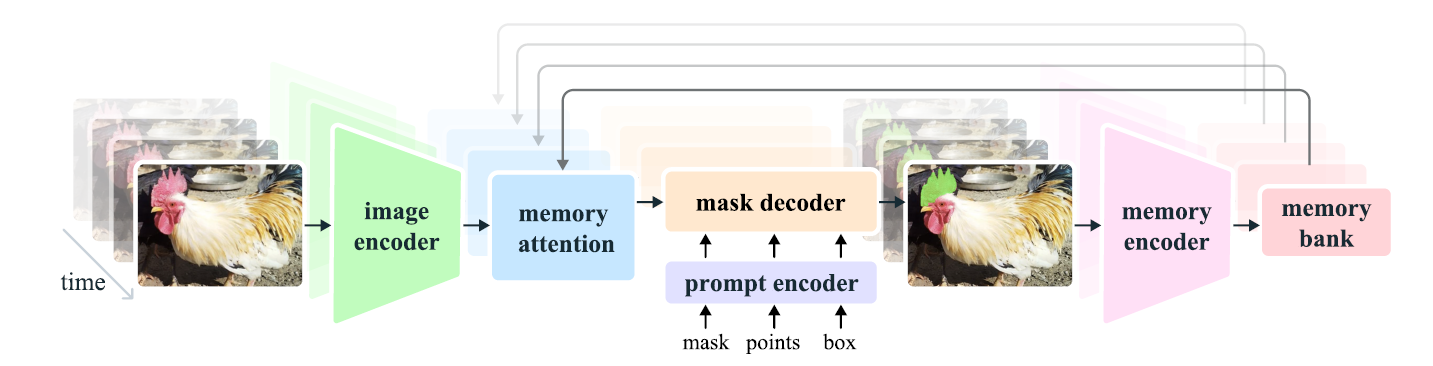
# Architecture and Working

## 2.1 Overview of SAM2 Architecture

SAM2 (Segment Anything Model 2) is a powerful extension of the original SAM that enhances generalization, scalability, and applicability to both **images and videos**. The key goal of SAM2 is to build a **unified promptable segmentation model** that works efficiently across a wide variety of visual inputs — including video frames with motion and changing contexts.

**Key Features of SAM2 Architecture:**

* Modular and promptable design
* Trained using both **static image data (SA-1B)** and **video data (SA-V)**
* Optimized for **both interactive (user-prompted)** and **automatic segmentation**
* Capable of producing **temporally coherent masks** across video frames



**Figure 2.1: Architecture of SAM2**

**Core Components of SAM2:**

1. **Image Encoder:** Processes the current frame and extracts spatial feature embeddings. These features are passed to both the Memory Attention and the Memory Encoder modules.
2. **Memory Attention:** Integrates relevant past information from earlier frames using the Memory Bank. Helps maintain temporal consistency in segmentation across video sequences.
3. **Prompt Encoder:** Embeds input prompts like points, boxes, or masks. These embeddings guide the mask generation process, ensuring the decoder focuses on user-specified regions.
4. **Mask Decoder:** Combines outputs from the Image Encoder, Prompt Encoder, and Memory Attention.Predicts segmentation masks for the current frame, guided by task tokens and prompt types.
5. **Memory Encoder:** Encodes current frame’s segmentation-related information and updates the Memory Bank. Allows the model to “remember” key features across time.
6. **Memory Bank:** Stores temporally encoded features across multiple past frames.Continuously updated and used for attention-based temporal modeling.

**Key Innovations:**

* **Promptable in Video**: Just like SAM worked with input prompts for images, SAM2 does so with videos.
* **Memory Modules**: Support long-term understanding and consistent segmentation.
* **Unified Framework**: No need for separate architectures for image and video segmentation.

**What’s New Compared to SAM?**

|  |  |  |
| --- | --- | --- |
| **Feature** | **SAM** | **SAM2** |
| Input Type | Static Images | Images & Videos |
| Prompt Support | Points, boxes, masks | Points, boxes, masks + task tokens |
| Dataset | SA-1B (Images) | SA-1B + SA-V (Videos) |
| Segmentation | One frame at a time | Multi-frame temporal segmentation |
| Promptable? | Yes | Yes |
| Generalization | Strong for images | Strong for images **and videos** |

**Table 2.1: Differences between SAM and SAM2**

## 2.2 Working of SAM2

**2.2.1 Image Encoder**

The image encoder transforms each input frame into a spatially-rich feature map. SAM2 employs a **Vision Transformer (ViT-H/16)** to perform this task.

**Working Mechanism:**

* **Patch Embedding**: The image is divided into small patches, each of which is flattened and embedded.
* **Position Encoding**: Positional information is added to maintain the spatial layout.
* **Transformer Layers**: Multiple self-attention layers compute contextual relationships among patches.

This module produces high-resolution feature maps that are forwarded to the **Mask Decoder** and **Memory Attention** modules.

**2.2.2 Prompt Encoder**

The Prompt Encoder converts input hints (prompts) into a vector representation that guides segmentation. These prompts help focus on specific areas or objects.

**Types of Prompts:**

* **Points**: User-specified positive/negative points.
* **Boxes**: Bounding boxes around regions of interest.
* **Masks**: Previously segmented masks.
* **Task Tokens**: Special embeddings denoting task types (image or video).

Each prompt is encoded using dedicated neural layers and fused with visual features.

**2.2.3 Mask Decoder**

The Mask Decoder is responsible for predicting the final segmentation mask. It takes into account all available embeddings and contextual features.

**Features:**

* **Cross-attention**: Incorporates visual, memory, and prompt information.
* **Hierarchical Reasoning**: Combines global and local context.
* **Multi-Mask Output**: Produces several mask hypotheses and selects the best using confidence scoring.

This module is the core decision-making unit of SAM2.

## 2.3 Task Tokens and Promptable Design

**2.3.1 What Are Task Tokens?**

Task tokens are learned embeddings that specify the type of task the model should perform. They act like switches within the network that tune its behavior.

**2.3.2 Types of Task Tokens:**

* T\_img: Static image segmentation
* T\_vid: Video segmentation
* T\_auto: Automatic segmentation without any prompt
* T\_interactive: Interactive segmentation with user prompts

This modular approach enables the same network to handle diverse segmentation tasks without retraining or architecture changes.

## 2.4 Temporal Integration and Video Segmentation

SAM2’s standout capability lies in its effective video segmentation. Unlike traditional models, it does not treat video frames independently. Instead, it utilizes memory to maintain coherence over time.

**2.4.1 Components Involved:**

* **Frame Encoder**: Processes each video frame.
* **Memory Encoder**: Converts current segmentation info into a memory-friendly format.
* **Memory Bank**: Stores memory vectors from past frames.
* **Memory Attention**: Attends to relevant past frames and injects temporal context.

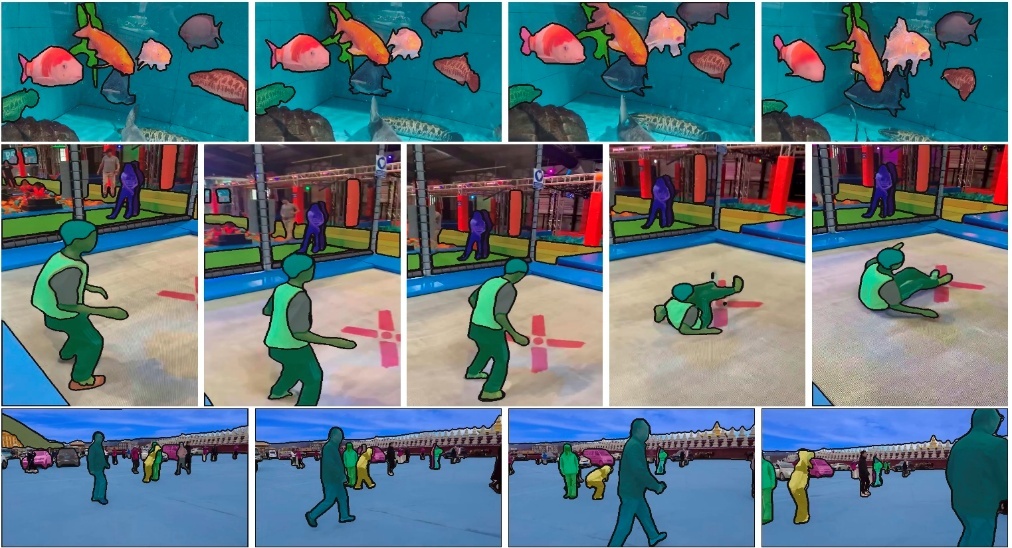
**2.4.2 Benefits:**

* **Temporal Coherence**: Avoids flickering or unstable masks.
* **Efficient Tracking**: Capable of tracking the same object across frames.
* **Robustness**: Handles occlusion and motion blur effectively.

This mechanism allows SAM2 to excel in video understanding tasks.

## 2.5 Training with SA-V Dataset

The training of SAM2 is powered by the **SA-V dataset**, which is a large-scale video segmentation dataset curated specifically to train promptable models.



**Figure 2.2: Example videos from the SA-V dataset with masklets overlaid**

**2.5.1 Dataset Characteristics:**

* Over 1,100 hours of annotated video
* More than 26 million segmentation masks
* Includes diverse settings: indoor, outdoor, static, dynamic, day/night

**2.5.2 Purpose:**

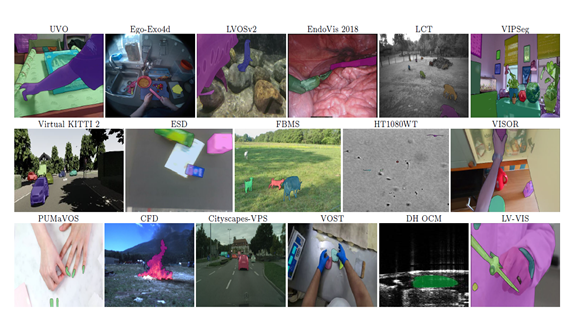
* Enhance model's generalization to diverse visual content
* Ensure robustness to lighting, occlusions, and object transformations
* Provide consistent labeling across frames

# Chapter 3

# Applications

## 3.1 Overview of Applications

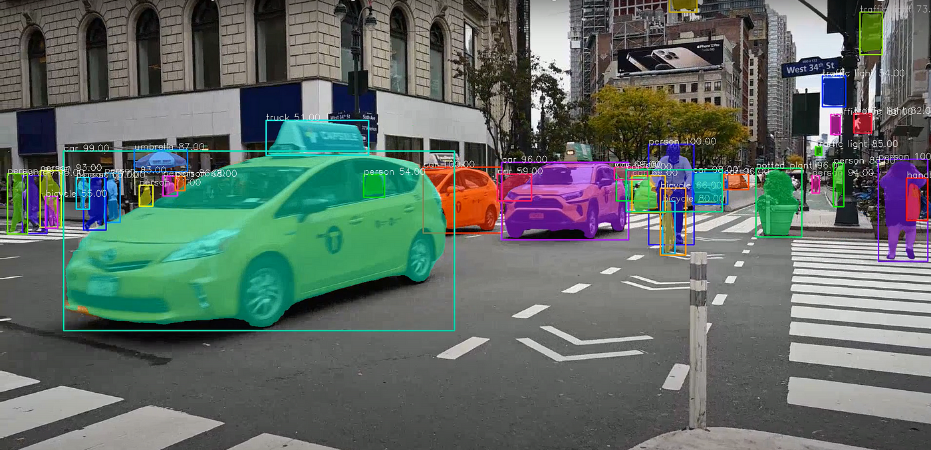
Segment Anything Model 2 (SAM2) extends the capabilities of the original SAM by enabling promptable segmentation across both static images and video sequences. Its robust architecture allows it to adapt to a wide variety of real-world use cases in computer vision. The ability to process both spatial and temporal information makes SAM2 particularly suitable for complex applications where precision, generalization, and user interactivity are essential.



**Fig 3.1: Examples from SAM 2 zero-shot video benchmark suite.**

## 3.2 Real-Time Video Object Segmentation

SAM2 supports video object segmentation in real-time, enabling dynamic scene understanding. This capability finds widespread application in several areas. In autonomous driving, SAM2 is used to identify moving objects such as pedestrians, vehicles, road signs, and traffic lanes, ensuring safer and more accurate decision-making by self-driving systems. Surveillance systems benefit from SAM2 by segmenting and tracking individuals or objects across different camera feeds, providing a consistent identity match. In sports analytics, it tracks players, balls, or specific gameplay elements, helping generate insightful data for performance analysis and strategic planning.

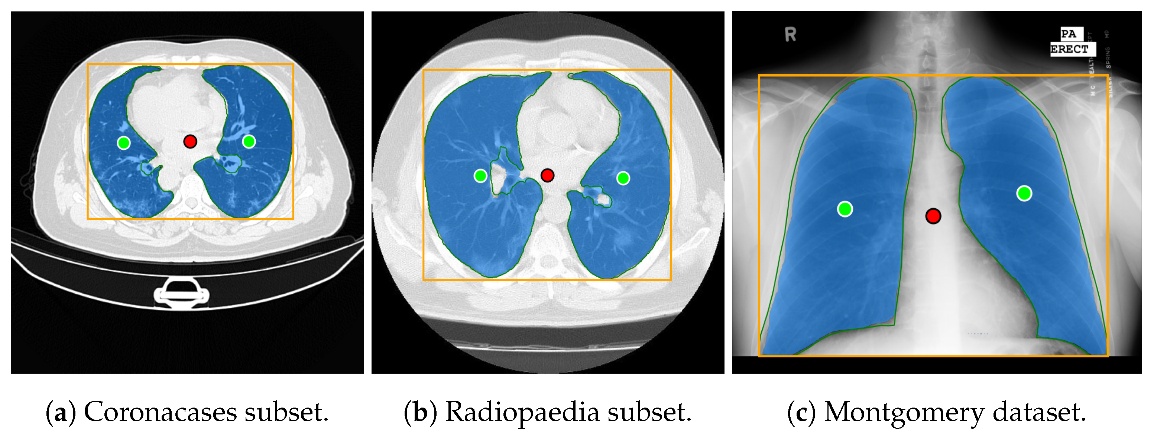


**Figure 3.2: Real-Time Video Object Segmentation**

By maintaining object consistency across video frames, handling occlusions and camera movements effectively, and minimizing flickering through memory-based temporal coherence, SAM2 has set a new standard in real-time video object segmentation.

## 3.3 Medical Imaging

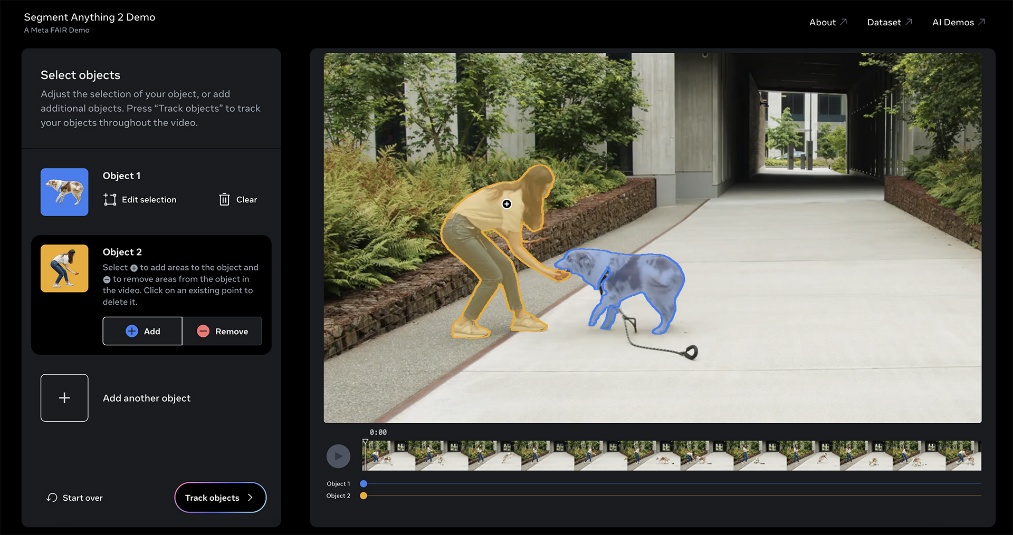
Promptable segmentation proves essential in the medical field, where precision and flexibility are critical. Clinicians and radiologists often need to manually highlight regions of interest in complex imaging data. SAM2 assists in segmenting tumors, organs, and lesions in MRI or CT scans using user-provided prompts. It plays a pivotal role in surgical planning by supporting the delineation of anatomical regions for augmented reality systems. Additionally, in cell biology, SAM2 enables real-time tracking of individual cells in high-resolution microscopy videos.



**Figure 3.3: Medical Imaging Segmentation**

This results in reduced manual effort, improved accuracy, and the ability to offer real-time feedback. Furthermore, it supports quantitative research in medicine by streamlining data annotation processes.

## 3.4 Content Creation and Editing



**Figure 3.4: Content Creation and Editing**

One major application is background removal, where SAM2 isolates the subject from the background in images and videos, allowing for seamless editing. It also enables object tracking, which helps apply visual effects that follow specific subjects across frames. Additionally, interactive annotation tools built using SAM2 provide intelligent, real-time segmentation features that respond to user interactions with high precision. These applications significantly enhance creativity and reduce time spent on manual editing tasks.

## 3.5 Web and App Integration

SAM2’s promptable nature allows for seamless integration into web and mobile applications. AI image editing apps use SAM2 to isolate objects, remove or replace backgrounds, and apply layer-based edits interactively. On social media platforms, real-time segmentation enables dynamic filters that respond to people’s movements. Video conferencing applications employ SAM2 to provide background blur or virtual backgrounds by segmenting the speaker in real time. These integrations enhance user experience through intelligent visual processing.

# Chapter 4

# Advantages and Disadvantages

## 4.1 Advantages

**4.1.1 Generalized Segmentation**

SAM2 supports zero-shot and few-shot learning paradigms, meaning it can perform well across a wide range of segmentation tasks without retraining. This makes it a versatile solution that can be deployed in diverse scenarios.

**4.1.2 Promptable Interface**

One of SAM2’s standout features is its ability to respond to prompts such as points, boxes, or text. This allows users to interactively guide the segmentation process, making it more intuitive and adaptable to specific tasks.

**4.1.3 Temporal Consistency**

Unlike traditional segmentation models that work on individual frames, SAM2 processes video sequences while maintaining temporal coherence. This significantly reduces flickering and inconsistency, especially in dynamic environments like video editing or surveillance.

**4.1.4 High Accuracy and Speed**

SAM2 achieves state-of-the-art results on several benchmark datasets. Its dual encoder-decoder structure and memory-efficient attention mechanisms contribute to both fast inference times and high segmentation accuracy.

**4.1.5 Scalability and Versatility**

SAM2 is suitable for multiple domains—ranging from autonomous vehicles to agriculture—due to its scalable architecture. It handles both high-resolution static images and low-resolution video streams with equal efficiency.

## 4.2 Disadvantages

**4.2.1 High Computational Requirements**

Despite its optimized design, SAM2 still demands significant computational resources, especially for video processing. Running SAM2 in real-time on edge devices remains a challenge.

**4.2.2 Limited Understanding of Semantics**

While SAM2 excels at boundary detection and object separation, it lacks deep semantic understanding. For example, it may segment a dog perfectly but not understand whether it is running, sitting, or playing.

**4.2.3 Training Complexity**

Although users benefit from a pre-trained model, training SAM2 from scratch is complex due to the massive size of the datasets and the intricate architecture. Fine-tuning also requires careful calibration.

**4.2.4 Prompt Sensitivity**

The model’s output can be highly sensitive to the type and placement of prompts. Small variations in prompt input can lead to significantly different segmentation results, affecting reliability in critical applications.

**4.2.5 Dataset Bias**

As with most deep learning models, SAM2's performance can be influenced by the data it was trained on. If underrepresented scenarios are present in a given domain, the model may not generalize well.

# Conclusion and Future Enhancement

**Conclusion**

The Segment Anything Model 2 (SAM2) signifies a major leap forward in the field of computer vision, particularly in segmentation tasks. By combining high generalization, promptable interfaces, and temporal consistency, SAM2 addresses key limitations found in earlier models. Its versatility across domains like healthcare, surveillance, agriculture, and media editing proves its broad potential.

However, SAM2 is not without challenges. It demands high computational resources and has limitations in semantic understanding and prompt stability. These shortcomings provide opportunities for further research and development. Still, SAM2 establishes a strong foundation for future segmentation models and paves the way for more interactive, intelligent, and scalable visual systems.

**Future Enhancements**

* **Efficient Edge Deployment**: Future work could aim to optimize SAM2 for mobile and embedded systems, allowing real-time inference on edge devices without compromising accuracy.
* **Deeper Semantic Integration**: Integrating multimodal understanding through language and context-aware prompts could enhance SAM2’s ability to distinguish between object behaviors and roles.
* **Adaptive Prompt Tuning**: Improving the model’s responsiveness to prompts using self-correcting mechanisms or reinforcement learning could address sensitivity and variability.
* **Expansion of Datasets**: Enriching training datasets to cover more diverse domains, lighting conditions, and environments can help reduce bias and improve generalization.
* **Cross-domain Transfer Learning**: SAM2 can be extended through transfer learning approaches that allow it to adapt quickly to new domains with minimal data.

# References

1. Meta AI. (2024). [Segment Anything Model 2 (SAM2)](https://ai.meta.com/SAM2/)
2. Meta AI Blog. (2024). [Introducing SAM 2: The next generation of Meta Segment Anything Model for videos and images](https://ai.meta.com/blog/segment-anything-2-video/)
3. Singh, J. (2024). SAM 2: Meta’s New Object Segmentation Model
4. Emanuilov, S. (2024). SAM 2 — The next generation of Meta’s Segment Anything Model for videos and images
5. Zephyrnet. (2024). [Meta's SAM-2: The Future Of Real-Time Visual Segmentation](https://zephyrnet.com/metas-sam-2-the-future-of-real-time-visual-segmentation/)
6. UnfoldAI. (2024). [SAM 2 — The next generation of Meta's Segment Anything Model for videos and images](https://unfoldai.com/meta-sam-2/)
7. Tech Driven Future. (2024). [Meta SAM 2: Structure, Functions & Limitations](https://technodrivenfuture.com/big-data/meta-sam-2-structure-functions-limitations/)
8. Ravi, N., et al. (2024). [SAM 2: Segment Anything in Images and Videos](https://arxiv.org/abs/2408.00714)