Object segmentation for Stretch 3 Robot Using SAM2

Abhinay Karade *IIT2022170* | Sachin Sarkar *IIT2022192* | Sambhav Sinha *IIT2022194* | Jayant Vinjamury *IIT2022218* | Dinesh Pradhan *IIT2022219*

***Abstract*—Major degradation in segmentation performance comes from several factors: high levels of noise, occlusions, and complex environments. Thus, tasks like navigation and interaction become considerably challenging. Using its excellent zero-shot capabilities and applying interactive segmentation techniques, this project aims to improve image segmentation using the addition of Segment Anything Model 2 (SAM 2) to the vision system of Stretch3 robot. Combining advanced memory mechanisms and promptable visual segmentation, SAM 2 effectively challenges the issues presented in video and image segmentation more than classical models by attaining more accurate adaptable segmentation and conserving efficient computation with benchmark datasets and real-world scenarios. Based on the results obtained, SAM 2 greatly enhances segmentation accuracy, especially in previously unseen object or dynamic environments, and therefore is highly suited for use in robotics, surveillance, and augmented reality, among other application domains. The present project emphasizes the fact that elaborate segmentation is necessary in maintaining the reliability of robotic systems operationally.**

I. Introduction

The basic purpose of this project is to integrate the Segment Anything Model 2 (SAM 2) with the STRETCH3 robot for It improves the segmentation capability of the latter on images. This requires the creation of SAM 2, a model at state-of-the-art level. It is current for accuracy and flexibility in the current one. camera-based perception system of STRETCH3. The goal Develop an objective to better improvement of the robot in object segmentation. This brings about improved performance in complex and dynamic situations. recognition and interaction. From the following comparison, SAM2 and the currently being used YOLO model, the project will thus prove SAM 2 in every respect, and quality, flexibility and strength, which should improve the The overall performance of the STRETCH3 robot.

A. Limitations of YOLO in Precision, Adaptability, and Unseen Object Segmentation in the Current Stretch 3 Setup

1) Limitations of YOLO in Fine Grain Precision: YOLO is famous for fast and efficient detection of objects and is not intrinsically designed for segmentation assignments involving pixel-level Precision is especially highlighted, for instance, in surgical labs or chemistry setups, where accurate boundary identification is critical. The lack of segmentation-focused design constrains Effectiveness of YOLO in those use cases.

2) Difficulty in Adapting to New or Unseen Scenarios: The effectiveness of YOLO depends highly on the data it has been exposed to. In response to new objects or settings, its accuracy often decreases. For Stretch 3, Functions within ever-changing and specialized contexts. Accordingly, this limitation may result in missing detections or false Classifications.

3) Depth Information Ineffectiveness: While Stretch 3 relies on the Intel RealSense D435 to capture depth information, YOLO's model architecture does not inherently make This research methodology use implies that YOLO is not fully sensible of spatial Relationships represent a crucial capability necessary for depth-awareness. tasks in 3D domains.

4) Underperformance on Domain-Specific Segmentation Tasks: Despite its enhancements, the framework of YOLO is not at all as fit for segregation works as more model specific, namely SAM 2. To illustrate, It was unable to rank YOLO v3 with the data of surgical images. adapt the necessary solution to adequately explain Complex areas, hence proved inappropriate for such Uses.

5) Lack of Built-In Segmentation Capability: By definition, YOLO v3 was written as an object-detection framework. rather than an object-segmentation framework. This is inherently limits it to applications where clear object boundary detection is therefore unnecessary. More recent revisions, such as YOLO v11, focus more on enhancing detection while leaving segmentation an ancillary property.

Confronting the aforementioned constraints in the Stretch 3 setup focus on underlining the significance of a supplementary fusion. Using SAM 2 for segmentation and Keeping YOLO as a detector means that this system has the best opportunity for more effective exploitation of its ecosystem, therefore releasing advanced robotics capability. abilities to interpret complex, real-world situations.

B. Using SAM 2 for improved segmentation precision and adaptability to diverse environments

SAM 2 is a more advanced version of an image segmen tation model, which outperforms traditional models such as YOLO due to the higher precision level and ability to be adaptable to different environments. This is how SAM 2

enhances the features of STRETCH3 to increase accuracy and versatility in its segmentation:

1) Increased Precision: SAM 2 achieves pixel-level seg presentation and therefore potentially more precise recognition performance. contours of object against the background of YOLO relies upon The bounding boxes provide better and more accurate object outline or delimitation. those especially the detailed object activties reference recognition, such as gripping small objects, picking up irregular bulky items, navigating detailed obstacles, or delicate goods. SAM 2 can work well with environments where details matter because because it can very well define objects with sharp edges.

2) Promptable Segmentation: One of the most recognizable A notable characteristic of SAM 2 is its ability to perform promptable segmentation. Such a capsule can take points, bounding boxes, or textual descriptions to guide segmentation. It is, therefore, highly She can adjust to many types of activities and environments. As an example, STRETCH3 may use SAM 2 to refine results from segmentation based on user inputs or dynamic changes In an environment, real-time decision-making is improved.

3) Adaptation to Novel and Unseen Objects: Unlike YOLO, it should be trained on new object categories SAM 2 can generalise to novel, unseen objects with minimal adjustment. This adaptation is critical for STRETCH3 in contexts in which objects may fluctuate frequently or comprise novel items that weren't part of the training data. SAM 2 can precisely categorize these unknown entities, thus reducing the need for constant retraining or model updates.

4) Complex Scenes: SAM 2 does an quite good job of complex scenes, with objects obscured, and overlap within. or is arranged in non-standard orientations. With its stronger algorithms, SAM 2 can detect and classify objects even in situations when or partly veiled, or nearly touching. This skill Give it an edge in extremely fluid settings where objects do tend to change location or position often.

5) Efficiency and Flexibility: Whereas SAM 2 increases segmentation accuracy, it does so efficiently by balancing the demand for precision with fast computation times, which is essential to deploy in real-time robotics applications. Given its general-purpose nature, it can be applied for numerous tasks without task-specific retraining, thereby making it a flexible solution for the diverse applications of STRETCH3.

By injecting SAM 2, STRETCH3 will be able to recognize objects at more detailed levels and operated upon appropriately in dynamic and complex scenarios, thus being more effective and autonomous.

II. Implementation Details

A. SAM 2 Features

The Segment Anything Model (SAM 2) integrates cutting-edge segmentation technologies, setting a benchmark for accuracy and adaptability. Below is a detailed exploration of its features and how they apply to the project's datasets:

1) Interactive Segmentation Using Points, Boxes, and Masks: SAM 2 enhances user control by supporting various interactive inputs:

a) Points: Key points allow for precise guidance, helping the model focus on specific regions within images.

b) Boxes: Bounding boxes define object regions, especially useful in complex datasets like LabPicsV1, where apparatus and chemical containers need isolation.

c) Masks: SAM 2 refines preexisting or auto-generated masks for pixel-perfect segmentation, ensuring high accuracy across diverse applications.

Fig. 1: INPUT IMAGE

Fig. 2: FINAL IMAGE MASK

2) Memory Mechanism for Video Segmentation: SAM 2's memory mechanism ensures consistent segmentation in video sequences

a) : Tracks objects across frames, ideal for dynamic lab settings captured using LabPicsV1.

b) : Maintains segmentation integrity even with shifting perspectives, complex interactions, or motion blur. This feature is invaluable for robotics and real-time analysis applications.

3) Dataset and Training Data Specifics: The robustness of SAM 2 stems from its training on diverse datasets and its fine-tuning capabilities for specialized datasets, such as

a) LabPicsV1: A chemistry lab dataset containing intricate apparatus and experimental setups. SAM 2's segmentation accurately isolates lab components, demonstrating its capability in structured environments.

b) BRAIN-TUMOR.v1i.yolov11: A medical imaging dataset focusing on brain tumor segmentation. Fine-tuning SAM 2 on this dataset highlights its adaptability to healthcare applications, showcasing precise segmentation even in highly detailed and sensitive imagery.

Fig. 3: FINE TUNING CODE

4) Zero-Shot Segmentation Capabilities and Results: SAM 2's zero-shot segmentation capabilities allow it to generalize to unseen data:

a) Performance: Without prior fine-tuning, SAM 2 achieves remarkable segmentation accuracy, identifying and isolating objects with minimal user input.

b) Applications: Its application in LabPicsV1 effectively segments experimental setups without requiring task-specific adjustments. Similarly, for BRAIN-TUMOR.v1i.yolov11, SAM 2 identifies tumor regions with high confidence, even in challenging cases.

c) Results: In comparative studies with YOLO v11, SAM 2 consistently outperformed in segmentation tasks, delivering results with greater precision and boundary accuracy.

By leveraging interactive tools, memory mechanisms, and zero-shot learning, SAM 2 has proven to be an unparalleled asset in diverse segmentation applications. Its performance on datasets like *LabPicsV1* and *BRAIN-TUMOR.v1i.yolov11* showcases its potential to revolutionize image analysis across domains.

B. Stretch3 Robot

The Stretch3 robot is a sophisticated mobile manipulation platform designed for both research and practical applications in fields such as robotics, automation, and assistive technologies. Equipped with a mobile base, a manipulator arm, and a variety of sensors, Stretch3 is capable of navigating and interacting with its environment autonomously. Its versatile design makes it well-suited for dynamic and cluttered environments, such as laboratories, offices, and homes, where precise manipulation and mobility are required.

Operating on the Robot Operating System (ROS), an open-source middleware framework, Stretch3 benefits from a modular and flexible approach to robotic system development. ROS provides a comprehensive suite of software libraries and tools that facilitate the creation of robust robotic applications. It supports communication between various robot components, ranging from low-level hardware control to high-level task execution. ROS nodes communicate through messages, topics, and services, ensuring seamless integration of functionalities such as navigation, object detection, and manipulation. This architecture enables Stretch3 to perform a wide range of tasks, including obstacle avoidance, object interaction, and environmental mapping, with ease.

The ROS-based setup provides an ideal environment for the integration of advanced perception systems, such as object detection models, that enhance the robot's ability to understand and respond to its surroundings.

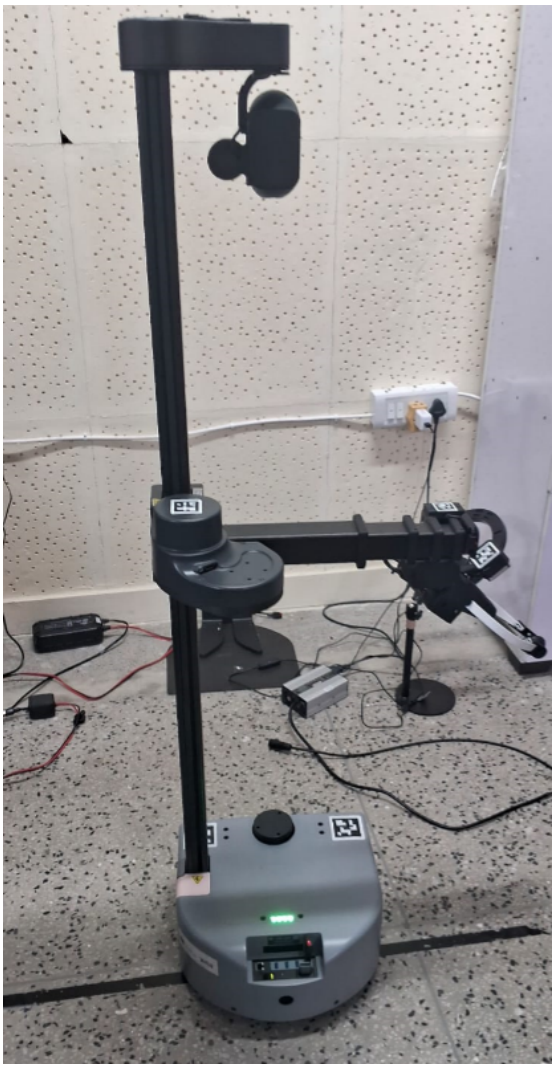

Fig. 4: STRETCH3 ROBOT

C. DataSet Acquisition

To evaluate and fine-tune the integration of SAM2 with the Stretch 3 robot for image segmentation, we carefully selected two datasets that offered diverse challenges for the model.

These datasets ensured comprehensive testing across different domains and scenarios.

1) Brain Tumor Dataset:

a) Overview: This dataset is widely used in medical research and comprises high-resolution images of brain scans. The images are detailed, often highlighting complex structures and irregular boundaries, making it ideal for testing the model's precision in segmentation.

b) Key Details:

- Number of Images: 1,671
- Image Resolution: 640x640 pixels

c) Purpose: The dataset was used to assess SAM2's ability to segment objects in a medical imaging context. It tests the model's capability to handle intricate patterns and differentiate fine details in the images.

2) Chemistry Lab Dataset:

a) Overview: This dataset features images of laboratory setups, including various objects like beakers, test tubes, and other equipment typically found in a chemistry lab. The scenes are often complex, with overlapping objects and diverse shapes, posing a significant challenge for segmentation.

b) Key Details:

- Number of Images: 7,900
- Image Resolution: 640x640 pixels

c) Purpose: This dataset was chosen to test SAM2 in scenarios with high object density and variability. It evaluates how well the model can segment multiple objects in cluttered and realistic environments.

3) Significance of the Datasets: These datasets complement each other by presenting distinct challenges:

- The Brain Tumor Dataset focuses on precision and detail in structured images, making it ideal for analyzing SAM2's accuracy in medical imaging.
- The Chemistry Lab Dataset adds complexity by introducing cluttered scenes with overlapping objects, helping evaluate the model's robustness in real-world applications.

4) Challenges and Preprocessing:

a) Challenges Encountered:

- Managing datasets with different contexts and ensuring the model adapts well to both.
- Maintaining consistent preprocessing steps across datasets to avoid introducing bias.

b) Preprocessing Steps:

- Standardized image dimensions where necessary.
- Normalized pixel intensities to align with SAM2's input requirements.
- Utilized these datasets to ensure a well-rounded evaluation of SAM2's segmentation capabilities, pushing its limits in diverse and demanding conditions.

D. Integration Process

1) Intel RealSense D435i Camera: The Intel RealSense D435i camera(used in STRETCH3)is a versatile depth camera designed for applications requiring precise 3D perception. Its key features include:

a) Depth Sensing: Utilizes stereo vision technology to generate high-resolution depth maps, enabling accurate detection of objects and their spatial positioning. Ideal for tasks that demand precise depth information, such as robotic navigation and object segmentation.

b) Wide Field of View (FOV): The D435 provides a field of view of approximately 85° x 58°, ensuring broad coverage for complex environments like labs or industrial setups.

c) RGB Camera: Equipped with a high-quality RGB sensor for capturing detailed images, supporting both depth-based and traditional segmentation tasks.

d) Compact and Lightweight Design: The camera's small form factor and ease of mounting make it ideal for robotic applications, including integration with Stretch3.

Fig. 5: Intel RealSense D435i

2) Implementation Details Using pyrealsense2: The pyrealsense2 library was pivotal in interfacing with the D435 camera. Key implementation highlights include:

a) Initialization: The library was used to initialize the camera, set resolution, and configure depth and RGB streams.

b) Depth and RGB Synchronization: Ensured that the depth and RGB data streams were synchronized for optimal segmentation results.

c) Image Processing : Captured images were preprocessed using standard techniques before being sent to SAM 2 for segmentation.

3) Steps Taken to Integrate SAM 2 with Stretch3: Integrating SAM 2 with the Stretch3 robot involved several stages to ensure seamless operation and precise object segmentation. The following steps outline the process:

a) Environment Setup: Two separate Python environments were created to handle compatibility issues:

- *Python 3.9* for camera access using pyrealsense2 (supported up to this version).
- *Python 3.12* for running SAM 2, which requires newer dependencies.

b) Hardware Configuration: The Stretch3 robot was equipped with an Intel RealSense D435i camera to capture high-quality depth and RGB images in real time.SAM 2 was configured to operate with the output images from the camera, enabling accurate object segmentation in various scenarios.

c) Image Capture Module: A custom script was developed using pyrealsense2 to interface with the Intel RealSense D435 camera. The captured images were then passed to SAM 2 for segmentation in the second environment, ensuring smooth interoperability between systems.

d) Data Flow: Real-time image data from the D435 camera was processed in Python 3.9 and seamlessly transferred to the Python 3.12 environment for segmentation. Results were evaluated to confirm the accuracy and relevance of the segmentation output in the context of Stretch3's operational requirements.

III. Evaluation Methodology

A. Metrics

1) Segmentation Precision: Segmentation performance was assessed using the Intersection over Union (IoU) metric. IoU is a common evaluation metric for segmentation tasks, given by the ratio of the overlapping area between the predicted segmentation mask and the ground truth mask to the total area covered by both masks. Mathematically,

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

IoU values lie between 0 and 1. The larger the IoU value is, the better the segmentation is. For both test cases

The results for the two datasets are as follows:

a) Chemistry Lab Dataset (LabPicsV1): LabPicsV1, SAM 2 was discovered to have an IoU of 0.6237. This signifies good success in the lab's intricate mechanical device and overlapping tools.

b) Brain Tumor Dataset (BRAIN-TUMOR.v1i.yolov11): SAM2 recorded an IoU of 0.7018 for the dataset. This demonstrates SAM Superior ability to segment detailed medical images with more precise boundary and higher accuracy

2) Resource Utilization: The experiments were conducted on a 4 GB GPU machine. For the experimental objective, it was up to the task to process image segmentation.

a) GPU Memory: SAM 2 utilized the GPU memory quite nicely and made sure that smooth performance was maintained even though the required computation was heavy.

b) Processing Time:

- Chemistry Lab Dataset: 1–2 seconds per image
- Brain Tumor Dataset: 1–2 seconds per image

All the results are demonstrating how SAM 2 can be used to deliver accurate and reliable segmentation results in resource-constrained environments, and therefore how it can be very a practical solution for many applications.

B. Test Setup

We are running object segmentation tests utilizing the SAM2 model on the STRETCH3 environment, under ideal laboratory conditions for the experiment. This setup ensures to have precise, efficient, and reproducible results. Detailed description of test scenarios is as follows:

1) Description Of Test Scenarios:

a) Lighting Conditions: The laboratory provides bright, even lighting with no shadows and objects have superior visibility. Illumination in every direction maintains constant illumination for the camera and lets it capture High-resolution images devoid of artifacts that might skew .Segmentation accuracy .

b) Object Diversity: The objects, whose performances are being evaluated are Of varying shapes, sizes, colours, and textures. The object varies from Having the slick, smooth surfaces to the ones having textured, matte. This will help Us check how well our model will perform on different kinds of objects.

c) Environmental Complexity: The lab environment is controlled but dynamic, with other robots in operation and manuals and tools available. This will include a realistic but controlled level of background complexity to ensure that the model is tested under conditions closest to reality.

d) Infrastructure and Support: The lab features the most advantageous facilities, highest-speed computers for model processing, and more than one robotic system. The above setup is coupled with the facility of guidance from our lab instructors whose guidance promises for precise experimentation and debugging.

2) Hardware Specifications: Our system ran on the latest hardware to function smoothly and for outputting accurately:

a) Stretch 3 Robot: Equipped with the Intel RealSense D435 camera, Stretch 3 came in as the main robotic platform for the imaging work. The fact that it could be taken around and captures the views formed an array of datasets.

b) GPU Setup: Segmentation tasks were run on a computational resource with high performance, equipped with a modern GPU NVIDIA RTX 3090. This permitted SAM 2 to fine-tune with gigantic datasets, though it never would have allowed any workflow to be anything but smooth and fluent. With the optimized environment, numerous scenarios and latest hardware, it became a basis to conduct further tests for the capabilities of SAM 2 from where information valuable enough was yielded. its flexibility and performance with real life conditions.

IV. Comparative Analysis: SAM 2 vs. YOLO

A. Advantages of SAM 2 Over YOLO

1) Accuracy: SAM 2 provides better segmentation accuracy, especially with the shapes such as irregularly or overlap objects. Where YOLO uses bounding-box based method of segmentation, SAM 2 is pixel-based segmentation that provides smooth and accurate object boundaries and recognition. To do that in STRETCH3, to be precise it has to pick a fragile object from the ground or to reach between the tight-packed objects.

2) Flexibility: This is accomplished by having SAM 2 to be class-agnostic at the zero-shot capability which allows the ability to generalize to unknown objects from limited training. Its flexibility assures that STRETCH3 could work well in dynamic environments, where novel categories of objects or changes in scenes are introduced.

3) User Interaction: SAM 2 supports interactive inputs, meaning segmentation user or robot results can be improved in real time. Inputs are points, bounding boxes, or text Prompts enable SAM 2 to be responsive to such specific needs or Correcting errors in a real-time process in order to improve flexibility in operations.

B. Performance Metrics

1) Accuracy: SAM2(0.7018) outperformed YOLO(0.6529) on the metrics for Intersection over Union (IoU), evidencing that it is better at mat

2) Computational Speed: YOLO is optimized for speed, typically processing frames many times faster than SAM 2, thereby suitable for simpler tasks that guarantee real-time detections. SAM 2, though computational, offers the possibility of optimization options like GPU acceleration to balance speed

C. Ability to Adapt to New and Unseen Objects

- SAM 2 proves to have greater flexibility in its zero-shot capabilities, as it segments objects outside the training dataset very effectively.
- YOLO is affected by unseen objects and will need to be retrained for every new class; this increases the development time and even the development effort.

Some of the output images comparing segmentation between SAM 2 and YOLO: While YOLO is still a great benchmark model for simple real-time object detection, SAM 2 outperforms YOLO in accuracy and adaptability as well as superior segmentation capabilities, providing a much more solid solution than the former for STRETCH3's challenging advanced image segmentation tasks. SAM 2's dynamic capability in environments, granular segmentation of segments, and adaptation to new scenarios helps improve the overall performance and flexibility of STRETCH3 to a great extent.

V. Results and Discussion

A. Observations on SAM 2's Performance in Real-World Scenarios

SAM 2 gave excellent performance in tasks concerning object segmentation that operated under controlled real-world

Fig. 6: SAM 2

Fig. 7: YOLO V11

environments, including labs and robotics scenarios. The model very efficiently segmented challenging objects at a high accuracy even in cases where

- *Cluttered laboratory environments*,with highly overlapped tools and apparatus created massive difficulties in segmentation.
- *Medical datasets*,even the most minute brain tumor boundaries were picked up successfully while outperforming models like YOLO v11 through IoU values in these experiments.

SAM 2's zero-shot segmentation skill gave it out to generalize well for objects and situations not encountered by the model at training time, further affirming its strength.

B. Insights on SAM 2's Ability to Handle Complex and Unseen Environments

1) Adaptability: SAM 2 demonstrated outstanding versatility for unseen environments, such as other lab settings and medical imaging datasets with high complexity. Its processing and ability to segment objects in fine-grained levels make the tool trustworthy for several domains, including robotics and health care.

2) Depth-Aware Segmentation: It efficiently utilized spatial information from the depth data from the Intel RealSense

Fig. 8: SAM 2

Fig. 10: SAM 2

Fig. 9: YOLO V11

Fig. 11: YOLO V11

D435 camera for enhanced accuracy in the segmentation for environments with geometrical complexities.

C. Challenges Observed

Apart from its positive sides, the The method faced some issues at its performance. With the integration and deployment of SAM 2 with the Stretch3 robot:

1) Compatibility Issues:

- The pyrealsense2 library, used for interaction with the Intel RealSense D435 camera, supports only Python versions less than 3.9.
- On the other hand, SAM 2 demands dependencies compatible with Python greater than 3.10.
- To overcome this, two independent environments were created, which added to the complexity in handling the workflow and data transfer between the environments.

2) Processing Speed: The system's 4 GB GPU was constraining: limitations on processing speed, leading to an average inference The processing time was 2–3 seconds per image. Although this was sufficiently good for most research objectives, it might be a problem in live applications where fast processing speeds are crucial.

D. Computational Demands

SAM 2 was memory intensive and thus needed careful optimization by utilization of GPU memory and processing pipelines; hardware upgrades or alternative strategies would be required to reduce computing overhead for real-time deployment in more resource-constrained settings.

The study indicates both the potential and limitations of the SAM 2 model in its application in the real world. In spite of showing promising performances, dealing with computational and compatibility challenges is a critical issue for broad applications of the model in robotics and other practical domains.

VI. Future Works

A. Proposed Workflow

Integration of YOLO and SAM 2 on Stretch 3 for Enhanced Object Detection and Segmentation This serial integration will integrate the detections from the bounding boxes of the objects by YOLO with the segmentation ability of SAM 2 for the performance of accurate localization of objects as well as refinement of boundaries.

1) Object Detection Using YOLO: The Stretch 3 camera has captured an image of the scene The image is passed through Object Detection Bounding Box Prediction YOLO detects objects in the scene and returns:

- Bounding Boxes: Coordinates of rectangular regions containing objects.
- Class Labels: Predicted category of each object.
- Confidence Scores: Probability of the detection being correct.

a) Bounding Box Adjustment: Optional resize YOLO's bounding boxes by a small amount (10–15 percent) so that the whole object is included for segmentation.

2) Crop ROIs for Segmentation:

a) Extract Regions of Interest (ROIs): For each of the bounding boxes detected by YOLO, extract the corresponding region from the original image.

b) Prepare the ROIs: Transform the extracted regions to the desired format for SAM 2, such as resizing or normalizing if necessary.

3) Fine Segmentation with SAM 2:

a) Input to SAM 2: Each cropped ROI is fed into SAM 2 to be segmented. The coordinates of the bounding box returned by YOLO can feed in prompts to SAM 2 to highlight of interest.

b) Segmentation Mask Generation: SAM 2 generates a detailed segmentation mask for every ROI. It captures the precise shape and boundaries that define the object in the bounding box.

- Mask for Object 1: A pixel-level binary mask marking the "Bottle" area.
- Mask for Object 2:A binary mask capturing the exact shape of the "Box".

4) Fine-Tuning YOLO's Bounding Box Mask-Based Refinement:

a) Refine YOLO's bounding box using SAM 2's segmentation masks: Conform the box tightly about the segmented object. Calculate box dimensions based on the size of the mask.

b) Validation: In case SAM 2 outputs an invalid mask for a YOLO detection label as probable false positive. Introduce a new detection when SAM 2 is segmenting an object outside the original bounding box of YOLO

5) Post-Processing and Integration:

a) Aggregating the Results: At object level, aggregate YOLO class label and confidence score with SAM 2's segmentation mask. Output a detection list sorted by:

- Refined Bounding Boxes.
- Segmentation Masks.
- YOLO's Class Labels and Confidence Scores.

b) Visualization: Overlay the refine bounding boxes and segmentation masks on the original image. Class label with confidence score annotate objects Stretch 3 Integration Use the ROI for executing functions such as object manipulation, such as:

- Identify grasp points for the arms of the robot.
- Avoid regions occluded by the object in a pick-and-place application.

6) Real-Time Optimization:

a) Batch Processing: Run several ROIs together in parallel, processing much faster, especially on real-time applications. Optimize YOLO and SAM 2 together on the GPU

b) Dynamic Tuning: Dynamically adjust the threshold of YOLO appropriate to detection quality versus overheads on computational. Call SAM 2 only for high-confidence YOLO detections or challenging cases.

c) Benefits of This Process:

- Higher Detection Accuracy: The segmentation of SAM 2 ensures tight-fit bounding boxes around the objects with improved accuracy.
- Better Object Manipulation: Stretch 3 will have knowledge about object boundaries using segmentation masks to make better plans for grasps.
- Reduction of False Positives: Using SAM 2 to cross-validate the detections of YOLO helps reduce false positives.
- Better Extension to Complex Scenes:SAM2 handles complex scenes fairly well, especially with matters of overlap and formed objects where YOLO does very badly.

VII. Conclusion and Recommendations

A. Conclusion

The integration of SAM 2 into the STRETCH3 system represents a significant leap forward in enhancing the robot's image segmentation capabilities. The comparative analysis has demonstrated that:

- SAM 2 excels in segmentation precision, particularly for irregularly shaped, overlapping, or complex objects, enabling STRETCH3 to handle intricate manipulation tasks with higher accuracy.
- Its class-agnostic and zero-shot capabilities provide unmatched adaptability, allowing the robot to operate effectively in dynamic environments with objects that are not seen before.
- The advanced video segmentation capabilities of SAM 2 ensure temporal consistency between frames, enhancing the performance of robots in real-time interaction tasks.
- Although more computationally intensive, the flexibility, accuracy, and robustness in SAM 2 largely outweigh what can be attained in YOLO. Therefore, it is recommended that STRETCH3 use SAM 2 for its future development.

Implementing SAM 2 in STRETCH3 will better face the challenges posed by modern robotics in tasks such as operating in cluttered environments and performing precision-driven operations.

B. Recommendations for Future Improvements

1) Optimize SAM 2 for Faster Inference:

a) Apply model optimization techniques: quantization or pruning-to reduce SAM 2's computational needs without sacrificing accuracy.

- Leverage on the board, onboard hardware accelerators, namely NVIDIA Jetson modules, to support the deployment of SAM 2 on STRETCH3 with real-time performance.

- Investigate lightweight variants of SAM 2 or design a task-specific model that learns from the data of operation generated by STRETCH3.

2) Hybrid Models: Investigate how the speed of YOLO can be married with the accuracy of SAM 2 in a hybrid pipeline: for example,

- Use YOLO for fast object detection and localization.
- Pass regions-of-interest to SAM 2 for detailed pixel-level segmentation.

Such an approach can thus offer the best balance of real-time processing and segmentation accuracy, making STRETCH3's operational efficiency maximized.

3) Real-Time Adaptation: Interactive features of SAM 2 must be designed in such a way that they allow the STRETCH3 to adapt segmentation dynamically during operation based upon the feedback or evolving requirements of the task.

Introduce advanced mechanisms of learning such as online learning or self-supervised learning that allow continuous improvement of the model as it experiences new objects and scenarios with STRETCH3.

4) System Scalability and Evaluation: Conduct rigorous field tests in multiple contexts to validate the performance of SAM 2 and to refine the system further.

Scaling up with provisions to incorporate new functionality like the simultaneous segmentation of multiple objects or interfacing with other sensors

By optimizing SAM 2 and developing novel hybrid solutions, the objective of STRETCH3 is to be the best in real-time performance while maintaining satisfactory precision and adaptation as required in advanced robotics applications.

VIII. Limitations

A. Object Detection using YOLOv3

This is YOLOv3, which is a bit older and less accurate compared to state-of-the-art object detection models like YOLOv5, YOLOv7 or YOLOv11. At times it may gripe incorrectly due to error in object detection; otherwise, the task will be inefficient.

B. Outdated Software Stack

Ubuntu 20.04 is being used by the robot which is aging and does not support new libraries and hardware opti mizations well. The pyrealsense2 library is also outdated, which severely limits compatibility with the latest Intel RealSense SDK features. Older dependency causes problems in the adaptation of newer technologies.

C. Higher Training and Processing Times

With SAM 2, very detailed types of segmentation for most objects are expressed, but at an incredibly high cost in terms of required computational resources. Training the model on new datasets is laborious, making it less ideal for scenarios demanding rapid adaptation. Processing a single image or video frame takes much longer than lightweight segmentation models, like YOLOv11.

D. Resource Constraints

SAM 2's operations demand high-end GPUs, which may not be feasible for edge devices or low-power systems. In scenarios like real-time segmentation for robotics, this computational overhead can hinder real-time performance.

E. Dependency on Extensive Pretraining

SAM 2 generalizes by leveraging huge pretraining datasets. Such reliance potentially constrains its performance in the case of in-domain or novel tasks unless further fine-tuned.

F. General Issues

1) Integration Issues: Integrating SAM 2 with the Stretch 3 robot adds interaction issues related to hardware-software compatibility and real-time performance. There might be technical challenges with updating software stacks - (e.g. Ubuntu 20.04) for older machines to state-of-the-art deep learning models such as SAM 2 or YOLOv11.

2) Power and Cost: High computational and power requirements of SAM 2 and modern YOLO models increase the operational cost for tasks requiring continuous use. Stretch 3's hardware, optimized for its current software ecosystem may need upgrades to take full advantage of newer models. These shortcomings motivate an exploration of a more balanced solution that shall combine the precise detection of SAM 2 with speed and efficiency of models such as YOLOv11, along with a software and hardware stack upgrade for the Stretch 3 robot.

IX. Ethical Considerations

Incorporating SAM 2 into STRETCH3, provides advanced functionalities while it raises a number of ethical considerations, which need to be treated in order that robots can be used responsibly with their technology:

A. Transparency and Accountability

1) User Understanding: The end-users are not aware of SAM 2 capabilities, and also their limitations: they tend to over reliance or even misuses.

2) Mitigation: Transparency in documentation along with clear guidelines and training of users would help the system being used responsibly.

B. Environmental Impact

1) Energy Consumption: Computational requirement high for SAM 2; thus, it consumes too much energy that might be unsustainable.

2) Mitigation: Minimize energy consumption from SAM 2 by exploring hardware alternatives that reduce energy consumption.

C. Accessibility and Fairness

1) Unequal Access: Advanced robotics solutions such as STRETCH3 with SAM 2 will be something of a luxury good for rich institutions or organizations and promise unequal technological leapfrogs.

2) Mitigation: Develop implementations at a cost-effective price and explore opportunities for open-source contributions to widely distribute the benefits from technology.

D. Safety and Reliability

1) Operational Failures: Segmentation errors may lead to false object identification or manipulation, which could result in property damage or user harm.

2) Mitigation: Ensure robust testing, fallback mechanisms, and safety protocols to mitigate the effects of segmentation errors.

E. Dual-Use Concerns

1) Misuse Potential: Advanced image segmentation capabilities can be abused for surveillance or applications that violate privacy or human rights.

2) Mitigation: Limit the use of SAM 2 with STRETCH3 to ethical applications via use policies and to comply with other regulations.

F. Conclusion

It will address such ethical issues in integrating SAM 2 into STRETCH3, not only enhancing its capabilities but also bringing forward greater sense of responsibility, fairness, and transparency. By addressing these potential risks through proactive approaches and maintaining that ethical integrity, the project can become a benchmark for responsible robotics innovation.

X. Acknowledgments

We would like to express our deepest gratitude to Prof. Vrijendra Singh and Mr. Ashok Kumar for their invaluable guidance and support throughout our project involving the Stretch 3 robot, SAM 2 model, and YOLOv11 model. Their expertise, insights, and encouragement have been instrumental in the successful execution of this work.

We are sincerely grateful for their mentorship and for providing us with the opportunity to learn and contribute to this challenging yet rewarding research. Their unwavering commitment to academic excellence has been a source of inspiration, and we are truly honored to have had the chance to work under their supervision.

We also extend our thanks to all those who provided assistance, constructive feedback, and resources that enabled us to accomplish this project.

XI. References

1) Ravi, N., Gabeur, V., Hu, Y. T., Hu, R., Ryali, C., Ma, T., Feichtenhofer, C. (2024). Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714.

2) Zhu, J., Qi, Y., Wu, J. (2024). Medical SAM 2: Segment medical images as video via Segment Anything Model 2. arXiv preprint arXiv:2408.00874.

3) Yan, Z., Sun, W., Zhou, R., Yuan, Z., Zhang, K., Li, Y., ... Sun, L. (2024). Biomedical SAM 2: Segment anything in biomedical images and videos. arXiv preprint arXiv:2408.03286.

4) Gilbert, S. D., Rambo, R. P., Van Tyne, D., Batey, R. T. (2008). Structure of the SAM-II riboswitch bound to S-adenosylmethionine. *Nature Structural Molecular Biology, 15*(2), 177-182.

5) Kemp, C. C., Edsinger, A., Clever, H. M., Matulevich, B. (2022, May). The design of Stretch: A compact, lightweight mobile manipulator for indoor human environments. In *2022 International Conference on Robotics and Automation (ICRA)* (pp. 3150-3157). IEEE.

6) Ali, M. F., Nchekwube, D. C., Genova, O., Freddi, A., Monteriù, A. (2023, October). An assistive robot in an indoor scenario: The Stretch Hello Robot as environment organizer. In *2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)* (pp. 676-681). IEEE.

7) Kadylak, T., Bayles, M. A., Galoso, L., Chan, M., Mahajan, H., Kemp, C. C., ... Rogers, W. A. (2021, September). A human factors analysis of the Stretch mobile manipulator robot. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 65, No. 1, pp. 442-446). Sage CA: Los Angeles, CA: SAGE Publications.

8) Kadylak, T., Bayles, M. A., Galoso, L., Chan, M., Mahajan, H., Kemp, C. C., ... Rogers, W. A. (2021, September). A human factors analysis of the Stretch mobile manipulator robot. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 65, No. 1, pp. 442-446). Sage CA: Los Angeles, CA: SAGE Publications.

9) Tsykunov, E., Ilin, V., Perminov, S., Fedoseev, A., Zainulina, E. (2020). Coupling of localization and depth data for mapping using Intel RealSense T265 and D435i cameras. *arXiv preprint arXiv:2004.00269*.

10) Shahmoradi, R. (2022). Investigating the feasibility of using a RealSense depth camera D435i by creating a framework for 3D pose analysis (Master's thesis, University of Twente).

11) Thongprasith, J., Separattananan, P., Meyer, P., Chanchareon, R. (2023, June). Portioning algorithm using the bisection method for slicing food. In *2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)* (pp. 1-6). IEEE.

12) Charngtong, C., Dheeravongkit, A., Vonzbunvona, S. (2024, June). A 3D camera protocol for object pose estimation from point cloud in robot operations. In *2024 21st International Joint Conference on Computer Science and Software Engineering (JCSSE)* (pp. 9-15). IEEE.

13) Tsykunov, E., Ilin, V., Perminov, S., Fedoseev, A.,

Zainulina, E. (2020). Coupling of localization and depth data for mapping using Intel RealSense T265 and D435i cameras. *arXiv preprint arXiv:2004.00269*.
14) Zhao, L., Li, S. (2020). Object detection algorithm based on improved YOLOv3. *Electronics, 9*(3), 537.
15) Mao, Q. C., Sun, H. M., Liu, Y. B., Jia, R. S. (2019). Mini-YOLOv3: Real-time object detector for embedded applications. *IEEE Access, 7*, 133529-133538.
16) Khanam, R., Hussain, M. (2024). YOLOv11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*.
17) Alif, M. A. R. (2024). YOLOv11 for vehicle detection: Advancements, performance, and applications in intelligent transportation systems. *arXiv preprint arXiv:2410.22898*.
18) Sharma, A., Kumar, V., Longchamps, L. (2024). Comparative performance of YOLOv8, YOLOv9, YOLOv10, YOLOv11, and Faster R-CNN models for detection of multiple weed species. *Smart Agricultural Technology, 100648*.
19) https://hello-robot.com/stretch-3-product
20) https://hello-robot.com/
21) https://github.com/hello-robot
22) https://ai.meta.com/sam2/
23) https://segment-anything.com/
24) https://github.com/facebookresearch/segment-anything
25) https://arxiv.org/abs/2408.00714
26) https://arxiv.org/abs/2408.00874
27) YOLOv11 Documentation
28) Ultralytics GitHub Repository
29) YOLOv3 Overview by Viso.ai
30) Darknet YOLO by PJ Reddie
31) YOLO Object Detection Explained - DataCamp
32) YOLO Research Paper on arXiv
33) Brain Tumor Dataset on Roboflow
34) Download LabPicsV1 Dataset (Zenodo)
35) Intel RealSense D435 Camera Information
36) pyrealsense2 Python Package
37) Introduction to pyenv - Real Python
38) Managing Multiple Python Versions Without Tools - Python Discussions
39) OpenCV Official Website
40) OpenCV GitHub Repository
41) GeeksforGeeks
42) PyTorch Official Website
43) ResearchGate