Biomedical SAM-2: Segment Anything in Biomedical Images and Videos

Zhiling Yan[*]
Lehigh University
zhy423@lehigh.edu

Weixiang Sun[*]
Northeastern University
kennysweson@gmail.com

Rong Zhou[*]
Lehigh University
roz322@lehigh.edu

Zhengqing Yuan[*]
University of Notre Dame
zyuan2@nd.edu

Kai Zhang
Lehigh University
kaz321@lehigh.edu

Yiwei Li
University of Georgia
yl80817@uga.edu

Sekeun Kim
Massachusetts General Hospital
and Harvard Medical School
skim207@mgh.harvard.edu

Sifan Song
Massachusetts General Hospital
and Harvard Medical School
ssong25@mgh.harvard.edu

Hui Ren
Massachusetts General Hospital
and Harvard Medical School
hren2@mgh.harvard.edu

Tianming Liu
University of Georgia
tliu@cs.uga.edu

Quanzheng Li
Massachusetts General Hospital
and Harvard Medical School
li.quanzheng@mgh.harvard.edu

Xiang Li
Massachusetts General Hospital
and Harvard Medical School
xli60@mgh.harvard.edu

Lifang He
Lehigh University
lih319@lehigh.edu

Lichao Sun[†]
Lehigh University
lis221@lehigh.edu

Abstract

Medical image segmentation and video object segmentation are essential for diagnosing and analyzing diseases by identifying and measuring biological structures. Recent advances in natural domain have been driven by foundation models like the Segment Anything Model 2 (SAM-2). To explore the performance of SAM-2 in biomedical applications, we designed three evaluation pipelines for single-frame 2D image segmentation, multi-frame 3D image segmentation and multi-frame video segmentation with varied prompt designs, revealing SAM-2's limitations in medical contexts. Consequently, we developed BioSAM-2, an enhanced foundation model optimized for biomedical data based on SAM-2. Our experiments show that BioSAM-2 not only surpasses the performance of existing state-of-the-art foundation models but also matches or even exceeds specialist models, demonstrating its efficacy and potential in the medical domain. The code has been made publicly available at https://github.com/ZhilingYan/Biomedical-SAM-2.

[*]Equal Contributions.
[†]Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

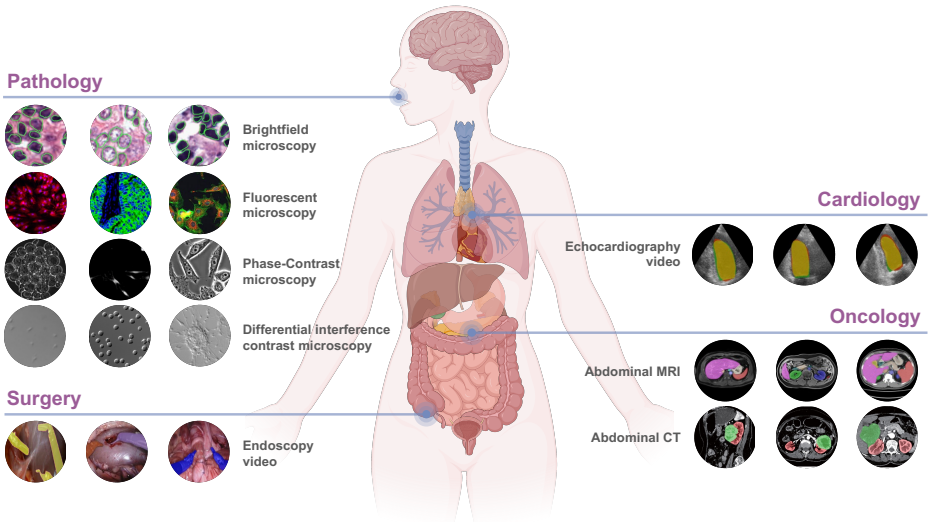

Figure 1: Overview of biomedical tasks in BioSAM-2.

1 Introduction

Medical image segmentation is crucial for identifying biological structures and measuring their morphology, aiding in the diagnosis and analysis of various diseases [1, 2]. Despite numerous advancements in medical imaging technologies, segmentation remains a formidable challenge due to the complexity of medical images and the extensive manual effort required for accurate annotation [3, 4]. Traditional methods often necessitate detailed manual annotation, which is not only time-consuming but also prone to human error [5].

Recently, the emergence of segmentation foundation models, such as the Segment Anything Model (SAM) [6] has driven significant advancements in natural image segmentation field. SAM demonstrates impressive zero-shot segmentation performance with prompt inputs, showing remarkable versatility and setting a new standard across various segmentation tasks. To further adapt SAM's capabilities to the medical field, numerous works have been proposed [7–10], with MedSAM [7] being a representative one. MedSAM modifies the SAM architecture by incorporating domain-specific knowledge to address the unique challenges of medical images, such as varying contrast, noise levels, and artifacts [11]. MedSAM has demonstrated significant improvements in segmentation performance on medical images, leveraging SAM's foundation while tailoring it for medical applications.

Recognizing the need to extend these capabilities to more complex scenarios, SAM-2 [12] is developed to expand SAM's functionality to include video inputs. This extension enables SAM-2 to process temporal sequences of images, making it suitable for tasks that require understanding spatial continuity over multiple frames. By handling both spatial and temporal dimensions, SAM-2 has demonstrated impressive zero-shot performance on natural image and video segmentation task.

However, SAM-2's potential on medical segmentation tasks has not yet been fully explored. We conducted a comprehensive evaluation to investigate its capability. Specifically, we assessed the performance of different SAM-2 variants across 8 medical modalities and 22 objects of interest. We designed three evaluation pipelines for single-frame 2D image segmentation, multi-frame 3D image segmentation and multi-frame video segmentation, respectively, incorporating diverse prompt designs. To further evaluate its performance, we compared it against various baseline models, including CNN-based, Transformer-based, and SSM-based models, using a range of metrics. Our findings indicate that SAM-2 cannot be directly utilized in medical domain. Main reasons are the significant domain gap between natural and medical data and its inability to associate segmentation regions with meaningful semantic classes. In other words, SAM-2 cannot perform semantic segmentation on medical data, which limits its application in computer-aided diagnosis.

Building on these observations, we further introduce BioSAM-2, a refined foundation model that significantly enhances the segmentation performance of SAM-2 on biomedical images and videos. BioSAM-2 incorporates a memory mechanism and a stream processing architecture, which is the same as SAM-2. This model can handle multi-frame segmentation tasks by retaining information

from past predictions, allowing it to make accurate predictions on slices without explicit prompts. Experimental results demonstrate that BioSAM-2 consistently outperforms the state-of-the-art (SOTA) segmentation foundation model [6, 12], while on par with, or even surpassing the performance of specialist models [13–16] trained on medical data from the same modality. These findings underscore the potential of BioSAM-2 as a new paradigm for versatile medical image and video segmentation.

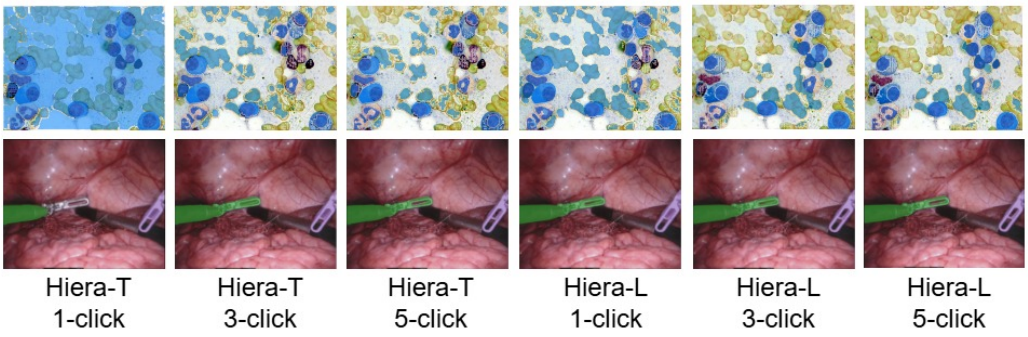

Figure 2: Image segmentation results of tiny SAM-2 (left) and large SAM-2 (right) based on different segmentation prompts.

Our contributions can be summarized as follows:

- We have developed three evaluation pipelines tailored for single-frame biomedical images, multi-frame biomedical images and multi-frame biomedical videos within 8 medical modalities and 22 objects of interest. These pipelines comprehensively assess the performance of SAM-2 in biomedical applications.
- To enhance the adaptability of SAM-2 in the biomedical domain, we introduce BioSAM-2, an optimized foundational model achieved by fine-tuning the original SAM-2. This refinement significantly boosts the segmentation performance of SAM-2. Remarkably, without any prompts, our automated segmentation consistently outperforms competitive SOTA foundation methods by a large margin.
- The results of experiments show that BioSAM-2 matches or even exceeds the performance of specialized models trained on medical data of the same modality. These findings highlight the potential of BioSAM-2 as a novel paradigm for versatile medical applications.

By integrating BioSAM-2 with medical image segmentation tasks, we anticipate significant improvements in segmentation accuracy and annotation efficiency, ultimately contributing to better clinical outcomes and facilitating the adoption of AI in medical imaging. This research aims to push the boundaries of current medical image segmentation techniques and explore the full potential of advanced AI models like SAM-2 in handling the complexities of medical imaging data.

2 Related Work

Medical Image Segmentation. CNN-based and Transformer-based models have significantly advanced medical image segmentation. U-Net [17], a notable CNN-based approach, features a symmetrical encoder-decoder architecture with skip connections to preserve details. Enhancements [14] like the self-configuring nnU-Net [13] have demonstrated robust performance across various medical segmentation challenges. In Transformer-based models, TransUnet [18] integrates the Vision Transformer (ViT)[19] for feature extraction and pairs it with a CNN for decoding, effectively processing global information. UNETR [15], and Swin-UNet [20] combine Transformer architectures with U-Net to enhance 3D imaging analysis, exploring Swin Vision Transformer blocks [21].SSM-based models like U-Mamba [22] have also been introduced for efficient long-sequence data analysis in medical imaging. Recently, SAM [6], a vision foundation model pre-trained on 1 billion masks, demonstrated impressive zero-shot capability across various segmentation tasks. Inspired by SAM's performance in natural images, adaptations quickly emerged for medical segmentation [23–26]. MedSAM [7] fine-tuned SAM on over 200,000 masks across 11 modalities, while SAM-Med2D [27] used comprehensive prompts for 2D medical images. SAMed [23] and MA-SAM [28] employed PETL [29] for fine-tuning, outperforming several existing medical segmentation methods.

Medical Video Object Segmentation. A large number of semantic segmentation models rely on single images to identify objects in a scene. This can lead to spatially and temporally inconsistent

predictions especially in multi-frame videos that require temporal context. To address this, Space Time Memory Networks (STM) [30] and its variants [31–33] use a memory network to extract vital information from a time-based buffer composed of all previous video sequences. Building on this methodology, DPSTT [34] integrates a memory bank with decoupled transformers to track temporal lesion movement in medical ultrasound videos. However, DPSTT requires substantial data augmentation to avoid overfitting and suffers from low processing speed. Subsequently, FLANet [35] introduced a frequency and location feature aggregation network, involving a large amount of memory occupancy. Optical flow methods for surgical videos [36, 37] are limited to using features between pairs of images and cannot leverage extended temporal context. Other methods employ a combination of 2D encoders and 3D convolutional layers in the temporal decoder [38] and Convolutional Long short-term Memory cells [39]. Alternative approaches enforce temporal consistency through a loss function during training [40] or use architectures that combine high and low frame rate model branches to integrate temporal context from different parts of the video [41].

The recently introduced SAM-2 [12] extended SAM to 3D, enhancing its capability to "segment anything in videos". Specifically, it is equipped with a memory that stores information of objects and previous interactions, allowing it to generate and correct masklet predictions throughout the video.

3 Method

3.1 Preliminary Study of SAM-2

Segment Anything Model 2 (SAM-2) is a unified transformer-based model for both image and video segmentation. For each video frame, the segmentation prediction leverages the current prompt and previously observed memories. Videos are processed sequentially, with each frame handled individually by **image encoder**, while **memory attention** conditions current frame features on past frames and predictions. The **mask decoder**, which can optionally take input prompts, predicts the segmentation mask for that frame. Finally, a **memory encoder** transforms the predictions and image embeddings into a form for future frames, ensuring temporal consistency among multiple frames.

The vision transformer in the image encoder is pretrained using the hierarchical masked autoencoder model Hiera [42], enabling multiscale feature decoding. Memory attention conditions the current frame features on past frames' features and predictions. Multiple transformer blocks are stacked, with the first block taking the current frame's image encoding as input. Each block performs self-attention, followed by cross-attention to memories of frames and object pointers stored in a memory bank. In SAM-2, prompts are encoded with positional encoding and two learnable tokens specifying foreground and background. The mask decoder comprises bi-directional transformer blocks that update prompt and frame embeddings. The model predicts multiple masks per frame, and in cases of ambiguity without subsequent prompt, it propagates only the mask with the highest predicted IoU. Additionally, an auxiliary prediction head determines the presence of the target object in the current frame. Finally, the memory encoder consolidates this process by downsampling the output mask using a convolutional module, which is then summed with the unconditioned frame embedding from the image encoder. This integrated data is retained in a memory bank, preserving essential information about past predictions for the target object throughout the sequence.

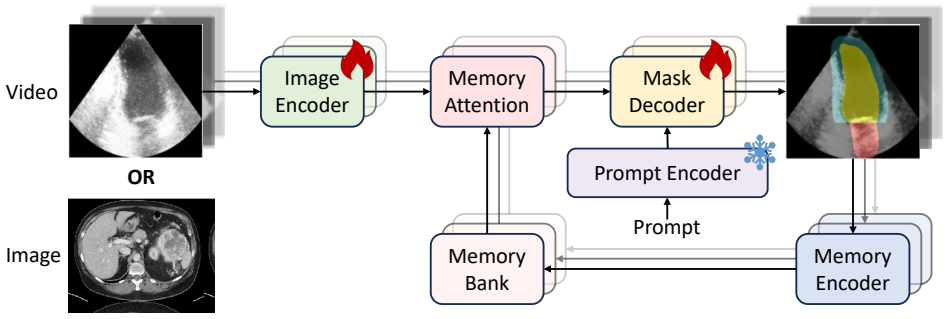

Figure 3: The workflow of BioSAM-2. We adapt SAM-2 for medical image and video segmentation by freezing the prompt encoder and only finetuning the image encoder and mask decoder.

3.2 Medical Applications of SAM-2

Building on the impressive zero-shot capabilities demonstrated by SAM-2 in natural images and videos, we explore its performance in medical applications. Specifically, we designed three pipelines, single-frame image segmentation, multi-frame image segmentation and multi-frame video segmentation to assess SAM-2's ability to handle zero-shot segmentation tasks within these medical contexts.

Single-frame 2D Image Segmentation. We design single-frame image segmentation in a non-iterative manner in which all prompts are determined without feedback from any prior predictions.

Firstly, we automate the generation of point prompts through a specific strategy. One point is added as the prompt by randomly selecting it from the initial mask. Given a set of candidate points P derived from the initial mask, the selected point p is:

$$p = random(P) \tag{1}$$

As highlighted in the SAM-2 documentation [12], employing a single point prompt can lead to segmentation ambiguity, as the model may associate the prompt with multiple valid masks without being able to discern the intended one. Although SAM-2 incorporates an ambiguity-resolving module that generates multiple masks and ranks them based on confidence scores, the use of multiple point prompts significantly mitigates this issue. Thus, we also assessed the performance of SAM-2 with additional randomly selected points from the initial mask candidates.

This sampling-from-mask method treats the initial segmentation mask as a reliable outcome and seeks to enhance segmentation accuracy by exploiting prompt selection invariance and incorporating additional point prompts.

Multi-frame 3D Image Segmentation. For multi-frame 3D image segmentation, we initially set the bounding box prompts on the middle slice of the image volume, following [43, 44]. The middle slice is chosen as the starting point in 3D image segmentation due to its tendency to contain the largest object and the greatest number of semantic classes among all slices in the conventional axial view. This strategic choice enhances the accuracy and reliability of subsequent segmentation. Once the middle frame is annotated, SAM-2 facilitates the propagation of annotations to the surrounding slices that do not have prompts. This is achieved through the video segmentation feature integrated into SAM-2. The process involves first propagating the annotations backward from the middle slice to the first slice, followed by a forward propagation from the middle slice to the last slice in the volume. This bidirectional propagation ensures consistent segmentation across the entire image volume.

Multi-frame Video Segmentation. In evaluating video segmentation, we employ a hybrid mode of offline and online evaluation. Specifically, we select the first n frames as interacted frames, on which click cues are added. Multiple click cues enable the model to more accurately determine the boundaries of objects, achieving higher segmentation accuracy. Additionally, multiple interacted frames effectively resolve the issue identified in SAM-2 where the model fails to track objects appearing in unmarked frames, making it more reasonable and effective for video scenarios. Overall, our evaluation traverses the video only once, subsequently yielding the final segmentation results.

To summarize, when directly employing SAM-2 for medical image and video segmentation, the generated mask can be ambiguous and necessitates multiple prompts or iterations for prediction and correction. Despite its promising potential, SAM-2 has encountered challenges in delivering satisfactory segmentation results across various medical image/video segmentation tasks. Additionally, since the video training data for SAM-2 predominantly consists of high-resolution footage, it may become entirely ineffective when dealing with low-resolution medical videos. In light of these limitations, the objective of this study is to develop a robust segmentation foundation model capable of effectively addressing a wide range of biomedical segmentation targets.

3.3 BioSAM-2: Dedicated biomedical segmentation foundation model

To tailor SAM-2 for medical segmentation, selecting the appropriate network components for fine-tuning is crucial. The architecture of SAM-2 comprises several key elements: image encoder, prompt encoder, memory attention, mask decoder, and memory encoder. It is possible to fine-tune any combination of these components. For our adaptation, the prompt encoder, which processes the information from given prompts, is retained from the pre-trained model and is frozen to maintain its original functionality. Conversely, the image encoder and mask decoder are fine-tuned to enhance SAM-2's suitability for the medical imaging domain. This adaptation strategy is depicted in Figure 3.

For the image encoder, we opted for a tiny-sized configuration to strike a balance between computational cost and performance efficacy. We initiated training for SAM-2 from the official checkpoint, conducting separate sessions for image and video data. During training, we employed the AdamW optimizer [45] and implemented layer decay strategies [46] on the image encoder to enhance its learning efficiency. In terms of the mask decoder, its configuration is simplified to generate a single mask per frame, given that the provided prompt distinctly identifies the expected segmentation target.

For loss design, we supervise the model's predictions using a combination of dice loss and binary cross-entropy (BCE) loss for the mask prediction. Specifically, p_i and g_i are the predicted and ground truth pixel values, respectively, and N is the total number of pixels. The dice loss is defined as:

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2 \times \sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i} \tag{2}$$

and BCE loss is:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^{N} [g_i \log(p_i) + (1 - g_i) \log(1 - p_i)] \tag{3}$$

During training, the model parameters are optimized using a combination of these losses:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{dice}} + \beta \mathcal{L}_{\text{BCE}} \tag{4}$$

If the ground-truth does not contain a mask for a frame, we do not supervise the mask outputs, but supervise the occlusion prediction head that predicts whether there should exist a mask in the frame.

4 Experiments

4.1 Biomedical Image Segmentation

4.1.1 Datasets

Table 1: Biomedical image dataset information.

Dataset	Dimension	#Training Image	#Testing Image	#Targets
Abdomen CT	3D	50 (4,794 slices)	50 (10,894 slices)	13
Abdomen MRI	3D	60 (5,615 slices)	50 (3357 slices)	13
Endoscopy Images	2D	1800	1200	7
Microscopy Images	2D	1000	101	1

To assess the performance and scalability of BioSAM-2, we utilize four datasets, including Abdomen CT dataset [47], Abdomen MR dataset [48], Endoscopy dataset [49] and Microscopy dataset [50].

Abdomen 3D images. For evaluating BioSAM-2, two datasets are implemented. The first is Abdomen CT [47] from the MICCAI 2022 FLARE challenge, which includes segmentation of 13 abdominal organs from 50 CT scans in both the training and testing sets. The organs include the liver, spleen, pancreas, kidneys, stomach, gallbladder, esophagus, aorta, inferior vena cava, adrenal glands, and duodenum. The second dataset, Abdomen MR [48] from the MICCAI 2022 AMOS Challenge, involves the same 13 organs with 60 MRI scans for training and 50 for testing.

Endoscopy Images. From MICCAI 2017 EndoVis Challenge [49], this dataset focuses on instrument segmentation in endoscopy images, featuring seven distinct instruments, including the large needle driver, prograsp forceps, monopolar curved scissors, cadiere forceps, bipolar forceps, vessel sealer, and a drop-in ultrasound probe. The dataset is split into 1800 training frames and 1200 testing frames.

Microscopy images. This dataset, from the NeurIPS 2022 Cell Segmentation Challenge [50], is used for cell segmentation in microscopy images, consisting of 1000 training images and 101 testing images. Following U-Mamba [22], we address this as a semantic segmentation task, focusing on cell boundaries and interiors rather than instance segmentation.

4.1.2 Experimental Setup

We compare BioSAM-2 with nnUNet [13], SegResNet [14], UNETR [15], SwinUNETR [16], and U-Mamba [22]. Each model is trained with its recommended optimizer, finetuned on each dataset,

Table 2: Results summary of 2D image segmentation tasks: instruments segmentation in endoscopy images, and cell segmentation in microscopy images.

Methods	Instruments in Endoscopy		Cells in Microscopy	
	DSC	NSD	F1	NSD
nnU-Net	0.6264±0.3024	0.6412±0.3074	0.5383±0.2657	**0.8332±0.1611**
SegResNet	0.5820±0.3268	0.5968±0.3303	0.5411±0.2633	0.7944±0.2356
UNETR	0.5017±0.3201	0.5168±0.3235	0.4357±0.2572	0.8201±0.2263
SwinUNETR	0.5811±0.3176	0.5973±0.3209	0.3880±0.2664	0.7981±0.1920
U-Mamba	**0.6303±0.3067**	**0.6451±0.3104**	0.5607±0.2784	0.8288±0.1706
SAM	0.1583±0.3300	0.1600±0.3313	0.3249±0.2285	0.3696±0.2544
SAM-2 (Hiera-T, 1-click)	0.4115±0.4092	0.4227±0.4189	0.0654±0.1220	0.0720±0.1372
SAM-2 (Hiera-T, 3-click)	0.5215±0.3802	0.5349±0.3864	0.3436±0.2400	0.3911±0.2640
SAM-2 (Hiera-T, 5-click)	0.5382±0.3568	0.5520±0.3616	0.3566±0.2496	0.4070±0.2761
SAM-2 (Hiera-L, 1-click)	0.4416±0.4217	0.4523±0.4312	0.0799±0.1540	0.0877±0.1722
SAM-2 (Hiera-L, 3-click)	0.5354±0.3750	0.5497±0.3813	0.3217±0.2481	0.3720±0.2750
SAM-2 (Hiera-L, 5-click)	0.5479±0.3629	0.5623±0.3681	0.3352±0.2598	0.3876±0.2849
BioSAM-2 (Hiera-T)	0.6251±0.2897	0.6427±0.3095	**0.5792±0.2666**	0.7436±0.2104

Table 3: Results summary of 3D organ segmentation on abdomen CT and MR datasets.

Methods	Organs in Abdomen CT		Organs in Abdomen MR	
	DSC	NSD	DSC	NSD
nnU-Net	0.8615±0.0790	0.8972±0.0824	0.8309±0.0769	0.8996±0.0729
SegResNet	0.7927±0.1162	0.8257±0.1194	0.8146±0.0959	0.8841±0.0917
UNETR	0.6824±0.1506	0.7004±0.1577	0.6867±0.1488	0.7440±0.1627
SwinUNETR	0.7594±0.1095	0.7663±0.1190	0.7565±0.1394	0.8218±0.1409
U-Mamba	**0.8638±0.0908**	**0.8980±0.0921**	**0.8501±0.0732**	**0.9171±0.0689**
SAM	0.2957±0.2761	0.2967±0.3018	0.3710±0.4321	0.3712±0.4321
SAM-2 (Hiera-T)	0.4431±0.2690	0.3875±0.2544	0.5534±0.2609	0.5572±0.2683
SAM-2 (Hiera-L)	0.4744±0.2472	0.4208±0.2306	0.5492±0.2315	0.5526±0.2341
BioSAM-2 (Hiera-T)	0.7632±0.0836	0.8568±0.0732	0.7439±0.0629	0.8317±0.0746

and applies the default preprocessing from nnUNet. SAM and SAM-2 performance is assessed by directly inferring image masks, and for SAM-2, we test two sizes using three prompt types.

We use an unweighted combination of Dice loss and cross-entropy loss across all datasets and employ the AdamW optimizer with an initial learning rate of 1e-4. Training is conducted for 200 epochs with a batch size of 8. For evaluation metrics, we employ the Dice Similarity Coefficient (DSC) and Normalized Surface Distance (NSD) to assess performance in abdominal multi-organ segmentation for CT and MR scans, as well as instrument segmentation in endoscopy images. For the cell segmentation task, we utilize the F1 score and NSD.

4.1.3 Results

As illustrated in Table 2 and Table 3, we conducted a thorough analysis of various SAM-2 variants. The results indicate a clear performance improvement with an increased number of clicks, particularly noticeable in microscopy datasets where the F1 metric improved from 0.0654 to a maximum of 0.3566. Similarly, significant progress was observed in other datasets. For instance, the DSC score of endoscopy dataset rises from a minimum of 0.4115 to 0.5382, and the NSD score increases from 0.4227 to 0.5520. Furthermore, we evaluated SAM-2's performance across two model sizes. The larger model size outperformed the tiny variant in terms of majority DSC and NSD scores, given the same prompt. This suggests that a larger SAM-2 model possesses superior segmentation capabilities.

When comparing the results of zero-shot of SAM-2 to other fine-tuned models specifically designed for medical image segmentation (e.g. nnU-Net), it is evident that the zero-shot performance is inferior. This discrepancy highlights that despite SAM2's robust transfer learning capabilities, there remains significant room for improvement in the medical imaging domain. This underscores the necessity of fine-tuning of SAM-2 to achieve optimal performance. In addition, it is observed that when SAM-2

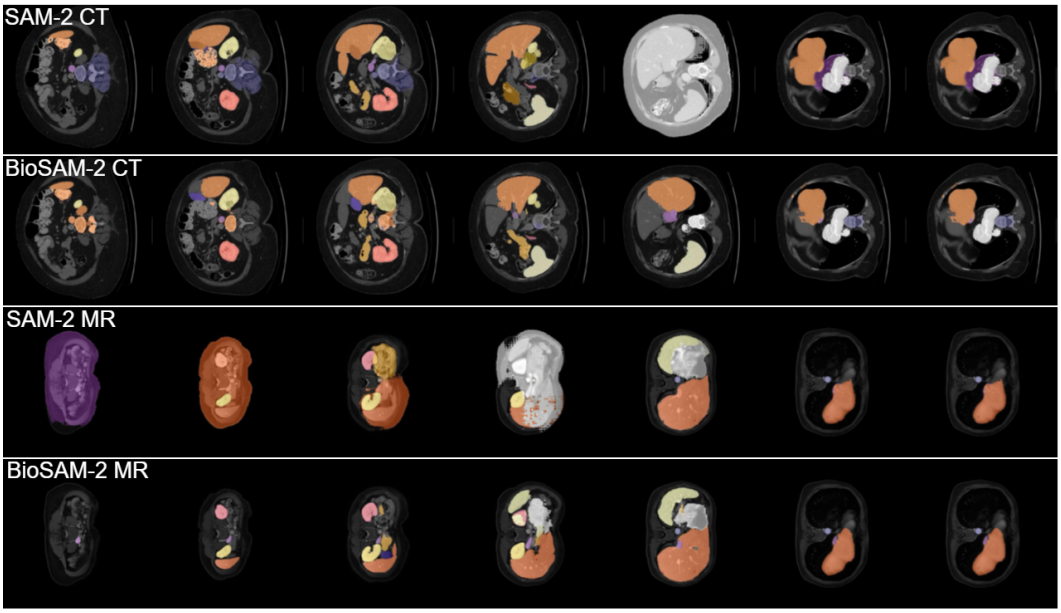

Figure 4: The visualization results of BioSAM-2 and SAM-2 on two medical scenarios.

employs proper prompts, SAM-2's zero-shot results outperform those of SAM, even though SAM utilizes its largest version, SAM_h. It reinforces the advantage of SAM-2's advanced design and adaptability, compared with SAM.

Table 2 and Table 3 also show the performance of our proposed method BioSAM-2. Comparative analysis between BioSAM-2 and SAM-2 reveals substantial improvements with BioSAM-2, achieving enhancements ranging from a minimum of 0.0772 to a maximum of 0.5138. This shows BioSAM-2's superior performance on biomedical imaging segmentation and demonstrates SAM-2's significant potential in this domain. SAM-2, being a universal model, requires adaptation, particularly in the medical field, due to its general-purpose design rather than being specialized. Its limited knowledge base on medical datasets and the constrained number of output masks are key factors. While SAM-2 can effectively segment image-level instances, it struggles with accurately segmenting class-level instances. For example, SAM-2 can easily delineate the boundaries of two cells but cannot determine if they belong to the same class, which is also shown in Figure 4. These limitations impact SAM-2's performance, especially in multi-class medical segmentation datasets. BioSAM-2 thus plays a critical role in bridging these gaps, enhancing the model's capability to handle the intricacies of medical image segmentation tasks.

Finally, according to Table 2, BioSAM-2 achieved a DSC score of 0.6251 and an NSD score of 0.6427 on the endoscopy dataset. On the microscopy dataset, it obtained an F1 score of 0.5792 and an NSD score of 0.7436. These results surpass most competing methods and are comparable to state-of-the-art models. This performance demonstrates BioSAM-2's exceptional capability in medical image segmentation, underscoring its potential to deliver high-quality results.

4.2 Biomedical Video Segmentation

4.2.1 Datasets

Table 4: Biomedical video dataset information.

Dataset	#Videos	#Frames/Video	#Train Video	#Test Video	#Targets
EchoNet-Dynamic	10,030	100-250	7,465	1,288	1
EndoVis 2018	19	300	15	4	7

To validate the performance of SAM-2 on biomedical video, we selected two datasets from medical scenarios, which include EndoVis 2018 [51] and EchoNet-Dynamic [52].

EndoVis 2018. This dataset is from Robotic Scene Segmentation Challenge [51]. It consists of video data from 16 robotic nephrectomy procedures performed using da Vinci Xi systems in porcine labs. The data is subsampled to 2 Hz to manage labeling costs, resulting in 149 frames per procedure after removing sequences with minimal motion. Each frame, presented in a resolution of 1280×1024, includes images from both left and right eye cameras along with stereo camera calibration parameters.

EchoNet-Dynamic. The EchoNet-Dynamic dataset [53] featuring 10,030 labeled echocardiogram videos was collected from routine clinical care at Stanford University Hospital. This dataset provides a unique resource for studying cardiac motion and chamber sizes, which is crucial for diagnosing cardiovascular diseases. Each video in the dataset captures the heart's dynamics from the apical-4-chamber view, meticulously cropped and masked to eliminate any extraneous text and external information, ensuring a focus solely on the cardiac imaging area. The videos have been uniformly resized to 112×112 pixels using cubic interpolation to standardize the dataset.

Table 5: $\mathcal{J}\&\mathcal{F}$ scores (%) of SAM-2 and BioSAM-2.

Method	EndoVis 2018					EchoNet-Dynamic				
	Box Prompt	1-click	3-clicks	5-clicks	GT-mask	Box Prompt	1-click	3-clicks	5-clicks	GT-mask
SAM-2 (Hiera-T)	50.2	58.4	69.9	71.4	71.6	50.1	6.6	56.2	69.8	70.0
SAM-2 (Hiera-S)	51.4	59.7	71.3	72.6	73.8	50.9	7.3	58.4	70.2	71.5
SAM-2 (Hiera-B+)	51.7	60.1	70.8	73.3	74.2	51.6	10.0	65.0	71.6	72.8
SAM-2 (Hiera-L)	52.5	60.9	71.2	74.8	77.4	55.4	7.1	65.9	67.3	73.9
BioSAM-2	**59.3**	-	-	-	-	**58.2**	-	-	-	-

Table 6: Different SAM-2 variants' $\mathcal{J}\&\mathcal{F}$ scores (%) under zero-shot conditions over multi-interacted frames under 3-clicks prompt.

Method	EndoVis 2018					EchoNet-Dynamic				
	1-frame	2-frames	4-frames	6-frames	8-frames	1-frame	2-frames	4-frames	6-frames	8-frames
SAM-2 (Hiera-T)	69.9	69.9	72.5	74.9	76.6	56.2	57.0	63.5	68.1	71.2
SAM-2 (Hiera-S)	71.3	71.0	73.0	75.4	77.8	58.4	58.5	64.0	69.5	70.9
SAM-2 (Hiera-B+)	70.8	71.2	74.1	77.6	78.2	65.0	64.9	67.9	70.0	73.4
SAM-2 (Hiera-L)	71.2	73.0	75.2	78.0	79.9	65.9	66.1	70.3	72.0	72.5

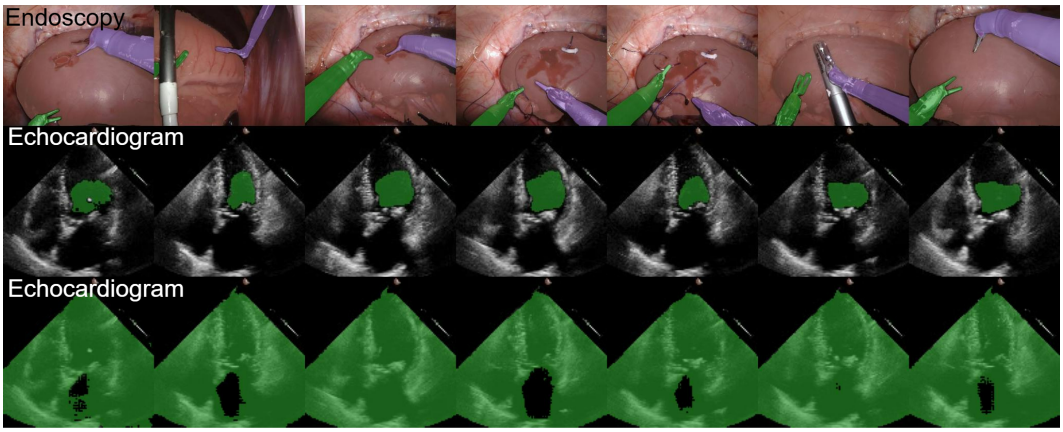

Figure 5: The visualization results of SAM-2 on two medical scenarios. The last row is the failure example of SAM-2 in 1-click application on low-resolution ultrasound videos.

4.2.2 Experimental Setup

In our evaluation of SAM-2 under the condition of zero-shot, we chose Jaccard and F-Score ($\mathcal{J}\&\mathcal{F}$) as our evaluation metrics. The Jaccard index describes the Intersection over Union (IoU) between the predicted mask and the ground truth (gt), while the F-Score quantifies the alignment between the boundaries of the predicted mask and the gt boundaries.

4.2.3 Results

The most significant breakthrough of SAM-2 compared to SAM is its capability to support video tracking of internal objects. As illustrated in Table 5, we conducted a detailed and comprehensive evaluation of the capabilities of SAM-2 in zero-shot segmentation of medical videos. The results indicate that click counts effectively enhance the accuracy of the outcomes. Just a few additional point prompts enable the model to accurately delineate the object boundaries. Notably, the EchoNet Dynamic [52] showed the most significant improvements. As shown in Figure 5, with only one click, SAM-2 segmented the imaging sector area. However, increasing the number of clicks resulted in

a substantial rise in the $\mathcal{J}\&\mathcal{F}$ scores from single digits to over 70. Moreover, the model generally performs better with larger sizes when the number of clicks remains the same. Additionally, we explored the performance of BioSAM-2 on two datasets using box prompts. BioSAM-2 outperforms all SAM-2 variants on both datasets, achieving scores of 59.3 and 58.2, respectively. This demonstrates the potential of BioSAM-2 in medical video segmentation tasks.

A major issue with SAM-2 video segmentation is that if the target object does not appear in the annotated frame, subsequent tracking and segmentation within the video are unfeasible. We addressed this problem by increasing the number of interacted frames. The results in Table 6 demonstrate the enhancements achieved by this method. With the addition of interacted frames, SAM-2 showed improved performance under a 3-click scenario, even surpassing results obtained using only the Ground Truth mask on the first frame. We observed that even if the same object was repeatedly marked in the interacted frames, it still enhanced the subsequent segmentation results. We hypothesize that this injection of information allows the model to better recognize the same object from different perspectives, thereby achieving better outcomes. Overall, while SAM-2 demonstrates some capability in zero-shot segmentation of medical videos, its lack of training in medical content renders it somewhat perplexed in certain medical scenarios. This underscores the importance of training BioSAM-2 on video data in subsequent efforts.

5 Conclusion

In conclusion, our development of BioSAM-2 represents a significant advancement in biomedical domain. Through the implementation of three specialized evaluation pipelines designed for biomedical images and videos, we have rigorously assessed the performance of SAM-2 across diverse medical scenarios and objects of interest. Our results indicate that the enhanced BioSAM-2 not only outperforms current state-of-the-art foundation methods, but also surpasses the performance of some specialized models trained specifically for the same medical modalities. These findings affirm the potential of BioSAM-2 as a novel biomedical segmentation approach towards more efficient, accurate, and adaptable diagnostic technologies.

Acknowledgements

In this work, Prof. Lichao Sun was partially supported by the National Science Foundation Grants CRII-2246067, ATD-2427915, NSF POSE-2346158, and Lehigh Grant FRGS00011497.

References

[1] Alireza Norouzi, Mohd Shafry Mohd Rahim, Ayman Altameem, Tanzila Saba, Abdolvahab Ehsani Rad, Amjad Rehman, and Mueen Uddin. Medical image segmentation methods, algorithms, and applications. *IETE Technical Review*, 31(3):199–213, 2014.

[2] Getao Du, Xu Cao, Jimin Liang, Xueli Chen, and Yonghua Zhan. Medical image segmentation based on u-net: A review. *Journal of Imaging Science & Technology*, 64(2), 2020.

[3] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023.

[4] KKD Ramesh, G Kiran Kumar, K Swapna, Debabrata Datta, and S Suman Rajest. A review of medical image segmentation algorithms. *EAI Endorsed Transactions on Pervasive Health and Technology*, 7(27):e6–e6, 2021.

[5] Shoulin Yin and Jing Bi. Medical image annotation based on deep transfer learning. In *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pages 47–49. IEEE, 2018.

[6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[7] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[8] Cheng Chen, Juzheng Miao, Dufan Wu, Zhiling Yan, Sekeun Kim, Jiang Hu, Aoxiao Zhong, Zhengliang Liu, Lichao Sun, Xiang Li, Tianming Liu, Pheng-Ann Heng, and Quanzheng Li. Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation, 2023.

[9] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation, 2023.

[10] Haixing Dai, Chong Ma, Zhiling Yan, Zhengliang Liu, Enze Shi, Yiwei Li, Peng Shu, Xiaozheng Wei, Lin Zhao, Zihao Wu, Fang Zeng, Dajiang Zhu, Wei Liu, Quanzheng Li, Lichao Sun, Shu Zhang Tianming Liu, and Xiang Li. Samaug: Point prompt augmentation for segment anything model, 2024.

[11] Andreas S. Panayides, Amir Amini, Nenad D. Filipovic, Ashish Sharma, Sotirios A. Tsaftaris, Alistair Young, David Foran, Nhan Do, Spyretta Golemati, Tahsin Kurc, Kun Huang, Konstantina S. Nikita, Ben P. Veasey, Michalis Zervakis, Joel H. Saltz, and Constantinos S. Pattichis. Ai in medical imaging informatics: Current challenges and future directions. *IEEE Journal of Biomedical and Health Informatics*, 24(7):1837–1857, 2020.

[12] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint*, 2024.

[13] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[14] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4*, pages 311–320. Springer, 2019.

[15] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

[16] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images, 2022.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[18] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[20] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[22] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

[23] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023.

[24] Risab Biswas. Polyp-sam++: Can a text guided sam perform better for polyp segmentation? *arXiv preprint arXiv:2308.06623*, 2023.

[25] Qi Wu, Yuyao Zhang, and Marawan Elbatel. Self-prompting large vision models for few-shot medical image segmentation. In *MICCAI workshop on domain adaptation and representation transfer*, pages 156–167. Springer, 2023.

[26] Haixing Dai, Chong Ma, Zhiling Yan, Zhengliang Liu, Enze Shi, Yiwei Li, Peng Shu, Xiaozheng Wei, Lin Zhao, Zihao Wu, et al. Samaug: Point prompt augmentation for segment anything model. *arXiv preprint arXiv:2307.01187*, 2023.

[27] Dongjie Cheng, Ziyuan Qin, Zekun Jiang, Shaoting Zhang, Qicheng Lao, and Kang Li. Sam on medical images: A comprehensive study on three prompt modes. *arXiv preprint arXiv:2305.00035*, 2023.

[28] Cheng Chen, Juzheng Miao, Dufan Wu, Zhiling Yan, Sekeun Kim, Jiang Hu, Aoxiao Zhong, Zhengliang Liu, Lichao Sun, Xiang Li, et al. Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. *arXiv preprint arXiv:2309.08842*, 2023.

[29] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[30] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks, 2019.

[31] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation, 2021.

[32] Kwanyong Park, Sanghyun Woo, Seoung Wug Oh, In So Kweon, and Joon-Young Lee. Per-clip video object segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1342–1351, 2022.

[33] Yongqing Liang, Xin Li, Navid Jafari, and Qin Chen. Video object segmentation with adaptive feature bank and uncertain-region refinement, 2020.

[34] Jialu Li, Qingqing Zheng, Mingshuang Li, Ping Liu, Qiong Wang, Litao Sun, and Lei Zhu. Rethinking breast lesion segmentation in ultrasound: A new video dataset and a baseline network. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 391–400, Cham, 2022. Springer Nature Switzerland.

[35] Junhao Lin, Qian Dai, Lei Zhu, Huazhu Fu, Qiong Wang, Weibin Li, Wenhao Rao, Xiaoyang Huang, and Liansheng Wang. Shifting more attention to breast lesion segmentation in ultrasound videos, 2023.

[36] Cristina González, Laura Bravo-Sánchez, and Pablo Arbelaez. Isinet: An instance-based approach for surgical instrument segmentation, 2020.

[37] Zixu Zhao, Yueming Jin, Xiaojie Gao, Qi Dou, and Pheng-Ann Heng. Learning motion flows for semi-supervised instrument segmentation from robotic surgical video, 2020.

[38] Juana González-Bueno Puyal, Kanwal K. Bhatia, Patrick Brandao, Omer F. Ahmad, Daniel Toth, Rawen Kader, Laurence Lovat, Peter Mountney, and Danail Stoyanov. Endoscopic polyp segmentation using a hybrid 2d/3d cnn. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 295–305, Cham, 2020. Springer International Publishing.

[39] Bowen Wang, Liangzhi Li, Yuta Nakashima, Ryo Kawasaki, Hajime Nagahara, and Yasushi Yagi. Noisy-lstm: Improving temporal awareness for video semantic segmentation, 2020.

[40] Yifan Liu, Chunhua Shen, Changqian Yu, and Jingdong Wang. Efficient semantic video segmentation with per-frame inference, 2020.

[41] Samvit Jain, Xin Wang, and Joseph Gonzalez. Accel: A corrective fusion network for efficient semantic segmentation on video, 2019.

[42] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. Hiera: A hierarchical vision transformer without the bells-and-whistles, 2023.

[43] Jun Ma, Sumin Kim, Feifei Li, Mohammed Baharoon, Reza Asakereh, Hongwei Lyu, and Bo Wang. Segment anything in medical images and videos: Benchmark and deployment, 2024.

[44] Haoyu Dong, Hanxue Gu, Yaqian Chen, Jichen Yang, and Maciej A. Mazurowski. Segment anything model 2: an application to 2d and 3d medical images, 2024.

[45] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[46] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

[47] Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Shihao Ma, Adamo Young, Cheng Zhu, Kangkang Meng, Xin Yang, Ziyan Huang, et al. Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. *arXiv preprint arXiv:2308.05862*, 2023.

[48] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in Neural Information Processing Systems*, 35:36722–36732, 2022.

[49] Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, et al. 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426*, 2019.

[50] Jun Ma, Ronald Xie, Shamini Ayyadhury, Cheng Ge, Anubha Gupta, Ritu Gupta, Song Gu, Yao Zhang, Gihun Lee, Joonkee Kim, et al. The multi-modality cell segmentation challenge: Towards universal solutions. *arXiv preprint arXiv:2308.05864*, 2023.

[51] Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al. 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190*, 2020.

[52] David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P Langlotz, Paul A Heidenreich, Robert A Harrington, David H Liang, Euan A Ashley, et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020.

[53] David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P. Langlotz, Paul A. Heidenreich, Robert A. Harrington, David H. Liang, Euan A. Ashley, and James Y. Zou. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020.