VideoSAM: A Large Vision Foundation Model for High-Speed Video Segmentation

Chika Maduabuchi
Massachusetts Institute of Technology
USA
chika691@mit.edu

Ericmoore Jossou
Massachusetts Institute of Technology
USA
ejossou@mit.edu

Matteo Bucci
Massachusetts Institute of Technology
USA
mbucci@mit.edu

***Abstract*—High-speed video (HSV) segmentation is essential for analyzing dynamic physical processes in scientific and industrial applications, such as boiling heat transfer. Existing models like U-Net struggle with generalization and accurately segmenting complex bubble formations. We present VideoSAM, a specialized adaptation of the Segment Anything Model (SAM), fine-tuned on a diverse HSV dataset for phase detection. Through diverse experiments, VideoSAM demonstrates superior performance across four fluid environments—Water, FC-72, Nitrogen, and Argon—significantly outperforming U-Net in complex segmentation tasks. In addition to introducing VideoSAM, we contribute an open-source HSV segmentation dataset designed for phase detection, enabling future research in this domain. Our findings underscore VideoSAM's potential to set new standards in robust and accurate HSV segmentation. The code and dataset used in this study are available at: https://github.com/chikap421/videosam.**

***Index Terms*—Computer vision, Video segmentation, Segment Anything Model, Phase detection**

I. Introduction

High-speed video (HSV) segmentation is essential for analyzing complex physical processes that are crucial in various scientific and industrial applications, such as chemical processes [1]–[4], bubble recognition in heat transfer [5]–[8], and high-speed imaging of dynamic events [9]–[13]. Manual segmentation of objects in HSVs, such as bubbles, is time-consuming, labor-intensive, and often subjective, requiring significant expertise [14]–[17]. Automated segmentation methods can significantly reduce the time and labor required, increase consistency, and enable the analysis of large-scale HSV datasets [18]–[22].

Convolutional neural networks (CNNs), particularly U-Net [18], have become the standard for HSV segmentation tasks [5]–[7], [17] due to their ability to learn complex features and deliver accurate segmentation results. However, these models are often highly task-specific, limiting their generalization to new tasks or varying imaging conditions [23]–[27]. This limitation is particularly challenging in HSV analysis, where dynamic nature, temporal dependencies, and real-time processing demands are prevalent [24], [28]–[31].

Recent advancements in segmentation models, particularly the emergence of foundation models like the Segment Anything Model (SAM) [32], have shown remarkable versatility and generalization across various tasks. Despite their success, the applicability of these models to scientific HSV tasks, especially for tasks like bubble segmentation, remains largely unexplored due to the significant differences between natural images and HSV frames [33]–[37].

Moreover, the diverse distribution of training data from different scientific HSV experiments often leads to domain shifts, which can result in suboptimal performance when using traditional CNNs for segmentation [5], [38]–[42]. To address these issues, specialized models are typically built for each data distribution, which may improve performance on similar test data but often fails to generalize well to new datasets [39], [40], [43]–[45]. Despite the dominance of models like U-Net in HSV segmentation, there is a significant gap in leveraging large vision foundation models like SAM to achieve better generalization in these tasks.

To overcome these limitations, we introduce **VideoSAM**, a refined vision foundation model that enhances SAM's segmentation performance specifically for scientific HSV tasks. VideoSAM is fine-tuned on a newly curated and extensive dataset of HSV frame-mask pairs, designed to cover a broad range of boiling modalities and dynamic behaviors. This dataset, which we introduce as a key contribution, facilitates the fine-tuning of large vision models for more robust and generalizable segmentation in HSV analysis.

Our experiments across four different fluids—Water, 3M™ Fluorinert™ Liquid (FC-72), Nitrogen, and Argon—demonstrate that VideoSAM significantly outperforms traditional models like U-Net, particularly in complex fluid environments. These results highlight the potential of VideoSAM to serve as a versatile and robust solution for HSV segmentation, offering improved generalization and accuracy in diverse scientific applications.

The remainder of this paper is organized as follows: Section 2 provides an overview of the related work in foundation models and HSV segmentation. Section 3 describes the methodology, including the dataset preparation, VideoSAM model, and training process. Section 4 presents the experimental results and discusses the performance of VideoSAM in comparison to existing models. Finally, Section 5 concludes the paper and outlines future research directions.

Identify applicable funding agency here. If none, delete this.

II. Related Works

Deep learning-based methods have shown significant promise in scientific HSV segmentation tasks, leveraging the hierarchical feature learning capabilities of deep neural networks. This section is divided into three parts: traditional HSV segmentation methods, large vision foundation models in segmentation, and the application of these models to scientific HSV tasks.

A. Traditional HSV Segmentation Methods

CNNs have become the dominant models for bubble segmentation in HSVs. U-Net [18] and Mask R-CNN [23] are particularly popular due to their robust performance in various image segmentation tasks. For instance, Passoni et al. [7] and Dunlap et al. [46] effectively employed U-Net and Mask R-CNN-based models for bubble segmentation, demonstrating their capacity to capture the complex dynamics of bubbles in HSVs. Similarly, Malakhov et al. [17] utilized a modified U-Net/Mask R-CNN architecture to segment bubbles in boiling experiments. Other researchers, such as Chernyavskiy et al. [47] and Hessenkemper et al. [48], explored various CNN variants for similar tasks, further validating the efficacy of CNNs in HSV segmentation.

Despite the advancements brought by these CNN architectures, they are inherently task-specific and often struggle to generalize to new HSV conditions, fluid properties, or imaging setups [23]–[27]. The diverse distribution of training data from various scientific HSV experiments can lead to domain shifts, resulting in suboptimal performance when these models are applied to unseen data distributions [5], [38]–[42]. Although specialized models tailored to specific data distributions can improve performance on similar test datasets, they often fail to generalize well beyond the trained distribution [39], [40], [43]–[45].

B. Large Vision Foundation Models in Segmentation

Recent advancements in image segmentation have introduced foundation models, such as the SAM [32], which are pretrained on large-scale, diverse datasets and have demonstrated strong performance across various segmentation tasks. These models are designed to be versatile and generalizable, capable of adapting to different segmentation challenges with minimal fine-tuning. SAM, for instance, leverages a novel prompt-based approach to achieve state-of-the-art results on natural image segmentation benchmarks, highlighting the potential of large vision foundation models to surpass traditional CNN-based approaches in flexibility and performance.

Other notable foundation models include SEEM [49], Mask2Former [50], HRNet [51], and Swin Transformer [52]. These models have set new benchmarks in natural image segmentation by incorporating architectural innovations such as multi-modal prompts, masked attention mechanisms, high-resolution representations, and hierarchical feature extraction. Their success in natural image segmentation has spurred applications in various domains, including medical imaging [53]–[57], remote sensing [58]–[62], and video tracking [63]–[67].

C. Existing Datasets in Boiling Phenomena

Existing datasets, such as the Boiling Dataset [68], focus primarily on broader boiling phenomena for classification tasks. However, these datasets do not cover phase detection data segmentation using HSV, lacking the detailed, high-resolution frame-mask pairs necessary for training and fine-tuning advanced large vision models in HSV research. We introduce a novel dataset specifically designed for phase detection in HSV segmentation, addressing these limitations and pushing forward the application of large vision models in boiling data analysis.

D. Application of Large Vision Models to Scientific HSV Segmentation

The application of large vision foundation models like SAM to scientific HSV segmentation remains largely unexplored, despite their potential to address the limitations of traditional CNN models in this domain. The unique challenges posed by HSVs, such as overlapping bubbles, weak boundaries, and dynamic bubble behavior, require more sophisticated segmentation approaches than what traditional CNN models can offer. Large vision models, with their ability to learn from diverse data and generalize across tasks, present an opportunity to significantly advance HSV segmentation.

However, the gap in applying these models to HSV segmentation is notable. Traditional CNN models dominate this field, yet they often fail to generalize well to new HSV conditions or fluid properties. The lack of open-source datasets specific to phase detection in HSVs has further hindered the development and fine-tuning of large vision models in this area. To bridge this gap, our work introduces **VideoSAM**, a refined segmentation foundation model that enhances SAM for HSV tasks. VideoSAM is fine-tuned on a specialized dataset of HSV frame-mask pairs, incorporating domain-specific adaptations that enable the model to capture the unique characteristics of bubbles effectively.

In addition to introducing VideoSAM, we contribute a novel HSV segmentation dataset specifically designed for phase detection. This dataset is made publicly available to facilitate further research and development of large vision foundation models for HSV segmentation. Through comprehensive experiments, we demonstrate that VideoSAM outperforms both SAM and specialist models like U-Net across various HSV datasets, making it a promising candidate for advancing segmentation tasks in scientific and industrial settings.

III. Methodology

A. Model Architecture

The model architecture, illustrated in Figure 1, follows a two-stage approach. Initially, specialized U-Net CNNs were built for each data modality (Argon, Nitrogen, FC-72, and Water) since the model could not generalize across all modalities. These U-Net models, originally trained on cellular images [18], were fine-tuned to each modality to generate initial segmentation masks. The image-mask pairs were then fed

into the VideoSAM transformer architecture to test its zero-shot generalization ability across different fluids. The model processes these inputs via its image encoder and mask decoder, producing refined segmentation masks for the final output. This approach leverages U-Net's initial mask generation and VideoSAM's capacity for handling complex high-speed video segmentation tasks.

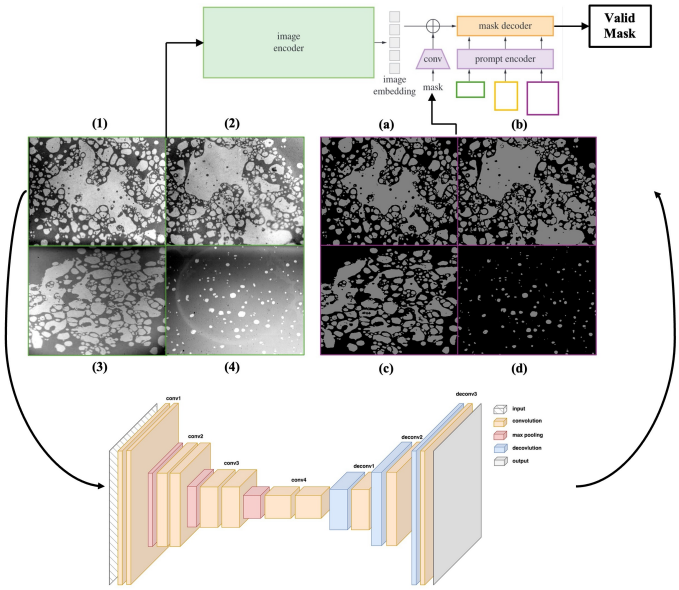


Figure 1: Illustration of the VideoSAM model architecture and integration with U-Net CNN. The initial segmentation masks generated by fine-tuned U-Net models for each modality are paired with their respective images and fed into the VideoSAM transformer. The image encoder and mask decoder process these inputs to refine the segmentation, leveraging the pre-trained SAM components for HSV segmentation.

Table I: High-Speed Video Modalities for Experimental Data Collection

Modality	**Boiling Conditions**	**Pressure (bar)**	**Heat Flux (kW/m^2)**	**Frames**
Argon	SPB	1	120	6K
Nitrogen	SPB	1	120	6K
FC-72	SPB	1	170	6K
Water	FB at 500 kg/m^{2}s	10	3000	7.5K

SPB: Saturated Pool Boiling, FB: Flow Boiling

B. Data Collection

To develop a versatile and generalizable foundation model for bubble segmentation in HSVs, we curated an extensive and diverse dataset of video frame-mask pairs from high-speed camera experiments. The dataset encompasses several boiling modalities with varied conditions, fluid properties, and imaging setups, ensuring robustness and applicability of the model across various scenarios. The HSV modalities are detailed in Table I, which includes data collected under different boiling conditions for liquid argon, liquid nitrogen, FC-72, and high-pressure water. Each modality was recorded with a specific resolution and frame count to capture the dynamic boiling processes effectively.

C. Data Processing

To ensure high-quality training data, 250 random frames were sampled from each data modality, yielding 1000 frames in total. Although this study focuses on frame-based segmentation, future work will incorporate temporal dynamics to enhance performance in HSV tasks. The training, testing, and validation sets were created using an 80:20 split of the remaining frames, ensuring diverse and representative samples across all modalities.

Raw images were converted to grayscale and normalized to enhance feature visibility by subtracting blank reference frames and adjusting contrast, which reduced background noise and improved mask extraction. Ground truth segmentation masks were created using a combination of manual annotation and semi-automatic techniques, with ImageJ [69] and adaptive thresholding algorithms [15] used by domain experts. A U-Net model was trained on these annotated frames to generate initial segmentation masks, which were then refined by human annotators to scale the dataset efficiently while maintaining accuracy.

The images and masks were then patchified using a 100x100 pixel grid, discarding patches without mask information. Remaining patches were resized to 256x256 pixels for model training, and masks were normalized to binary values (0 or 1). Random patches were visualized to confirm preprocessing integrity, as shown in Figure 2.

This comprehensive data processing pipeline ensured the quality, diversity, and representativeness of the training data, improving the robustness and generalizability of the VideoSAM model.

D. Training Process

VideoSAM was fine-tuned by freezing the pre-trained vision and prompt encoder layers of the `facebook/sam-vit-base` model while allowing updates to the mask decoder. A custom `SAMDataset` class managed the data, incorporating bounding box generation for masks. The dataset was wrapped in `DataLoader` objects to enable efficient batch processing.

The model was trained using the Adam optimizer (1×10^{-5}, no weight decay) with a combination of Dice Coefficient and Cross-Entropy Loss. Mixed precision training with `GradScaler` was applied to accelerate training and reduce memory usage, with gradient clipping used to maintain stability. At each epoch, the model's performance was evaluated on a validation set using metrics like IoU, precision, and recall. Learning rate scheduling (`ReduceLROnPlateau`) was applied based on validation loss to optimize training.

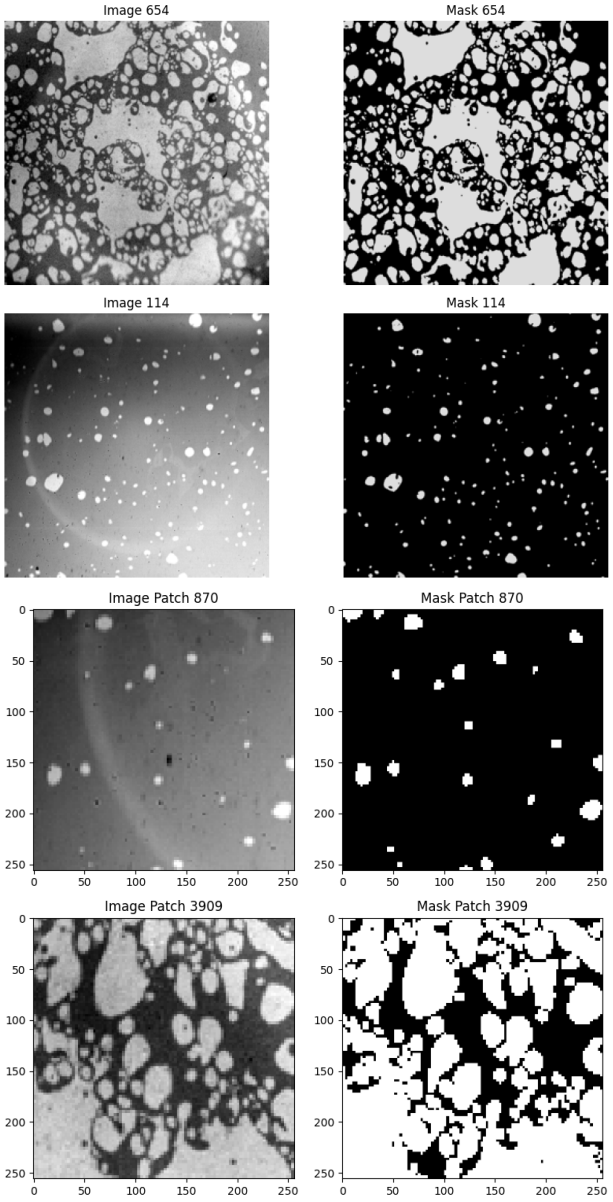


Figure 2: Left: Original high-speed video frames showcasing randomly sampled frames from the large training dataset. The images illustrate the difference between the modalities of water (image 114) and gas (image 654). Notice the difference in bubble footprints, with gases exhibiting more bubbles with complex shapes compared to water. Right: Visualization of the patched images resulting from the patchification process. This process highlights the segmentation of the original images into smaller patches for detailed analysis.

Training and validation losses, as well as key metrics, were logged, and model checkpoints were saved periodically. This meticulous process ensured that VideoSAM was robust and capable of performing high-accuracy bubble segmentation across diverse data modalities.

E. Inference Pipeline

The inference pipeline for VideoSAM evaluated performance on different data modalities using mask extraction and metrics evaluation for both single and composite frames.

Mask extraction began by converting high-speed video frames to grayscale and normalizing them. For single frames, the `SAMInferenceDataset` class segmented images into smaller patches using grid-based bounding boxes. These patches were processed through both VideoSAM and the base SAM model, and the predicted masks were stitched together to reconstruct full image masks. For composite frames, the same approach was extended to sequences, ensuring dynamic video data was consistently evaluated over time.

Metrics evaluation compared predicted masks with ground truth using F1 Score, IoU, and Precision for single frames. For composite frames, these metrics were aggregated across sequences to provide detailed insights into performance consistency, including mean, minimum, maximum, and standard deviation.

This pipeline combined mask extraction and detailed metrics to thoroughly assess VideoSAM's segmentation capabilities across diverse modalities and ensure robust evaluation.

F. Experimental Setup

In this study, we conducted three key experiments to evaluate the performance of VideoSAM. These experiments were designed to test the model's zero-shot/generalization capabilities, its performance across different data modalities, and to compare its results against a diversified CNN model.

1) Experiment 1: Zero-Shot Generalization Across Modalities: The first experiment aimed to test the ability of VideoSAM to generalize to unseen data modalities. For this, we trained VideoSAM on the high-speed video frames from the Argon boiling modality. After training, we tested the model on all other data modalities, including Nitrogen, FC-72, and Water. The goal was to evaluate how well the model could perform zero-shot segmentation on data modalities it had not encountered during training. We inspected the results through both visual analysis and by quantifying key metrics, including Intersection over Union (IoU) and F1 Score, which are standard metrics for segmentation tasks. We expected VideoSAM to outperform the baseline Segment Anything Model (SAM) across all modalities. The experiment demonstrated that VideoSAM achieved superior segmentation quality compared to SAM, as confirmed by both visual inspection and the quantified metrics.

2) Experiment 2: Performance Across Multiple Modalities: In the second experiment, we evaluated VideoSAM's ability to handle multiple boiling modalities. The model was trained on a combination of four different datasets, representing boiling processes in Argon, Nitrogen, FC-72, and Water. After training, the model was tested on unseen data from these same modalities to determine how well it could generalize across diverse conditions. This experiment was crucial for assessing the robustness of VideoSAM in handling a variety of fluids with different boiling characteristics. We expected VideoSAM to demonstrate high performance across all datasets, consistently

outperforming SAM in both IoU and F1 Score. As anticipated, VideoSAM excelled across all test datasets, showing superior performance in capturing the complexities of each boiling modality.

3) Experiment 3: Comparison with U-Net CNN: In this experiment, we compared VideoSAM with U-Net, a well-established CNN architecture frequently employed for high-speed video segmentation tasks [5]–[7]. U-Net's proven success in segmenting complex cellular structures, which share similarities with the bubble footprints in HSV data, makes it a conventional and strong baseline for comparison in these tasks. Both VideoSAM and U-Net were trained on the same four datasets—Argon, Nitrogen, FC-72, and Water—and evaluated using IoU and F1 Score. We expected VideoSAM to excel in complex fluids like FC-72, Nitrogen, and Argon, where dynamic and intricate boiling behaviors dominate, while U-Net was anticipated to perform better on simpler tasks, such as those in the Water dataset. As predicted, VideoSAM delivered superior results in the more challenging fluid environments, while U-Net slightly outperformed in the Water dataset, reinforcing its effectiveness in handling simpler segmentation tasks.

IV. Results and Discussions

A. Experiment 1: Zero-Shot Generalization Across Modalities

The zero-shot performance of VideoSAM in generalizing to unseen data modalities was evaluated on Nitrogen, FC-72, and Water datasets, as shown in Figure 3. The figure combines both the qualitative analysis of segmentation masks (Figure 3a) and the quantitative performance metrics (Figure 3b) to offer a comprehensive assessment of the model's zero-shot generalization capabilities.

In the **qualitative analysis** (Figure 3a), VideoSAM exhibits remarkable segmentation performance in the Nitrogen and FC-72 datasets, closely matching the ground truth masks. The contours and boundaries of the bubbles are significantly better preserved in the binary masks generated by VideoSAM compared to the baseline SAM model, which struggles with boundary delineation and introduces noticeable artifacts. In particular, VideoSAM effectively captures the complex bubble structures in the Nitrogen dataset, maintaining consistent bubble segmentation even in high-density regions. Similarly, for the FC-72 dataset, VideoSAM provides a more accurate segmentation of overlapping and irregularly shaped bubbles, further reinforcing its robustness in generalizing to diverse fluid environments.

However, the model's performance degrades on the Water dataset, where fewer objects of interest and substantial background noise negatively affect its segmentation accuracy. The binary mask for the Water dataset generated by VideoSAM fails to distinguish bubbles clearly, highlighting a significant challenge in simpler datasets with low contrast and fewer distinct objects.

The **quantitative analysis** (Figure 3b) supports these observations. VideoSAM outperforms SAM across the Nitrogen and FC-72 datasets in terms of accuracy, precision, IoU, and Dice coefficient. The model achieves notably higher IoU and F1 scores, confirming its ability to generalize well to unseen data with complex bubble distributions. For example, in the Nitrogen dataset, VideoSAM shows substantial improvement in accuracy and specificity, translating to better boundary detection and fewer false positives.

Conversely, in the Water dataset, VideoSAM's performance falls significantly below that of the other datasets, as reflected in its poor IoU and Dice scores. The simplicity of the Water dataset, characterized by fewer bubbles and more uniform backgrounds, proves to be a challenge for the model, suggesting that its architecture may be overfitted to the complexities of dense and overlapping structures, thus making it less effective in simpler scenarios.

In conclusion, while VideoSAM demonstrates impressive zero-shot generalization in datasets with intricate fluid dynamics like Nitrogen and FC-72, its performance on simpler datasets like Water reveals a limitation that may require additional architectural or preprocessing adjustments to handle different types of modalities effectively.

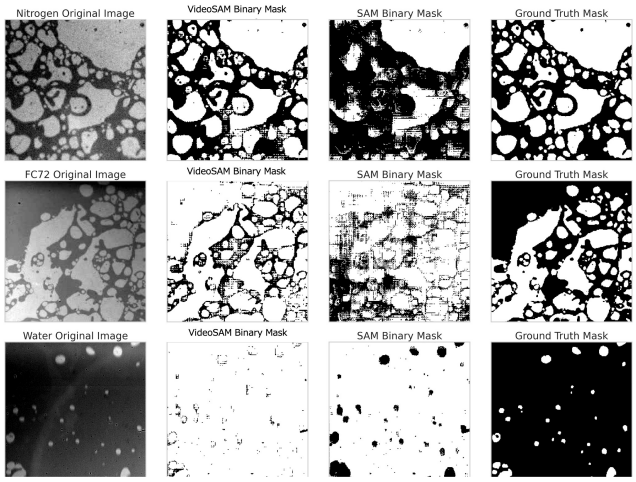
B. Experiment 2: Performance Across Multiple Modalities

In this experiment, the goal was to evaluate the VideoSAM model's generalization performance across various high-speed video modalities, specifically for fluids like Water, FC-72, Nitrogen, and Argon. The results were evaluated based on single-frame and composite-frame segmentation tasks. In both analyses, we compare VideoSAM with the baseline SAM model.

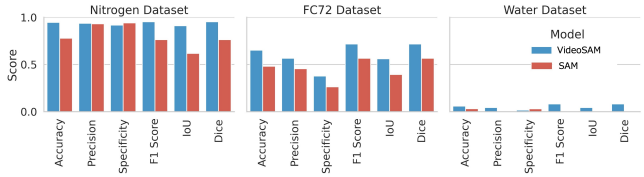
Table 4a presents the results of the single-frame analysis for VideoSAM across different datasets. As indicated, VideoSAM consistently outperforms SAM in the more complex datasets such as Nitrogen, FC-72, and Argon. These datasets exhibit complex bubble dynamics and intricate boundaries, and VideoSAM's architecture proves more effective in capturing these challenging fluid environments. For instance, in the Nitrogen dataset, VideoSAM achieves an IoU of 0.8317 and an F1 Score of 0.9080, compared to SAM's IoU of 0.6702 and F1 Score of 0.8025. Similarly, in the FC-72 dataset, VideoSAM outperforms SAM with an IoU of 0.7997 and F1 Score of 0.8885, showing better segmentation accuracy and boundary delineation.

The Water dataset, however, presents a unique challenge for both models due to fewer objects of interest and significant background noise. VideoSAM achieves an IoU of 0.1894, which, although an improvement over SAM's IoU of 0.0620, highlights the difficulty in segmenting this simpler environment. The lack of distinct features in the Water dataset reduces the effectiveness of both models in distinguishing bubbles from the background.

Additionally, the composite-frame analysis (Figure 4b) further supports these findings. By aggregating the model performance across multiple frames within each dataset, VideoSAM maintains its robustness and superior accuracy in complex environments, particularly in the Nitrogen, FC-72, and Argon datasets. The box plots highlight that VideoSAM's IoU and



(a) Comparison of Binary Masks for Nitrogen, FC-72, and Water datasets between VideoSAM, SAM, and Ground Truth Masks.



(b) Performance comparison between VideoSAM and SAM across various metrics for the Nitrogen, FC-72, and Water datasets.

Figure 3: Combined results of Experiment 1: Qualitative and quantitative analysis of VideoSAM's zero-shot generalization performance.

F1 Score distributions have higher median values across these datasets, showcasing more reliable and consistent performance than SAM. Nevertheless, the results on the Water dataset show a wider variance and lower median values, indicating more significant performance fluctuations due to the simplicity of the scene and background noise.

Experiment 2 demonstrates that VideoSAM excels in handling complex fluid dynamics with intricate bubble boundaries and overlapping structures, especially in datasets like FC-72, Nitrogen, and Argon. However, simpler environments, like the Water dataset, expose the model's limitations, particularly when faced with fewer distinct objects and more background noise. These results suggest that VideoSAM is a powerful tool for high-speed video segmentation in complex environments, but further refinement is needed to improve its performance in simpler scenes. Future directions could involve exploring techniques such as domain adaptation, multi-scale feature extraction, or data augmentation to address these limitations.

C. Experiment 3: Comparison with U-Net CNN

This experiment compares the performance of VideoSAM, U-Net, and SAM models across four fluid datasets: Water, FC-72, Nitrogen, and Argon. The primary objective was to assess how these models perform in both complex and simpler fluid environments. The metrics used for comparison are IoU and F1 Score, both of which measure segmentation accuracy.

Table 4a summarizes the IoU and F1 Scores for U-Net, VideoSAM, and SAM models across all datasets. The results clearly demonstrate VideoSAM's superior performance in more complex environments, such as FC-72, Nitrogen, and Argon. For example, in the Argon dataset, VideoSAM achieved a mean IoU of 0.8384 and an F1 Score of 0.9120, outperforming both U-Net and SAM. Similarly, VideoSAM's performance on the Nitrogen dataset (IoU: 0.8317, F1: 0.9080) also exceeded that of U-Net and SAM.

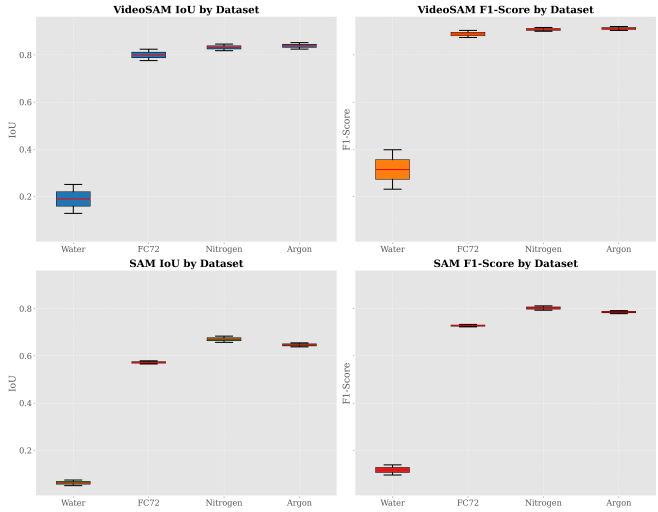
In contrast, U-Net showed the best performance in the simpler Water dataset, achieving an IoU of 0.5619 and an F1 Score of 0.7191, while VideoSAM only managed an IoU of 0.1894 and an F1 Score of 0.3143. This can be attributed to U-Net's architecture, which has been fine-tuned for cellular-level data segmentation, making it more effective in simpler environments where the fluid characteristics closely resemble its training data.

The results from Table 4a indicate that while VideoSAM excels in handling more complex fluids, U-Net remains a strong contender in simpler datasets like Water. This suggests that the architecture of U-Net, particularly its cellular-level segmentation capabilities, makes it highly effective in environments with fewer objects and more consistent fluid structures, as seen in the Water dataset.

On the other hand, VideoSAM's ability to handle complex fluid dynamics and intricate bubble boundaries gives it a

Fluid	Model	IoU	F1 Score
Water	U-Net	**0.5619**	**0.7191**
	SAM	0.0620	0.1165
	VideoSAM	0.1894	0.3143
FC-72	U-Net	0.7244	0.8400
	SAM	0.5721	0.7278
	VideoSAM	**0.7997**	**0.8885**
Nitrogen	U-Net	0.7547	0.8602
	SAM	0.6702	0.8025
	VideoSAM	**0.8317**	**0.9080**
Argon	U-Net	0.7815	0.8773
	SAM	0.6464	0.7852
	VideoSAM	**0.8384**	**0.9120**

(a) Comparison of IoU and F1 Score for U-Net, VideoSAM, and SAM across fluids.



(b) Box plots comparing IoU and F1 Score across datasets.

Figure 4: Combined table and figure layout comparing the performance of U-Net, VideoSAM, and SAM across different datasets.

distinct advantage in more challenging datasets like FC-72, Nitrogen, and Argon. The consistent superiority of VideoSAM in these datasets—demonstrated by its higher IoU and F1 Scores—validates its architecture's robustness in segmenting intricate fluid behaviors and dynamic properties.

In summary, VideoSAM outperforms both U-Net and SAM in complex environments involving intricate fluid behaviors and dynamic conditions, such as FC-72, Nitrogen, and Argon. However, in simpler fluid environments like Water, U-Net's architecture shines, surpassing VideoSAM in both IoU and F1 Score. These findings suggest that VideoSAM is particularly well-suited for complex scientific applications, while U-Net may be more appropriate for simpler segmentation tasks.

D. Weaknesses and Potential Solutions

Despite the promising results, VideoSAM exhibits certain weaknesses, particularly in handling simpler HSV datasets like Water. The model, which was fine-tuned primarily on complex fluid dynamics, tends to overfit to these intricate scenarios, leading to reduced performance on datasets with less dynamic behavior. To address this, we propose the following solutions:

1. Hybrid Model Architecture: A potential solution is the development of a hybrid model that integrates both traditional CNN layers, such as those in U-Net, and transformer-based layers from VideoSAM. The CNN layers could effectively handle simpler, static features, while the transformer layers focus on complex, dynamic features. This hybrid approach could leverage the strengths of both architectures, enhancing the model's ability to generalize across different types of datasets.

2. Curriculum Learning: Curriculum learning is another approach that could significantly improve VideoSAM's performance. By initially training the model on simpler datasets and gradually introducing more complex scenarios, the model

could build a robust understanding of basic features before tackling more challenging segmentation tasks. This gradual increase in data complexity during training could improve the model's generalization capabilities, making it more versatile across diverse HSV datasets.

3. Multi-Scale Feature Aggregation: Incorporating multi-scale feature aggregation into VideoSAM could enhance its ability to capture features at different levels of detail. This approach allows the model to effectively segment both large, simple structures and small, intricate ones, addressing the varying bubble sizes and textures across different HSV datasets. Multi-scale aggregation could be particularly beneficial in improving segmentation performance on simpler datasets like Water, where the model currently struggles.

V. Conclusion and Future Work

In this work, we introduced VideoSAM, a refined large vision foundation model designed for HSV segmentation, with a focus on complex boiling phenomena. VideoSAM significantly outperformed traditional models like U-Net in challenging fluid environments such as FC-72, Nitrogen, and Argon. Despite its success in intricate tasks, VideoSAM showed limitations in simpler datasets like Water, where it tended to overfit to complex scenarios.

To address this, we propose several future enhancements. These include hybrid architectures that integrate CNN layers with transformers, curriculum learning to improve generalization across data complexities, and multi-scale feature aggregation to handle varying object sizes. Expanding real-time segmentation capabilities and incorporating temporal dependencies are also key areas for future exploration.

In addition to the model, we contribute an open-source dataset for HSV segmentation, specifically curated for phase detection in boiling experiments, which will enable and facilitate further research in this domain. This dataset, along with VideoSAM's performance improvements, advances the field of HSV analysis by providing valuable resources for future studies. Further work will also involve exploring advanced fine-tuning techniques, such as LoRA, VPT, and SSF, to enhance VideoSAM's generalization and adaptability across different datasets.

References

[1] K. Rajan, H. O. Brinkhaus, M. Sorokina, A. Zielesny, and C. Steinbeck, "DECIMER-Segmentation: Automated extraction of chemical structure depictions from scientific literature," *Journal of Cheminformatics*, vol. 13, no. 1, p. 20, Mar. 2021. [Online]. Available: https://doi.org/10.1186/s13321-021-00496-1

[2] K. Brzozowski, E. Matuszyk, A. Pieczara, J. Firlej, A. Nowakowska, and M. Baranska, "Stimulated raman scattering microscopy in chemistry and life science – development, innovation, perspectives," *Biotechnology Advances*, vol. 60, p. 108003, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0734975022000994

[3] J. Zhang, J. Zhao, H. Lin, Y. Tan, and J.-X. Cheng, "High-Speed Chemical Imaging by Dense-Net Learning of Femtosecond Stimulated Raman Scattering," *The Journal of Physical Chemistry Letters*, vol. 11, no. 20, pp. 8573–8578, Oct. 2020, publisher: American Chemical Society. [Online]. Available: https://doi.org/10.1021/acs.jpclett.0c01598

[4] S. Eppel, H. Xu, M. Bismuth, and A. Aspuru-Guzik, "Computer Vision for Recognition of Materials and Vessels in Chemistry Lab Settings and the Vector-LabPics Data Set," *ACS Central Science*, vol. 6, no. 10, pp. 1743–1752, Oct. 2020, publisher: American Chemical Society. [Online]. Available: https://doi.org/10.1021/acscentsci.0c00460

[5] J. H. Seong, M. Ravichandran, G. Su, B. Phillips, and M. Bucci, "Automated bubble analysis of high-speed subcooled flow boiling images using u-net transfer learning and global optical flow," *International Journal of Multiphase Flow*, vol. 159, p. 104336, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0301932222002956

[6] M. Ravichandran, A. Kossolapov, G. M. Aguiar, B. Phillips, and M. Bucci, "Autonomous and online detection of dry areas on a boiling surface using deep learning and infrared thermometry," *Experimental Thermal and Fluid Science*, vol. 145, p. 110879, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0894177723000353

[7] S. Passoni, R. Mereu, and M. Bucci, "Integrating machine learning and image processing for void fraction estimation in two-phase flow through corrugated channels," *International Journal of Multiphase Flow*, vol. 177, p. 104871, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0301932224001484

[8] L. Zhang, C. Wang, G. Su, A. Kossolapov, G. Matana Aguiar, J. H. Seong, F. Chavagnat, B. Phillips, M. M. Rahman, and M. Bucci, "A unifying criterion of the boiling crisis," *Nature Communications*, vol. 14, no. 1, p. 2321, Apr. 2023, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41467-023-37899-7

[9] Y. Zhou, T. Zhang, S. Ji, S. Yan, and X. Li, "Dvis-daq: Improving video segmentation via dynamic anchor queries," 2024.

[10] J. Portillo-Portillo, G. Sanchez-Perez, L. K. Toscano-Medina, A. Hernandez-Suarez, J. Olivares-Mercado, H. Perez-Meana, P. Velarde-Alvarado, A. L. S. Orozco, and L. J. García Villalba, "Fassvid: Fast and accurate semantic segmentation for video sequences," *Entropy*, vol. 24, no. 7, 2022. [Online]. Available: https://www.mdpi.com/1099-4300/24/7/942

[11] Z. Zhu, L. Qiu, J. Wang, J. Xiong, and H. Peng, "Video object segmentation using multi-scale attention-based siamese network," *Electronics*, vol. 12, no. 13, 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/13/2890

[12] G. Balachandran and J. V. G. Krishnan, "Machine learning based video segmentation of moving scene by motion index using io detector and shot segmentation," *Image and Vision Computing*, vol. 122, p. 104443, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885622000725

[13] B. Gao, Y. Zhao, F. Zhang, B. Luo, and C. Yang, "Video object segmentation based on multi-level target models and feature integration," *Neurocomputing*, vol. 492, pp. 396–407, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231222004118

[14] Y. Suh, S. Chang, P. Simadiris, T. B. Inouye, M. J. Hoque, S. Khodakarami, C. Kharangate, N. Miljkovic, and Y. Won, "Vision-it: A framework for digitizing bubbles and droplets," *Energy and AI*, vol. 15, p. 100309, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666546823000812

[15] H. Zhang, Z. Tang, Y. Xie, X. Gao, and Q. Chen, "A watershed segmentation algorithm based on an optimal marker for bubble size measurement," *Measurement*, vol. 138, pp. 182–193, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0263224119301149

[16] B.-L. Chen, T.-F. Yang, U. Sajjad, H. M. Ali, and W.-M. Yan, "Deep learning-based assessment of saturated flow boiling heat transfer and two-phase pressure drop for evaporating flow," *Engineering Analysis with Boundary Elements*, vol. 151, pp. 519–537, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0955799723001376

[17] I. Malakhov, A. Seredkin, A. Chernyavskiy, V. Serdyukov, R. Mullyadzanov, and A. Surtaev, "Deep learning segmentation to analyze bubble dynamics and heat transfer during boiling at various pressures," *International Journal of Multiphase Flow*, vol. 162, p. 104402, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0301932223000253

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[19] R. Yuan, J. Xu, Q. Li, Y. Zhang, R. Feng, X. Zhang, T. Zhang, and S. Gao, "Semi-medseq: Semi-supervised semantic segmentation for medical image sequences," in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2023, pp. 1662–1669.

[20] Y. Fu, J. E. Ippolito, D. R. Ludwig, R. Nizamuddin, H. H. Li, and D. Yang, "Technical note: Automatic segmentation of ct images for ventral body composition analysis," *Medical Physics*, vol. 47, no. 11, p. 5723–5730, Oct. 2020. [Online]. Available: http://dx.doi.org/10.1002/mp.14465

[21] M. E. Rayed, S. S. Islam, S. I. Niha, J. R. Jim, M. M. Kabir, and M. Mridha, "Deep learning for medical image segmentation: State-of-the-art advancements and challenges," *Informatics in Medicine Unlocked*, vol. 47, p. 101504, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914824000601

[22] I. Rizwan I Haque and J. Neubert, "Deep learning approaches to biomedical image segmentation," *Informatics in Medicine Unlocked*, vol. 18, p. 100297, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S235291481930214X

[23] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," 2018.

[24] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," 2020.

[25] Y. Xiong, Z. Li, Y. Chen, F. Wang, X. Zhu, J. Luo, W. Wang, T. Lu, H. Li, Y. Qiao, L. Lu, J. Zhou, and J. Dai, "Efficient deformable convnets: Rethinking dynamic and sparse operator for vision applications," 2024.

[26] Z. Zuo, J. Smith, J. Stonehouse, and B. Obara, "Robust and explainable fine-grained visual classification with transfer learning: A dual-carriageway framework," 2024.

[27] A. Mahbod, G. Dorffner, I. Ellinger, R. Woitek, and S. Hatamikia, "Improving generalization capability of deep learning-based nuclei instance segmentation by non-deterministic train time and deterministic test time stain normalization," *Computational and Structural Biotechnology Journal*, vol. 23, p. 669–678, Dec. 2024. [Online]. Available: http://dx.doi.org/10.1016/j.csbj.2023.12.042

[28] I. Papadeas, L. Tsochatzidis, A. Amanatiadis, and I. Pratikakis, "Real-time semantic image segmentation with deep learning for autonomous driving: A survey," *Applied Sciences*, vol. 11, no. 19, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/19/8802

[29] R. A. Zeineldin, M. E. Karar, J. Coburger, C. R. Wirtz, and O. Burgert, "Deepseg: deep neural network framework for automatic brain tumor segmentation using magnetic resonance flair images," *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, no. 6, p. 909–920, May 2020. [Online]. Available: http://dx.doi.org/10.1007/s11548-020-02186-z

[30] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," 2018.

[31] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," 2018.

[32] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023.

[33] S. M. S. Hassan, A. Feeney, A. Dhruv, J. Kim, Y. Suh, J. Ryu, Y. Won, and A. Chandramowlishwaran, "Bubbleml: A multi-physics dataset and benchmarks for machine learning," 2023.

[34] C. Ramaswamy, Y. Joshi, W. Nakayama, and W. Johnson, "High-speed visualization of boiling from an enhanced structure," *International Journal of Heat and Mass Transfer*, vol. 45, no. 24, pp. 4761–4771, 2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0017931002001965

[35] J. Manin, S. A. Skeen, and L. M. Pickett, "Performance comparison of state-of-the-art high-speed video cameras for scientific applications," *Optical Engineering*, vol. 57, no. 12, p. 124105, 2018. [Online]. Available: https://doi.org/10.1117/1.OE.57.12.124105

[36] S. T. Thoroddsen, T. G. Etoh, and K. Takehara, "High-Speed Imaging of Drops and Bubbles," *Annual Review of Fluid Mechanics*, vol. 40, no. Volume 40, 2008, pp. 257–285, Jan. 2008, publisher: Annual Reviews. [Online]. Available: https://www.annualreviews.org/content/journals/10.1146/annurev.fluid.40.111406.102215

[37] X. Duan, B. Phillips, T. McKrell, and J. Buongiorno, "_USSynchronized High-Speed Video, Infrared Thermometry, and Particle Image Velocimetry Data for Validation of Interface-Tracking Simulations of Nucleate Boiling Phenomena," *_USBuongiorno via Chris Sherratt*, May 2013, accepted: 2014-01-13T18:09:43Z Publisher: Taylor & Francis. [Online]. Available: https://dspace.mit.edu/handle/1721.1/83904

[38] M. Schepperle, S. Junaid, and P. Woias, "Computer-vision- and deep-learning-based determination of flow regimes, void fraction, and resistance sensor data in microchannel flow boiling," *Sensors*, vol. 24, no. 11, 2024. [Online]. Available: https://www.mdpi.com/1424-8220/24/11/3363

[39] J. Lin, Y. Tang, J. Wang, and W. Zhang, "Mitigating both covariate and conditional shift for domain generalization," 2022.

[40] S. Park, O. Bastani, J. Weimer, and I. Lee, "Calibrated prediction with covariate shift via unsupervised domain adaptation," 2020.

[41] W. Yan, Y. Wang, S. Gu, L. Huang, F. Yan, L. Xia, and Q. Tao, "The domain shift problem of medical image segmentation and vendor-adaptation by unet-gan," 2019.

[42] Y. Ju, L. Wu, M. Li, Q. Xiao, and H. Wang, "A novel hybrid model for flow image segmentation and bubble pattern extraction," *Measurement*, vol. 192, p. 110861, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S026322412200149X

[43] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *Advances in Neural Information Processing Systems*, B. Schölkopf, J. Platt, and T. Hoffman, Eds., vol. 19. MIT Press, 2006. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2006/file/b1b0432ceafb0ce714426e9114852ac7-Paper.pdf

[44] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1–20, 2022. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2022.3195549

[45] W. Cho, J. Park, and T. Kim, "Complementary domain adaptation and generalization for unsupervised continual domain shift learning," 2023.

[46] C. Dunlap, C. Li, H. Pandey, N. Le, and H. Hu, "Bubbleid: A deep learning framework for bubble interface dynamics analysis," 2024.

[47] A. N. Chernyavskiy and I. P. Malakhov, "CNN-based visual analysis to study local boiling characteristics," *Journal of Physics: Conference Series*, vol. 2119, no. 1, p. 012068, Dec. 2021, publisher: IOP Publishing. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/2119/1/012068

[48] H. Hessenkemper, S. Starke, Y. Atassi, T. Ziegenhein, and D. Lucas, "Bubble identification from images with machine learning methods," *International Journal of Multiphase Flow*, vol. 155, p. 104169, Oct. 2022, arXiv:2202.03107 [physics]. [Online]. Available: http://arxiv.org/abs/2202.03107

[49] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Wang, L. Wang, J. Gao, and Y. J. Lee, "Segment everything everywhere all at once," 2023.

[50] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," 2022.

[51] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," 2020.

[52] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021.

[53] V. Maquiling, S. A. Byrne, D. C. Niehorster, M. Nyström, and E. Kasneci, "Zero-shot segmentation of eye features using the segment anything model (sam)," *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, vol. 7, no. 2, p. 1–16, May 2024. [Online]. Available: http://dx.doi.org/10.1145/3654704

[54] Y. Huang, X. Yang, L. Liu, H. Zhou, A. Chang, X. Zhou, R. Chen, J. Yu, J. Chen, C. Chen, S. Liu, H. Chi, X. Hu, K. Yue, L. Li, V. Grau, D.-P. Fan, F. Dong, and D. Ni, "Segment anything model for medical images?" *Medical Image Analysis*, vol. 92, p. 103061, Feb. 2024. [Online]. Available: http://dx.doi.org/10.1016/j.media.2023.103061

[55] Y. Zhang, Z. Shen, and R. Jiao, "Segment anything model for medical image segmentation: Current applications and future directions," 2024.

[56] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, Jan. 2024. [Online]. Available: http://dx.doi.org/10.1038/s41467-024-44824-z

[57] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, "Segment anything model for medical image analysis: An experimental study," *Medical Image Analysis*, vol. 89, p. 102918, Oct. 2023. [Online]. Available: http://dx.doi.org/10.1016/j.media.2023.102918

[58] L. P. Osco, Q. Wu, E. L. de Lemos, W. N. Gonçalves, A. P. M. Ramos, J. Li, and J. Marcato, "The segment anything model (sam) for remote sensing applications: From zero to one shot," *International Journal of Applied Earth Observation and Geoinformation*, vol. 124, p. 103540, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1569843223003643

[59] R. Yang, G. He, R. Yin, G. Wang, Z. Zhang, T. Long, Y. Peng, and J. Wang, "A novel weakly-supervised method based on the segment anything model for seamless transition from classification

to segmentation: A case study in segmenting latent photovoltaic locations," *International Journal of Applied Earth Observation and Geoinformation*, vol. 130, p. 103929, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1569843224002838

[60] X. Zhou, F. Liang, L. Chen, H. Liu, Q. Song, G. Vivone, and J. Chanussot, "Mesam: Multiscale enhanced segment anything model for optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[61] L. Ding, K. Zhu, D. Peng, H. Tang, K. Yang, and L. Bruzzone, "Adapting segment anything model for change detection in vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, p. 1–11, 2024. [Online]. Available: http://dx.doi.org/10.1109/TGRS.2024.3368168

[62] W. Ji, J. Li, Q. Bi, T. Liu, W. Li, and L. Cheng, "Segment anything is not always perfect: An investigation of sam on different real-world applications," 2023.

[63] Z. Yang, Y. Wei, and Y. Yang, "Associating objects with transformers for video object segmentation," 2021.

[64] Y. Cheng, L. Li, Y. Xu, X. Li, Z. Yang, W. Wang, and Y. Yang, "Segment and track anything," 2023.

[65] A. Maalouf, N. Jadhav, K. M. Jatavallabhula, M. Chahine, D. M. Vogt, R. J. Wood, A. Torralba, and D. Rus, "Follow anything: Open-set detection, tracking, and following in real-time," 2024.

[66] J. Yang, M. Gao, Z. Li, S. Gao, F. Wang, and F. Zheng, "Track anything: Segment anything meets videos," 2023.

[67] T. Zhou, W. Luo, Q. Ye, Z. Shi, and J. Chen, "Sam-pd: How far can sam take us in tracking and segmenting anything in videos by prompt denoising," 2024.

[68] C. Dunlap, H. Pandey, and H. Hu, "Supervised and unsupervised learning models for detection of critical heat flux during pool boiling," *ASME 2022 Heat Transfer Summer Conference*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252504004

[69] C. T. Rueden, J. Schindelin, M. C. Hiner, B. E. DeZonia, A. E. Walter, E. T. Arena, and K. W. Eliceiri, "Imagej2: Imagej for the next generation of scientific image data," 2017.