Medical SAM 2: Segment Medical Images as Video via Segment Anything Model 2

Jiayuan Zhu
University of Oxford
jiayuan.zhu@ieee.org

Abdullah Hamdi
University of Oxford
abdullah.hamdi@eng.ox.ac.uk

Yunli Qi
University of Oxford
yunli.qi@wolfson.ox.ac.uk

Yueming Jin
National University of Singapore
ymjin@nus.edu.sg

Junde Wu
University of Oxford
jundewu@ieee.org

Abstract

Medical image segmentation plays a pivotal role in clinical diagnostics and treatment planning, yet existing models often face challenges in generalization and in handling both 2D and 3D data uniformly. In this paper, we introduce Medical SAM 2 (MedSAM-2), a generalized auto-tracking model for universal 2D and 3D medical image segmentation. The core concept is to leverage the Segment Anything Model 2 (SAM2) pipeline to treat all 2D and 3D medical segmentation tasks as a video object tracking problem. To put it into practice, we propose a novel self-sorting memory bank *mechanism that dynamically selects informative embeddings based on confidence and dissimilarity, regardless of temporal order. This mechanism not only significantly improves performance in 3D medical image segmentation but also unlocks a* One-Prompt Segmentation *capability for 2D images, allowing segmentation across multiple images from a single prompt without temporal relationships. We evaluated MedSAM-2 on five 2D tasks and nine 3D tasks, including white blood cells, optic cups, retinal vessels, mandibles, coronary arteries, kidney tumors, liver tumors, breast cancer, nasopharynx cancer, vestibular schwannoma, mediastinal lymph nodules, cerebral artery, inferior alveolar nerve, and abdominal organs, comparing it against state-of-the-art (SOTA) models in task-tailored, general and interactive segmentation settings. Our findings demonstrate that MedSAM-2 surpasses a wide range of existing models and updates new SOTA on several benchmarks. Our code has been released.*

1. Introduction

Artificial intelligence has significantly transformed various industries, and healthcare is poised for a substantial revolution driven by advancements in medical image understanding [10, 38, 64–66]. Medical image segmentation, which involves partitioning images into meaningful regions, is crucial for applications like diagnosis, treatment planning, and image-guided surgery [72–74]. Despite the progress made with automated segmentation methods using deep learning models such as convolutional neural networks (CNNs) and vision transformers (ViTs), significant challenges remain [4, 57]. One primary issue is model generalization; models trained on specific targets like certain organs or tissues often struggle to adapt to other targets or modalities. Additionally, many deep learning architectures are designed for 2D images, whereas medical imaging data frequently exist in 3D formats (e.g., CT, MRI), creating a gap when applying these models to 3D data [17, 50].

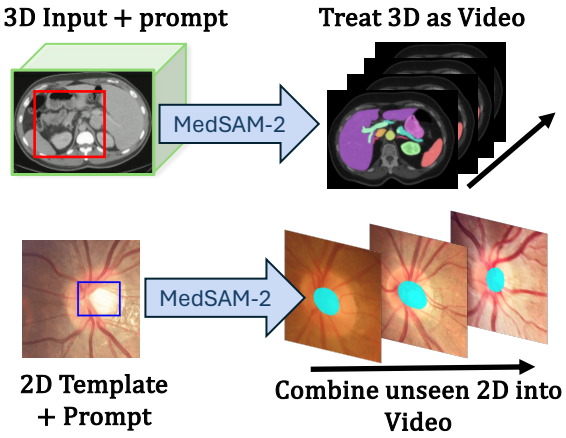

Figure 1. **Segmentation Capabilities of MedSAM-2**. When provided with a prompt in one 3D slice, MedSAM-2 can segment all later spatial-temporal 3D frames. When given a prompt in one 2D image, MedSAM-2 can accurately segment other 2D images that are not temporally related using the same criteria, which is an emergence of One-prompt Segmentation capability.

Recent developments in promptable segmentation models, particularly the Segment Anything Model (SAM) [38] and its enhanced version SAM 2 [55], have shown promise in addressing some challenges. SAM has demonstrated remark-

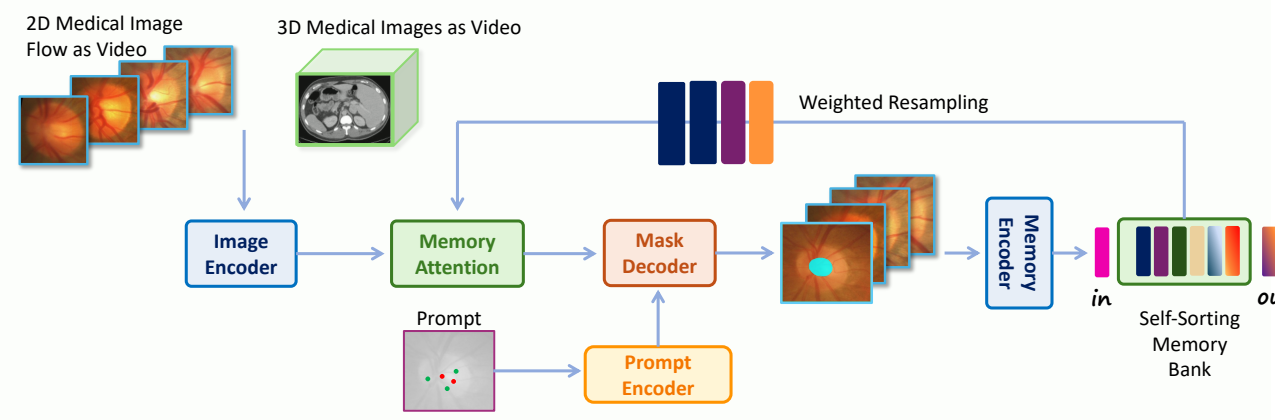

Figure 2. **MedSAM-2 Framework.** Building on the SAM 2 framework, we propose treating 3D medical images and 2D medical image flows as videos to facilitate memory-enhanced medical image segmentation. This approach not only improves performance in 3D medical image segmentation but also unlocks One-Prompt Segmentation capability for 2D medical image flows. This is achieved by incorporating our proposed Self-Sorting Memory Bank, which selects the most confident embeddings based on the confidence predictions (α, β, γ) from the mask decoder.

able zero-shot capabilities in image segmentation tasks by leveraging user-provided prompts to segment objects without prior training on specific targets. However, this approach requires user interaction for each image, which can be labor-intensive and impractical in clinical settings where large volumes of data are common [48]. SAM 2 extends SAM's capabilities to videos, introducing real-time object tracking with reduced user interaction time. Yet, it still relies on temporal relationships between frames, limiting its applicability to unordered medical images and failing to fully address the generalization challenges in medical image segmentation.

In this work, we introduce MedSAM-2, a generalized auto-tracking model for universal medical image segmentation. MedSAM-2 tackles these challenges by treating medical images as videos and incorporating a novel **self-sorting memory bank**. This mechanism dynamically selects informative embeddings based on confidence and dissimilarity, allowing the model to handle unordered medical images effectively. By rethinking the memory mechanism in SAM 2, MedSAM-2 not only improves performance in 3D medical image segmentation but also unlocks the *One-Prompt Segmentation* capability [70] for 2D medical images. This capability enables the model to generalize from a single prompt to segment across multiple images without temporal relationships, significantly reducing user interaction and enhancing convenience for clinicians.

We evaluated MedSAM-2 across 14 different benchmarks, encompassing 25 distinct tasks for validation. Compared with previous fully-supervised segmentation models and SAM-based interactive models, MedSAM-2 demonstrated superior performance across all tested methods and achieved state-of-the-art results in both 2D and 3D medical image segmentation tasks. Specifically, under the one-prompt segmentation setting, MedSAM-2 outperformed previous foundation segmentation models, thereby showcasing its exceptional generalization capabilities. Our contributions can be summarized as follows:

Contributions **(i)** We introduce MedSAM-2, the first SAM-2-based generalized auto-tracking model for universal medical image segmentation, capable of uniformly handling both 2D and 3D medical imaging tasks with minimal user intervention. **(ii)** We propose a novel *self-sorting memory bank* mechanism that dynamically selects informative embeddings based on confidence and dissimilarity, enhancing the model's ability to handle unordered medical images and improves generalization. **(iii)** We evaluate MedSAM-2 across 15 different benchmarks, including 25 distinct tasks, demonstrating superior performance compared to previous fully-supervised and SAM-based interactive models.

2. Related Works

Medical Image Segmentation Traditionally, medical image segmentation models have been task-specific, designed and optimized for particular targets like specific organs or tissues [12, 14]. These task-tailored models leverage the unique characteristics of each task to achieve high performance. For instance, uncertain-aware modules have been utilized to handle the ambiguity in optic cup segmentation in fundus images [35] However, the reliance on task-specific models presents significant challenges. Designing and training a unique model for each segmentation task is labor-intensive and time-consuming. Moreover, many deep learning architectures are designed for 2D images, whereas medical imaging data often exist in 3D formats (e.g., CT, MRI), creating a gap when applying these models to 3D data [17, 24–26, 50]. To address these limitations, there has been growing interest in developing generalized medical image segmentation

models capable of handling multiple tasks and modalities [28, 72, 74]. These models aim to generalize across different targets without the need for task-specific adaptations. However, achieving robust generalization remains a significant challenge due to the diverse nature of medical images, which can vary greatly in appearance, resolution, and anatomical structures. On the other hand, our MedSAM-2 is a generalized model that tackles multiple domains and can be used for both 2D and 3D medical image segmentation.

Prompting Segment Anything Models (SAMs) The introduction of the Segment Anything Model (SAM) [38] marked a significant advancement in the field of image segmentation. SAM leverages user-provided prompts to segment objects in images without prior training on specific targets, demonstrating remarkable zero-shot capabilities. In medical imaging, early applications of SAM involved fine-tuning the model to adapt to different medical segmentation tasks [15, 18, 58, 73]. However, these approaches still required user interaction for each image, which can be impractical in clinical settings with large volumes of data. To reduce the reliance on extensive user prompting, researchers have explored few-shot and zero-shot segmentation methods [20, 44, 52, 71], enabling adaptation to new tasks with minimal annotated samples. For example, UniSeg [11] requires only a few annotated samples to process an entire unseen segmentation task during inference. One-Prompt Segmentation [70] further combines SAM's interactive setting with zero-shot segmentation, requiring only one visual prompt for a template sample to segment similar targets in subsequent samples without retraining. Nonetheless, these models still primarily focus on 2D medical images and are not yet tailored for the unique requirements of 3D medical imaging. Our MedSAM-2 uses a self-sorting memory bank allowing the model to generalize better on unsorted medical images while leveraging the context-rich video pretraining of SAM 2 with minimal prompting requirements.

3. Method

We introduce MedSAM-2, an advanced segmentation model based on SAM 2 [55], tailored for medical image segmentation tasks.

3.1. Preliminaries on Segment Anything Model (SAM 2)

SAM 2 [55] is a promptable visual segmentation model designed for image and video tasks. Given an input sequence of frames or images $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{T}$ and optional prompts $\mathbf{P} = \{\mathbf{P}_t\}_{t=1}^{T}$, the model predicts segmentation masks $\mathbf{Y} = \{\mathbf{y}_t\}_{t=1}^{T}$ for each frame $\mathbf{x}_t$. The architecture comprises an image encoder $\mathcal{E}_{\text{img}}$ that encodes each frame $\mathbf{x}_t$ into a feature embedding $\mathbf{F}_t = \mathcal{E}_{\text{img}}(\mathbf{x}_t)$; a prompt encoder $\mathcal{E}_{\text{prompt}}$ that processes user prompts $\mathbf{P}_t$, generating embeddings $\mathbf{Q}_t = \mathcal{E}_{\text{prompt}}(\mathbf{P}_t)$; a memory bank $\mathcal{M}_t$ that stores K past embeddings $\mathbf{E}_i$ before frame $\mathbf{x}_t$; a memory attention mechanism $\mathcal{A}$ that combines $\mathbf{F}_t$, $\mathcal{M}_t$, and $\mathbf{Q}_t$; and a mask decoder $\mathcal{D}$ that predicts the segmentation mask $\mathbf{y}_t$. Mathematically, the segmentation process can be formulated:

$$\begin{aligned} \mathbf{y}_t &= \mathcal{D}\left(\mathcal{A}\left(\mathbf{F}_t, \mathcal{M}_t, \mathbf{Q}_t\right)\right), \quad \text{for } t = 1, \ldots, T, \\ \mathcal{M}_t &= \left\{ \mathbf{E}_i \mid i \in \{\max(j, 0)\}_{j=t-K-1}^{t-1} \right\}, \end{aligned} \tag{1}$$

3.2. MedSAM-2: Self-Sorting SAM2 for Medical Imaging

Although SAM2 has been highly successful with natural images, directly applying it to medical images is not straightforward. In medical imaging, the order of slices or images may not be meaningful due to varying acquisition protocols and orientations. Moreover, 2D medical images are often unordered, and each orientation in 3D imaging can be considered as an independent sequence to be integrated with different order. To address this, we propose a **self-sorting memory bank** $\mathcal{M}_t^{\text{sort}}$ that dynamically selects and retains the most informative embeddings, rather than simply using the most recent K frames as in SAM 2 [55].

Memory Bank Update with Confidence and Dissimilarity At each time step t, we update the self-sorting memory bank $\mathcal{M}_t^{\text{sort}}$ based on $\mathcal{M}_{t-1}^{\text{sort}}$ and the embedding $\mathbf{E}_{t-1}$ of the previous frame. First, the model predicts the segmentation mask $\mathbf{y}_{t-1}$ and computes IOU confidence score c_{t-1} for frame $t-1$, estimated by the model itself. If the confidence score satisfies $c_{t-1} \geq c_{\text{thresh}}$, we consider adding $\mathbf{E}_{t-1}$ to the memory bank. We form a candidate set $\mathcal{C} = \mathcal{M}_{t-1}^{\text{sort}} \cup \{\mathbf{E}_{t-1}\}$. To maintain diversity, we compute the total dissimilarity for each embedding in $\mathcal{C}$:

$$D_i = \sum_{\substack{\mathbf{E}_j \in \mathcal{C} \\ j \neq i}} 1 - \text{sim}(\mathbf{E}_i, \mathbf{E}_j), \quad \forall \mathbf{E}_i \in \mathcal{C} = \mathcal{M}_{t-1}^{\text{sort}} \cup \{\mathbf{E}_{t-1}\}, \tag{2}$$

where $\text{sim}(\cdot, \cdot)$ is a similarity function (e.g., cosine similarity). We then select the top K embeddings with the highest total dissimilarity to form the updated memory bank:

$$\mathcal{M}_t^{\text{sort}} = \underset{\mathbf{E}_i \in \mathcal{C}}{\text{TopK}}(D_i). \tag{3}$$

If the confidence condition is not met ($c_{t-1} < c_{\text{thresh}}$), we keep the memory bank unchanged $\mathcal{M}_t^{\text{sort}} = \mathcal{M}_{t-1}^{\text{sort}}$.

Resampling the Memory Bank Before computing the attention for frame t, we resample the memory bank to emphasize embeddings similar to the current embedding $\mathbf{F}_t$, enhancing relevance. This is achieved by assigning higher selection probabilities to embeddings more similar to $\mathbf{F}_t$. We calculate the similarity scores between $\mathbf{F}_t$ and each embedding $\mathbf{E}_i$ in the memory bank $\mathcal{M}_t^{\text{sort}}$ using a similarity function $\text{sim}(\cdot, \cdot)$ (e.g., cosine similarity):

$$p_{i,t} = \frac{\text{sim}(\mathbf{F}_t, \mathbf{E}_i)}{\sum_{\mathbf{E}_j \in \mathcal{M}_t^{\text{sort}}} \text{sim}(\mathbf{F}_t, \mathbf{E}_j)}, \quad \forall \mathbf{E}_i \in \mathcal{M}_t^{\text{sort}}. \tag{4}$$

Using the probability distribution $\{p_{i,t}\}$, we perform resampling with replacement to create the importance-weighted memory bank $\tilde{\mathcal{M}}_t^{\text{sort}}$. Specifically, we sample K embeddings from $\mathcal{M}_t^{\text{sort}}$, where each embedding $\mathbf{E}_i$ is selected independently with probability $p_{i,t}$:

$$\tilde{\mathcal{M}}_t^{\text{sort}} = \{\mathbf{E}_{i_k} \mid i_k \sim \text{Categorical}(\{p_{i,t}\}),\ k = 1, \ldots, K\}. \tag{5}$$

This resampling process effectively prioritizes embeddings more similar to current embedding $\mathbf{F}_t$, enhancing the relevance of the memory bank in the attention mechanism.

MedSAM-2 Pipeline The segmentation process in MedSAM-2 incorporates the self-sorting memory bank and resampled embeddings into SAM 2. With the fixed prompt $\mathbf{P}_1$ from the first frame, we modify (1) as:

$$\mathbf{y}_t = \mathcal{D}\left(\mathcal{A}\left(\mathbf{F}_t, \tilde{\mathcal{M}}_t^{\text{sort}}, \mathbf{Q}_1\right)\right), \quad \text{for } t = 1, \ldots, T, \tag{6}$$

where $\mathbf{F}_t = \mathcal{E}_{\text{img}}(\mathbf{x}_t)$, $\mathbf{Q}_1 = \mathcal{E}_{\text{prompt}}(\mathbf{P}_1)$, and $\tilde{\mathcal{M}}_t^{\text{sort}}$ is the resampled memory bank from 5. This modification allows MedSAM-2 to handle unordered medical images effectively, leveraging informative and relevant embeddings for segmentation, thus enhancing performance in both 2D and 3D medical imaging tasks after training with standard segmentation loss [48].

Self-Sorting Works because of Entropy and Mutual Information By utilizing the self-sorting memory bank, we ensure that the memory bank contains the most reliable and informative embeddings, regardless of their temporal order. This self-sorting mechanism not only handles the unordered nature of medical images but also forces the extracted features to have higher entropy due to the increased randomness introduced by the "learned shuffle“ based on confidence rather than inherent temporal order. This increased entropy coincides with an increase in the mutual information between the memory bank features and the output, improving the robustness and generalization of the model according to principle of maximum entropy [34]. Consequently, the model is better equipped to handle unordered medical images. We show mathematically in how entropy-increase and mutual information by self-sorting improve learning generalization.

3.3. Unified Approach for 2D and 3D Images

MedSAM-2 leverages a self-sorting memory bank to improve robustness and effectively utilize context in both 2D and 3D medical imaging segmentation tasks. This unified framework allows MedSAM-2 to perform effectively across diverse medical imaging scenarios, unlocking the 'One-Prompt Segmentation' capability for 2D medical images and improving performance in 3D segmentation.

MedSAM-2 for 3D Medical Imaging For 3D medical images, such as MRI or CT scans represented as volumes $\mathbf{V} \in \mathbb{R}^{H \times W \times D}$, we treat the volume as a sequence of 2D slices along various orientations, similar to frames in a video. We define six orientations for processing the 3D volume:

1. **Axial**: $\mathbf{X}^{(1)} = \{\mathbf{x}_t = \mathbf{V}(:, :, t)\}_{t=1}^{D}$.
2. **Coronal**: $\mathbf{X}^{(2)} = \{\mathbf{x}_t = \mathbf{V}(:, t, :)\}_{t=1}^{H}$.
3. **Sagittal**: $\mathbf{X}^{(3)} = \{\mathbf{x}_t = \mathbf{V}(t, :, :)\}_{t=1}^{W}$.
4. **Reverse Axial**: $\mathbf{X}^{(4)} = \{\mathbf{x}_t = \mathbf{V}(:, :, D - t + 1)\}_{t=1}^{D}$.
5. **Reverse Coronal**: $\mathbf{X}^{(5)} = \{\mathbf{x}_t = \mathbf{V}(:, H - t + 1, :)\}_{t=1}^{H}$.
6. **Reverse Sagittal**: $\mathbf{X}^{(6)} = \{\mathbf{x}_t = \mathbf{V}(W - t + 1, :, :)\}_{t=1}^{W}$.

Processing the volume with all orientations combined $\mathbf{X} = \bigcup_{o=1}^{6} \mathbf{X}^{(o)}$ exposes the model to diverse anatomical perspectives, enhancing its ability to generalize and capture anisotropic structures. However, the best *order* of picking these directions is still *unknown*. Hence, our self-sorting memory bank order embeddings from different orientations based on the mean direction features and their confidences as before. This allows our MedSAM-2 model to *jointly* capture the 3D context and reap the benefits of the self-sorting mechanism.

During inference, the model processes the input data in multiple orientations with combined input $\mathbf{X}$, obtaining segmentation predictions where the final output for the 3D volume is obtained by aggregating these predictions: $\mathbf{Y}_{\text{3D}} = \text{Aggregate}\left(\{\mathbf{Y}^{(o)}\}_{o=1}^{6}\right)$, where Aggregate is a function such as averaging or majority voting applied pixel-wise.

MedSAM-2 for One-prompt 2D Segmentation For 2D medical images, which may consist of independent slices or images lacking temporal connections, we treat sets of images as pseudo-video sequences. By processing them sequentially using MedSAM-2's memory mechanism, we achieve a *One-Prompt Segmentation* capability [70], where providing a prompt on a single image template $(\mathbf{X}_1, \mathbf{P}_1)$ allows the model to propagate the segmentation across the entire set. Our MedSAM-2 approach leverages the self-sorting memory bank to associate the prompt more closely with intrinsic features in each frame, improving generalization and efficiency. This ability of one-prompt-segmentation is less restrictive than the universal interactive video object segmentation (iVOS) [48], where its target is to learn a universal function for any single input image $\mathbf{x}_i$ and prompt $\mathbf{P}_i$, to predict the output mask $\mathbf{y}_i$.

4. Experiment

4.1. Dataset

To build a foundation model with strong generalization on unseen tasks, we train and test our model on the One-Prompt dataset [70], a large-scale and diverse collection of 2D and 3D medical images assembled from publicly accessible datasets with clinicians-annotated prompts. This data source comprises 78 datasets across various medical domains and imaging modalities, covering a wide range of organs such as lungs [59–61], eyes [21, 32, 49, 51], brain [7, 23, 31, 39, 40], and abdominal organs [8, 29, 36, 37, 41–43, 45–47, 54, 61].

Table 1. **3D Medical Images Segmentation Performance**. We show the comparison of MedSAM-2 with SOTA segmentation methods over BTCV dataset [22] evaluated by Dice Score (%). Task-tailored models, interactive generalized models, auto-tracking generalized models are marked in yellow, green, blue.

Model	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Stom.	Aorta	IVC	Veins	Panc.	AG	Ave
TransUNet	0.952	0.927	0.929	0.662	0.757	0.969	0.889	0.920	0.833	0.791	0.775	0.637	0.838
UNetr	0.968	0.924	0.941	0.750	0.766	0.971	0.913	0.890	0.847	0.788	0.767	0.741	0.856
Swin-UNetr	0.971	0.936	0.943	0.794	0.773	0.975	0.921	0.892	0.853	0.812	0.794	0.765	0.869
nnUNet	0.942	0.894	0.910	0.704	0.723	0.948	0.824	0.877	0.782	0.720	0.680	0.616	0.802
EnsDiff	0.938	0.931	0.924	0.772	0.771	0.967	0.910	0.869	0.851	0.802	0.771	0.745	0.854
SegDiff	0.954	0.932	0.926	0.738	0.763	0.953	0.927	0.846	0.833	0.796	0.782	0.723	0.847
MedSegDiff	0.973	0.930	0.955	0.812	0.815	0.973	0.924	0.907	0.868	0.825	0.788	0.779	0.879
SAM	0.518	0.686	0.791	0.543	0.584	0.461	0.562	0.612	0.402	0.553	0.511	0.354	0.548
MedSAM	0.751	0.814	0.885	0.766	0.721	0.901	0.855	0.872	0.746	0.771	0.760	0.705	0.803
SAM-U	0.868	0.776	0.834	0.690	0.710	0.922	0.805	0.863	0.844	0.782	0.611	0.780	0.790
SAM-Med3D	0.873	0.884	0.932	0.795	0.790	0.943	0.889	0.872	0.796	0.813	0.779	0.797	0.847
SAMed	0.862	0.710	0.798	0.677	0.735	0.944	0.766	0.874	0.798	0.775	0.579	0.790	0.776
VMN	0.803	0.788	0.801	0.783	0.712	0.870	0.821	0.832	0.825	0.742	0.655	0.710	0.779
FCFI	0.876	0.834	0.889	0.795	0.781	0.945	0.887	0.921	0.897	0.829	0.780	0.760	0.858
SAM 2	0.861	0.882	0.913	0.864	0.832	0.861	0.891	0.835	0.822	0.855	0.831	0.871	0.860
TrackAny	0.818	0.762	0.760	0.805	0.730	0.824	0.841	0.829	0.815	0.780	0.701	0.728	0.783
iMOS	0.801	0.738	0.759	0.844	0.762	0.855	0.861	0.843	0.828	0.810	0.744	0.672	0.793
UniverSeg	0.824	0.862	0.889	0.774	0.835	0.918	0.826	0.899	0.831	0.804	0.819	0.818	0.842
OnePrompt	0.881	0.869	0.894	0.900	0.868	0.891	0.908	0.837	0.829	0.850	0.832	0.870	0.869
MedSAM-2	0.922	0.931	0.919	0.932	0.918	0.865	0.871	0.896	0.834	0.847	0.840	0.911	**0.890**

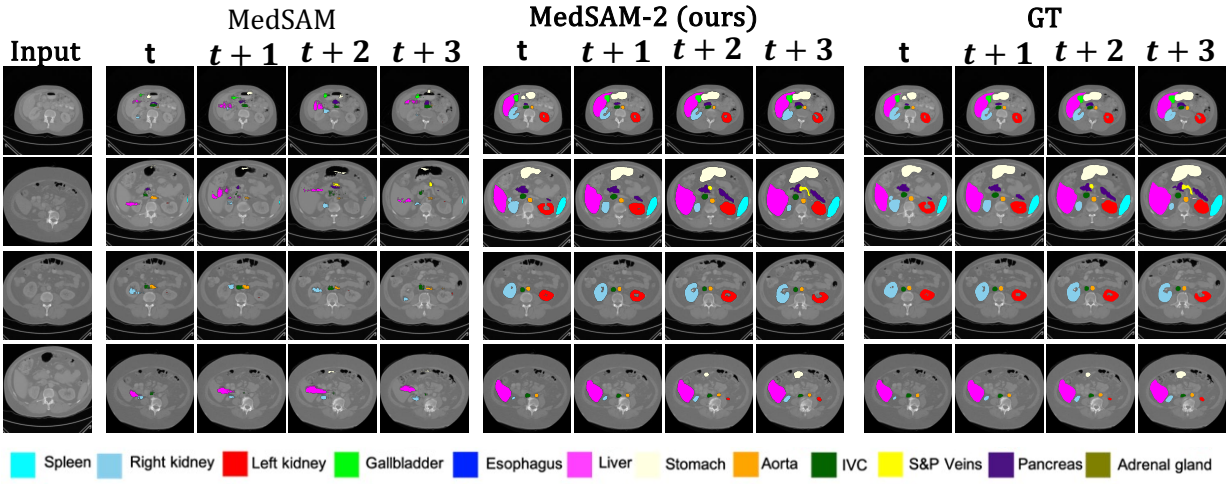

Figure 3. **Qualitative Comparison on 3D Medical Image Segmentation.** We show comparison of MedSAM [48], our MedSAM-2, and ground truth on sequential 3D medical image segmentation on the BTCV dataset [22]. Note how our MedSAM-2 produce more consistent 3D predictions leveraging the 3D context and maintaining high generalization capability compared to MedSAM [48].

Each dataset includes at least one image or slice annotated by a professional clinician, with over 3,000 samples collectively annotated by clinicians. A detailed list of the One-Prompt datasets is provided in the supplementary materials.

We follow the default split of the One-Prompt dataset, using 64 datasets for training and 14 for testing. The test set includes 8 MICCAI2023 Challenge tasks across diverse anatomies—kidney tumor [30], liver tumor [53], breast cancer [1], nasopharynx cancer [6], vestibular schwannoma [2], mediastinal lymph node [56], cerebral artery [13], and inferior alveolar nerve [9]—along with 6 other tasks for structures like white blood cells [79], optic cups [21], mandibles [3], coronary arteries [63], abdominal organs [22], and retinal vessels [32]. We evaluate model performance on each test dataset using task-specific prompts: the *Click* prompt for KiTS23, ATLAS23, TDSC, and WBC; the *BBox* prompt for SegRap, CrossM23, REFUGE, Pendal, LNQ23, and CAS23; and the *Mask* prompt for CadVidSet, STAR, BTCV-test and

Table 2. **Universal 2D Medical Images Segmentation Performance.** We show the comparison of MedSAM-2 with task-tailored models, interactive generalized models, and auto-tracking generalized models. Evaluated on 11 unseen tasks by Dice Score (%).

Methods	KiTS	ATLAS	WBC	SegRap	CrossM	REFUGE	Pendal	LQN	CAS	CadVidSet	ToothFairy	Ave
TransUNet	38.2	34.5	49.1	25.5	37.7	36.3	31.2	23.3	24.5	31.6	37.9	33.6
Swin-UNetr	37.2	26.5	32.1	25.6	29.7	28.9	31.4	17.2	20.5	22.6	32.1	28.5
nnUNet	39.8	30.3	40.4	26.8	35.0	34.9	42.9	18.9	37.4	41.8	35.3	34.9
MedSegDiff	40.1	30.5	42.9	34.7	37.7	31.9	42.6	21.1	38.3	34.7	33.5	35.3
MSA	54.6	48.9	55.9	47.3	51.7	49.2	54.2	41.0	48.9	53.5	47.6	50.3
MedSAM	62.4	53.1	67.8	52.3	59.3	54.5	58.7	42.5	41.5	45.7	56.2	53.9
SAM-Med2D	56.3	51.4	52.6	43.5	47.2	52.0	50.8	47.4	44.3	49.0	55.1	50.0
SAM2	64.6	58.3	69.1	54.8	60.7	55.8	61.4	45.1	51.6	53.9	59.0	57.6
TrackAny	63.1	56.2	66.6	51.3	57.8	54.5	60.1	43.6	42.5	41.4	54.0	53.7
iMOS	62.6	53.8	61.4	52.5	54.3	57.0	58.7	42.2	44.8	46.1	50.7	53.1
UniverSeg	63.8	54.2	74.0	70.8	74.2	71.1	69.2	47.7	60.4	66.8	65.1	65.2
One-Prompt	75.3	66.8	77.5	81.2	83.8	77.4	72.8	51.9	61.6	79.3	76.4	73.1
MedSAM-2	**78.2**	**71.8**	**80.5**	**86.2**	**85.7**	**79.9**	**76.0**	**53.6**	**66.5**	**86.1**	**80.8**	**76.8**

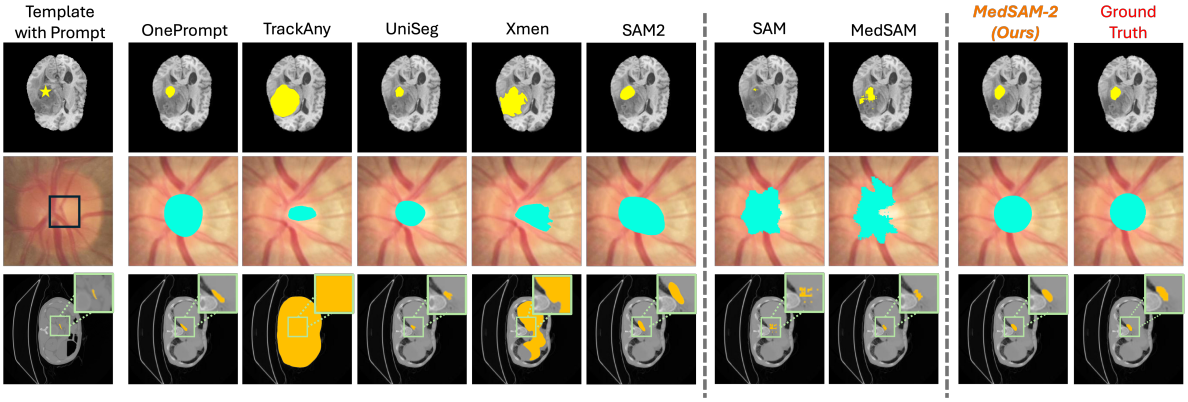

Figure 4. **Qualitative Examples of MedSAM-2 for 2D One-Prompt Segmentation & 3D Segmentation.** We show several examples of 2D segmentation on diverse datasets.

ToothFairy. Among these, STAR, BTCV-test, and TDSC involve tasks seen in training, while the remaining 11 tasks are used for zero-shot testing.

4.2. Human-User Prompted Evaluation

For evaluation, we engaged human users to simulate real-world interactions in prompt-based segmentation. Fifteen users were assigned to prompt approximately 10% of the test images, including 5 laypersons with a clear understanding of the task but no clinical background, 7 junior clinicians, and 3 senior clinicians. This setup aims to reflect real-world prompting scenarios, such as clinical training or semi-automated annotation.

4.3. Implementation

We conduct training and testing on the PyTorch platform, leveraging 64 NVIDIA A100 GPUs for distributed training. Optimization uses the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with a linear learning rate warmup followed by cosine decay. Our training simulates an interactive environment by sampling 8-frame sequences, randomly selecting up to 2 frames (including the first) for corrective clicks. Prompts are generated from ground-truth masks and model predictions, with initial prompts consisting of the ground-truth mask with 50% probability, a positive click from the mask with 25%, or a bounding box input with 25%. To maintain diversity across tasks and prompts, we use a balanced sampling strategy that avoids equal representation across all tasks, as certain image modalities, tasks, or prompt types are more frequent. To prevent overfitting to these dominant elements, we uniformly select tasks and sequence states, starting with a random task selection, then narrowing the pool to data associated with that task. We proceed by selecting an image modality available for the task, refining the pool to ensure homogeneity, and finally selecting a sample from the filtered set. All comparison models are trained and tested under the same setting. Additional details on data processing and training are provided in the supplementary material.

5. Results

In this section, we present a comprehensive evaluation of MedSAM-2 on both 2D and 3D medical image segmentation tasks. We compare our model with a range of state-of-the-art (SOTA) methods, including task-specific, diffusion-based, and interactive segmentation models. Performance is quan-

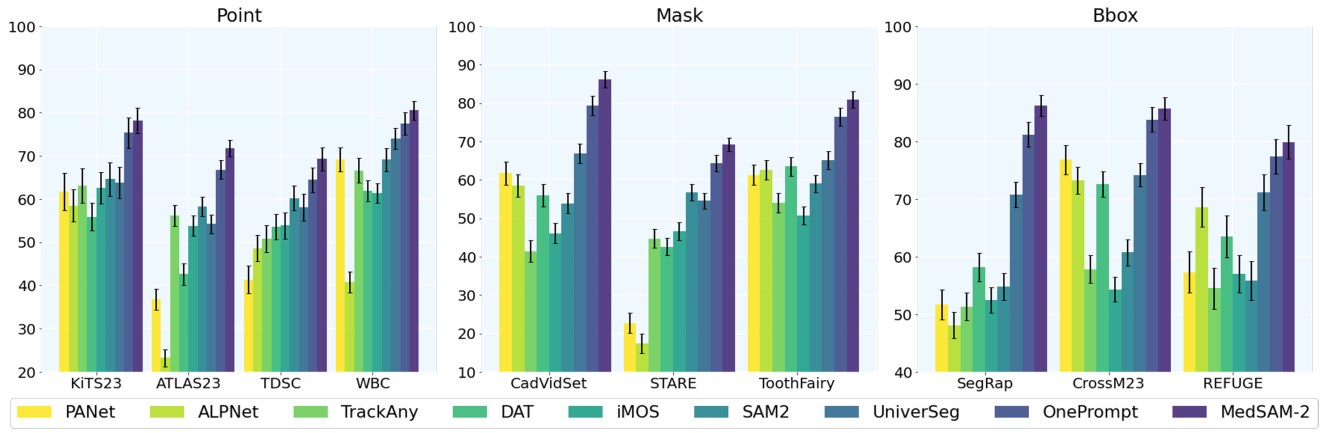

Figure 5. **One-prompt 2D Segmentation Performance.** We show MedSAM-2 *v.s.* Few/One-shot Models under One-prompt Segmentation setting on 10 datasets with different prompts. Our MedSAM-2 colored by the darkest blue on the right of each bar group.

tified using the Dice coefficient, Intersection over Union (IoU), and Hausdorff Distance (HD95) where appropriate.

5.1. Performance of Universal Medical Image Segmentation

For 3D medical images, prompts are provided to frames with a probability of 0.25, meaning each frame has a 25% likelihood of receiving a prompt. For 2D images, prompts are provided with a probability of 0.3. Results for 3D medical image segmentation are presented in Table 1, while 2D results are presented in Table 2.

On 3D Medical Images To assess the general performance of MedSAM-2 on 3D medical images, we conducted experiments on the BTCV multi-organ segmentation dataset (3). We compare MedSAM-2 with established SOTA segmentation methods such as nnUNet [33], TransUNet [14], UNetr [28], Swin-UNetr [27], and diffusion-based models like EnsDiff [68], SegDiff [5], and MedSegDiff [72]. Additionally, we evaluate interactive segmentation models including SAM [38], MedSAM [48], SAMed [77], SAM-Med2D [16], SAM-U [19], VMN [80], and FCFI [67]. For FCFI, ConvNext-v2-H [69] is used as the backbone. We also compare MedSAM-2 with auto-tracking generalized models, such as SAM 2 [55], TrackAny [76], iMOS [75], UniverSeg [11], OnePrompt [70]. Table 1 presents the quantitative results on the BTCV dataset. MedSAM-2 achieves a Dice score of 89.0%, outperforming all compared methods. Specifically, MedSAM-2 surpasses the previous SOTA model MedSegDiff by a margin of 1.10%. Among interactive models, MedSAM-2 maintains the lead, outperforming the previously leading interactive model, FCFI, by 3.20%. It is important to note that all these competing interactive models require prompts for each frame, whereas MedSAM-2 achieves better results with far fewer user prompts.

On 2D Medical Images We further evaluate MedSAM-2 in a zero-shot setting on 11 unseen 2D medical image segmentation tasks. Similar to 3D medical image segmentation

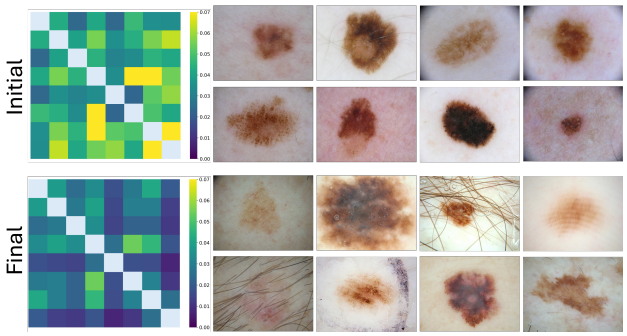

Figure 6. **Mutual Information Analysis of the Self-Sorting Memory Bank.** We show pairwise mutual information analysis and visualizations of memory bank samples at the start and final stages. The total mutual information between the memory bank samples decreases from 2.54 in the initial stage to 1.43 in the final stage.

ask, we compare the results with task-tailored models, interactive models that require prompts for each image, and auto-tracking models. Table 2 summarizes the results across different datasets. MedSAM-2 consistently outperforms the compared methods, demonstrating its superior generalization capability across diverse medical imaging modalities. For instance, MedSAM-2 improves the Dice score by 2.5% on optic disc segmentation and 2.9% on brain tumor segmentation compared to the previous best models. Even when compared to interactive models that require prompts for each image, MedSAM-2 maintains its lead, highlighting the effectiveness of our proposed self-sorting memory bank mechanism. See 4 for visualizations.

5.2. One-prompt Segmentation Performance under different prompts

We further assess MedSAM-2 under the One-Prompt segmentation setting by comparing it to various few/one-shot learning baselines that use different prompts. We benchmark against few/one-shot models such as PANet [62], ALPNet [52],TrackAny [76], DAT [78], iMOS [75], turned

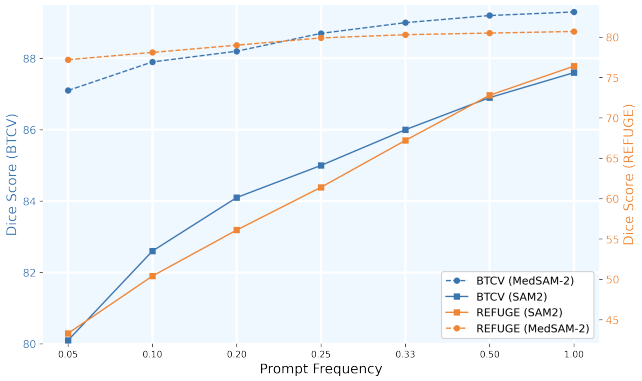

Figure 7. **Impact of Prompt Frequency on 2D and 3D Medical Images.** We report the Dice Score (%) of MedSAM-2 on REFUGE (2D) and BTCV (3D) datasets with varying prompt frequencies.

SAM2 [55], UniverSeg [11], and One-Prompt [70]. We further evaluate the models by testing them 5 times with different prompted images and input sequences to observe performance variance. Figure 5 presents the average Dice scores and variance per task for each method. MedSAM-2 not only consistently achieves higher average performance but also demonstrates significantly lower variance in most cases, underscoring its robust generalization across various tasks and prompt types.

5.3. Analysis and Ablation Study

Mutual Information Analysis of Memory Bank We analyze the effectiveness of the self-sorting memory bank in MedSAM-2 by examining the mutual information of stored embeddings over time using the ISIC dataset. The mutual information analysis assesses the diversity of embeddings in the memory bank. Figure 6 illustrates the pairwise mutual information of memory bank samples at different stages. Initially, the total mutual information is high (2.54), indicating redundancy among the stored embeddings. As the memory bank evolves, the mutual information decreases to 1.43, showing that the embeddings become more diverse and representative of different features. This confirms that the self-sorting mechanism effectively captures a diverse set of informative embeddings, enhancing the model's generalization capability. The right side of Figure 6 visually supports this trend, showing an increasingly diverse set of samples as the memory bank develops from the initial to the final stage.

Prompt Frequency Analysis on 2D and 3D Medical Images We conduct experiments to study the impact of prompt frequency on the performance of 2D (REFUGE) and 3D (BTCV) datasets. The performance improves with increasing prompt given frequency, as seen in the progressive increase in Dice scores for both datasets (Figure 7). Compared to SAM 2, our model demonstrates greater robustness under varying prompt frequencies. On 3D images, the performance gap between 5% prompting and full prompting is only 2% for our model, while SAM 2 shows a 7.5% gap. This difference is even more pronounced in 2D medical images, where our model maintains a 3.5% gap, whereas SAM 2 shows a substantial 33.1% drop. This highlights how our self-sorting memory bank significantly enhances model robustness, achieving strong performance even with minimal human interaction.

Table 3. **Ablation Study of MedSAM-2.** We evaluate each of the novel components of MedSAM-2 pipeline on the CadVidSet dataset [63] and the aorta task in BTCV dataset [22].

IOU Confidence Threshold	Dissimilar Templates	Memory Resampling	CadVidSet	BTCV-Aorta
			53.9	83.5
✓			57.8	86.2
✓	✓		64.5	88.4
✓	✓	✓	**72.9**	**89.6**

Ablation Study In the ablation study, we evaluate several key design choices for the MedSAM-2 model, including the use of an IOU confidence threshold c_{thresh} from Section 3.2 for storing samples, the storage of dissimilar templates in the memory bank, and the application of resampling strategies on the memory bank. This study is conducted with with CadVidSet dataset and the aorta task in the BTCV dataset. Table 3 presents the results of the ablation study. Using an IOU confidence threshold for the memory bank improves CadVidSet dataset's average Dice score to 57.8%, whereas without the threshold, it reaches only 53.9%. This selective storage based on confidence enhances the quality of retained samples and reduces redundancy. In addition, further storing dissimilarity templates results in a Dice score of 64.5% and 88.4% for CadVidSet and BTCV dataset, respectively. The dissimilarity-based storage captures a broader range of features, allowing the memory bank to adapt more effectively to diverse inputs. Applying the resampling conditioned on feature relevance raises the Dice score to 72.9% and 89.6% for CadVidSet and BTCV datasets, by prioritizing relevant samples for specific segmentation tasks.

6. Conclusion

In this work, we introduced MedSAM-2, an generalized auto-tracking segmentation model for both 2D and 3D medical images. By treating medical images as videos and incorporating a novel *self-sorting memory bank*, MedSAM-2 effectively handles unordered medical images and enhances generalization across diverse tasks. This innovation unlocks the *One-Prompt Segmentation* capability, allowing MedSAM-2 to generalize from a single prompt to segment similar structures across multiple images without temporal relationships. Comprehensive evaluations across 14 benchmarks and 25 tasks demonstrate that MedSAM-2 consistently outperforms state-of-the-art models in both 2D and 3D medical image segmentation. It achieves superior performance while re-

ducing the need for continuous user interaction, making it particularly advantageous in clinical settings.

References

[1] Tumor detection, segmentation and classification challenge on automated 3d breast ultrasound (abus) 2023. *https://tdsc-abus2023.grand-challenge.org*, 20203. 5

[2] Cross-modality domain adaptation for medical image segmentation. *https://crossmoda-challenge.ml*, 2023. 5

[3] Amir Hossein Abdi, Shohreh Kasaei, and Mojdeh Mehdizadeh. Automatic segmentation of mandible in panoramic x-ray. *Journal of Medical Imaging*, 2(4):044003–044003, 2015. 5

[4] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020. 1

[5] Tomer Amit, Eliya Nachmani, Tal Shaharbany, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021. 7

[6] Mehdi Astaraki, Simone Bendazzoli, and Iuliana Toma-Dasu. Fully automatic segmentation of gross target volume and organs-at-risk for radiotherapy planning of nasopharyngeal carcinoma. *arXiv preprint arXiv:2310.02972*, 2023. 5

[7] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021. 4

[8] Nicholas Bloch, Anant Madabhushi, Henkjan Huisman, John Freymann, Justin Kirby, Michael Grauer, Andinet Enquobahrie, Carl Jaffe, Larry Clarke, and Keyvan Farahani. Nci-isbi 2013 challenge: automated segmentation of prostate structures. *The Cancer Imaging Archive*, 370(6):5, 2015. 4

[9] Federico Bolelli. Tooth fairy: A cone-beam computed tomography segmentation challenge. *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2023*, 2023. 5

[10] Emmanuelle Bourigault, Abdullah Hamdi, and Amir Jamaludin. X-diffusion: Generating detailed 3d mri volumes from a single image using cross-sectional diffusion models, 2024. 1

[11] Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Universeg: Universal medical image segmentation. *arXiv preprint arXiv:2304.06131*, 2023. 3, 7, 8

[12] Bingzhi Chen, Yishu Liu, Zheng Zhang, Guangming Lu, and Adams Wai Kin Kong. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2023. 2

[13] Huijun Chen. Cerebral artery segmentation challenge. *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2023.*, 2023. 5

[14] 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. 2, 7

[15] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023. 3

[16] Junlong Cheng et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023. 7

[17] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016. 1, 2

[18] Guoyao Deng, Ke Zou, Kai Ren, Meng Wang, Xuedong Yuan, Sancong Ying, and Huazhu Fu. Sam-u: Multi-box prompts triggered uncertainty estimation for reliable sam in medical image. *arXiv preprint arXiv:2307.04973*, 2023. 3

[19] Guoyao Deng et al. Sam-u: Multi-box prompts triggered uncertainty estimation for reliable sam in medical image. *arXiv preprint arXiv:2307.04973*, 2023. 7

[20] Hao Ding, Changchang Sun, Hao Tang, Dawen Cai, and Yan Yan. Few-shot medical image segmentation with cycle-resemblance attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2488–2497, 2023. 3

[21] Huihui Fang, Fei Li, Huazhu Fu, Xu Sun, Xingxing Cao, Jaemin Son, Shuang Yu, Menglu Zhang, Chenglang Yuan, Cheng Bian, et al. Refuge2 challenge: Treasure for multi-domain learning in glaucoma assessment. *arXiv preprint arXiv:2202.08994*, 2022. 4, 5

[22] Xi Fang and Pingkun Yan. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. *IEEE Transactions on Medical Imaging*, 39(11):3619–3629, 2020. 5, 8

[23] Randy L Gollub, Jody M Shoemaker, Margaret D King, Tonya White, Stefan Ehrlich, Scott R Sponheim, Vincent P Clark, Jessica A Turner, Bryon A Mueller, Vince Magnotta, et al. The mcic collection: a shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia. *Neuroinformatics*, 11:367–388, 2013. 4

[24] Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. Mvtn: Multi-view transformation network for 3d shape recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–11, 2021. 2

[25] Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. Voint cloud: Multi-view point cloud representation for 3d understanding. In *The Eleventh International Conference on Learning Representations*, 2023.

[26] Abdullah Hamdi, Faisal AlZahrani, Silvio Giancola, and Bernard Ghanem. Mvtn: Learning multi-view transformations for 3d understanding. *International Journal of Computer Vision*, 2024. 2

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

images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2022. 7

[28] 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. 3, 7

[29] Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical image analysis*, 67:101821, 2021. 4

[30] Nicholas Heller, Fabian Isensee, Dasha Trofimova, Resha Tejpaul, Zhongchen Zhao, Huai Chen, Lisheng Wang, Alex Golts, Daniel Khapun, Daniel Shats, Yoel Shoshan, Flora Gilboa-Solomon, Yasmeen George, Xi Yang, Jianpeng Zhang, Jing Zhang, Yong Xia, Mengran Wu, Zhiyang Liu, Ed Walczak, Sean McSweeney, Ranveer Vasdev, Chris Hornung, Rafat Solaiman, Jamee Schoephoerster, Bailey Abernathy, David Wu, Safa Abdulkadir, Ben Byun, Justice Spriggs, Griffin Struyk, Alexandra Austin, Ben Simpson, Michael Hagstrom, Sierra Virnig, John French, Nitin Venkatesh, Sarah Chan, Keenan Moore, Anna Jacobsen, Susan Austin, Mark Austin, Subodh Regmi, Nikolaos Papanikolopoulos, and Christopher Weight. The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct, 2023. 5

[31] Moritz R Hernandez Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, Waldo Valenzuela, Mauricio Reyes, Maria Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, et al. Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific data*, 9(1):762, 2022. 4

[32] AD Hoover, Valentina Kouznetsova, and Michael Goldbaum. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical imaging*, 19(3):203–210, 2000. 4, 5

[33] 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. 7

[34] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, 1957. 4, 1

[35] Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, and Yefeng Zheng. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12341–12351, 2021. 2

[36] Yuanfeng Ji, Haotian Bai, Jie Yang, Chongjian Ge, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *arXiv preprint arXiv:2206.08023*, 2022. 4

[37] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021. 4

[38] 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. *arXiv preprint arXiv:2304.02643*, 2023. 1, 3, 7

[39] Hugo J Kuijf, J Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M Jorge Cardoso, Adria Casamitjana, et al. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. *IEEE transactions on medical imaging*, 38(11): 2556–2568, 2019. 4

[40] Maria Kuklisova-Murgasova, Paul Aljabar, Latha Srinivasan, Serena J Counsell, Valentina Doria, Ahmed Serag, Ioannis S Gousias, James P Boardman, Mary A Rutherford, A David Edwards, et al. A dynamic 4d probabilistic atlas of the developing brain. *NeuroImage*, 54(4):2750–2763, 2011. 4

[41] Zoé Lambert, Caroline Petitjean, Bernard Dubray, and Su Kuan. Segthor: Segmentation of thoracic organs at risk in ct images. In *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2020. 4

[42] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, page 12, 2015.

[43] Guillaume Lemaître, Robert Martí, Jordi Freixenet, Joan C Vilanova, Paul M Walker, and Fabrice Meriaudeau. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: a review. *Computers in biology and medicine*, 60:8–31, 2015. 4

[44] Yiwen Li, Yunguan Fu, Iani JMB Gayo, Qianye Yang, Zhe Min, Shaheer U Saeed, Wen Yan, Yipei Wang, J Alison Noble, Mark Emberton, et al. Prototypical few-shot segmentation for cross-institution male pelvic structures with spatial registration. *Medical Image Analysis*, 90:102935, 2023. 3

[45] Geert Litjens, Robert Toth, Wendy Van De Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram Van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014. 4

[46] Xiangde Luo, Wenjun Liao, Jianghong Xiao, Tao Song, Xiaofan Zhang, Kang Li, Guotai Wang, and Shaoting Zhang. Word: Revisiting organs segmentation in the whole abdominal region. *arXiv preprint arXiv:2111.02403*, 1(2):13, 2021.

[47] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6695–6714, 2021. 4

[48] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 2, 4, 5, 7, 1

[49] Yuhui Ma, Huaying Hao, Jianyang Xie, Huazhu Fu, Jiong Zhang, Jianlong Yang, Zhen Wang, Jiang Liu, Yalin Zheng, and Yitian Zhao. Rose: a retinal oct-angiography vessel segmentation dataset and new model. *IEEE transactions on medical imaging*, 40(3):928–939, 2020. 4

[50] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 1, 2

[51] José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel Van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis*, 59:101570, 2020. 4

[52] Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 762–780. Springer, 2020. 3, 7

[53] Félix Quinton, Romain Popoff, Benoît Presles, Sarah Leclerc, Fabrice Meriaudeau, Guillaume Nodari, Olivier Lopez, Julie Pellegrinelli, Olivier Chevallier, Dominique Ginhac, et al. A tumour and liver automatic segmentation (atlas) dataset on contrast-enhanced magnetic resonance imaging for hepatocellular carcinoma. *Data*, 8(5):79, 2023. 5

[54] Perry Radau, Yingli Lu, Kim Connelly, Gideon Paul, Alexander J Dick, and Graham A Wright. Evaluation framework for algorithms segmenting short axis cardiac mri. *The MIDAS Journal*, 2009. 4

[55] 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. 1, 3, 7, 8

[56] Erik Ziegler Ron Kikinis, Steve Pieper. Mediastinal lymph node quantification (lnq): Segmentation of heterogeneous ct data. *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2023. 5

[57] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 1

[58] Saikat Roy, Tassilo Wald, Gregor Koehler, Maximilian R Rokuss, Nico Disch, Julius Holzschuh, David Zimmerer, and Klaus H Maier-Hein. Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model. *arXiv preprint arXiv:2304.05396*, 2023. 3

[59] Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, SQ Truong, CD Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, AY Ng, et al. Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation. *MedRxiv*, 2021. 4

[60] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas Van Den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical image analysis*, 42:1–13, 2017.

[61] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019. 4

[62] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 9197–9206, 2019. 7

[63] Lu Wang, Dongxue Liang, Xiaolei Yin, Jing Qiu, Zhiyun Yang, Junhui Xing, Jianzeng Dong, and Zhaoyuan Ma. Coronary artery segmentation in angiographic videos utilizing spatial-temporal information. *BMC medical imaging*, 20: 1–10, 2020. 5, 8

[64] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2023. 1

[65] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023.

[66] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022. 1

[67] Qiaoqiao Wei et al. Focused and collaborative feedback integration for interactive image segmentation. In *CVPR*, pages 18643–18652, 2023. 7

[68] Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C Cattin. Diffusion models for implicit image segmentation ensembles. *arXiv preprint arXiv:2112.03145*, 2021. 7

[69] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *CVPR*, pages 16133–16142, 2023. 7

[70] Junde Wu and Min Xu. One-prompt to segment all medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11302–11312, 2024. 2, 3, 4, 7, 8

[71] Junde Wu, Shuang Yu, Wenting Chen, Kai Ma, Rao Fu, Hanruo Liu, Xiaoguang Di, and Yefeng Zheng. Leveraging undiagnosed data for glaucoma classification with teacher-student learning. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 731–740. Springer, 2020. 3

[72] Junde Wu, Huihui Fang, Yu Zhang, Yehui Yang, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion

probabilistic model. *arXiv preprint arXiv:2211.00611*, 2022. 1, 3, 7

[73] Junde Wu, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, Yueming Jin, and Tal Arbel. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023. 3

[74] Junde Wu, Wei Ji, Huazhu Fu, Min Xu, Yueming Jin, and Yanwu Xu. Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6030–6038, 2024. 1, 3

[75] Zhongnuo Yan, Tong Han, Yuhao Huang, Lian Liu, Han Zhou, Jiongquan Chen, Wenlong Shi, Yan Cao, Xin Yang, and Dong Ni. A Foundation Model for General Moving Object Segmentation in Medical Images, 2024. arXiv:2309.17264. 7

[76] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track Anything: Segment Anything Meets Videos, 2023. arXiv:2304.11968. 7

[77] Kaidong Zhang et al. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023. 7

[78] Amy Zhao, Guha Balakrishnan, Fredo Durand, John V Guttag, and Adrian V Dalca. Data augmentation using learned transformations for one-shot medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8543–8553, 2019. 7

[79] Xin Zheng, Yong Wang, Guoyou Wang, and Jianguo Liu. Fast and robust segmentation of white blood cell images by self-supervised learning. *Micron*, 107:55–71, 2018. 5

[80] Tianfei Zhou et al. Volumetric memory network for interactive medical image segmentation. *Medical Image Analysis*, 83:102599, 2023. 7

A. Why Does Self-Sorting Work?

The effectiveness of the self-sorting memory bank $\mathcal{M}^{\text{sort}}$ in MedSAM-2 can be understood through the lens of information theory, particularly in terms of entropy and mutual information.

Let x_t denote the input image at time t, let Y_t denote the predicted segmentation mask at time t, and let Z_t denote the ground truth at time t. The mutual information $I(Y_t; Z_t|x_t)$ measures the amount of information that the predicted segmentation mask contain about the ground truth:

$$I(Y_t; Z_t|x_t) = I(\mathcal{D}(\mathcal{A}(\mathbf{F}_t(x_t), \mathcal{M}_t, \mathbf{Q}_t)); Z_t|x_t) \quad (7)$$

Given that the input image x_t is specified, both $\mathbf{F}_t$ and $\mathbf{Q}_t$ remain constant. Consequently, the only variable is the selected memory bank $\mathcal{M}_t$. Therefore, increasing the mutual information between $\mathcal{M}_t$ and Y_t, conditioned on x_t, will lead to an improved predicted mask.

By leveraging the relationship between mutual information and conditional entropy, the following decomposition can be derived:

$$I(M_t; Z_t|x_t) = H(Z_t) - H(Z_t|M_t, x_t) \quad (8)$$

where $H(Z_t|x_t)$ denotes the entropy of the ground truth given the input image, and $H(Z_t|M_t, x_t)$ represents the conditional entropy of the ground truth Z_t given the known M_t, defined as:

$$H(Z_t|M_t, x_t) = -\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x, y)\log\frac{p(x, y)}{p(x)} \quad (9)$$

$$= \mathbb{E}[-\log\frac{p(x, y)}{p(x)}] \quad (10)$$

$$= \mathbb{E}[-\log(p(y|x))] \quad (11)$$

Since $-\log(p(y|x))$ can be interpreted as the amount of information required to describe the random variable y given the value of x, the conditional entropy $H(Z_t|M_t, x_t)$ can thus be viewed as the expected information needed to describe the ground truth Z_t given the selected memory bank M_t.

By selecting embeddings based on the highest confidence scores, the self-sorting memory bank seeks to maximize $I(M_t; Z_t|x_t)$ by minimizing $H(Z_t|M_t, x_t)$. Given that $H(Z_t|x_t)$ is assumed to be constant, high-confidence embeddings provide more information regarding the output, thereby reducing the information required to describe Z_t. This reduction leads to a smaller $H(Z_t|M_t, x_t)$ and, consequently, an increase in mutual information. This increase in mutual information suggests that the model is able to make more accurate and reliable predictions based on the embeddings stored in the memory bank.

Furthermore, the self-sorting mechanism introduces variability in the selection of memory embeddings, as it is not limited to the most recent frames but instead selects from all past frames based on confidence scores. This variability enhances the diversity of information within the memory bank M_t, potentially decreasing the additional information needed to infer Z_t, especially in contexts where frames change rapidly and significantly. As a result, the self-sorting mechanism can yield a smaller $H(Z_t|M_t, x_t)$, thereby increasing the mutual information $I(M_t; Z_t|x_t)$.

This increased entropy in the memory embeddings enhances the model's ability to generalize. According to the principle of maximum entropy [34], a model that considers a broader distribution of features is less likely to overfit to specific patterns in the training data and is better equipped to handle variability in unseen data. By increasing both the mutual information between the memory embeddings and the output and the entropy of the memory embeddings themselves, the self-sorting memory bank improves the robustness and generalization of MedSAM-2.

Consequently, the model is better suited to handle unordered medical images, as it leverages the most informative and diverse embeddings for segmentation. This leads to enhanced performance across diverse medical imaging tasks after training with standard segmentation loss [48].

B. Experimental Details

B.1. Evaluation Metrics

We use Intersection over Union (IoU) and Dice Score to assess the performance of models in medical image segmentation.

Intersection over Union (IoU) Intersection over Union (IoU), also known as the Jaccard Index, is a measure used to evaluate the accuracy of an object detector on a specific dataset. It quantifies the overlap between two datasets by dividing the area of overlap between the predicted segmentation and the ground truth by the area of their union. The formula for IoU is given by:

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

IoU provides a clear metric at the object level, assessing both the size and position accuracy of the prediction relative to the actual data, which is particularly useful for understanding detection model performance.

Dice Score The Dice Score, or Dice Coefficient, is a statistical tool that compares the similarity between two samples. It is particularly prevalent in medical image analysis due to its sensitivity to the size of the objects being examined. The Dice Score is calculated by taking twice the area of overlap between the predicted and actual segmentations, divided by

the total number of pixels in both the prediction and the ground truth. The formula for the Dice Score is:

$$\text{Dice Score} = \frac{2 \times \text{Area of Overlap}}{\text{Area of Prediction} + \text{Area of Ground Truth}}$$

This score ranges from 0 to 1, where a score of 1 indicates perfect agreement between the model's predictions and the ground truth. The Dice Score is known for its robustness against the size variability of the segmented objects, making it extremely valuable in medical applications where such variability is common.

Both metrics, IoU and Dice Score, provide comprehensive insights into model accuracy, with Dice Score being particularly effective in scenarios involving significant variations in object size.

Hausdorff Distance (HD95) Metric The Hausdorff Distance (HD95) is a metric used to determine the extent of discrepancy between two sets of points, typically used to evaluate the accuracy of object boundaries in image segmentation tasks. It is particularly useful for quantifying the worst-case scenario of the distance between the predicted segmentation and the ground truth boundary.

The Hausdorff Distance measures the maximum distance of a set to the nearest point in the other set. For image segmentation, this means calculating the greatest of all the distances from a point in the predicted boundary to the closest point in the ground truth boundary and vice versa. The formula for the Hausdorff Distance is given by:

$$\text{HD} = \max\left(\sup_{x \in A} \inf_{y \in B} d(x, y), \sup_{y \in B} \inf_{x \in A} d(x, y)\right)$$

where A and B represent the sets of boundary points of the ground truth and the predicted segmentation, respectively, and $d(x, y)$ denotes the Euclidean distance between points x and y.

While the Hausdorff Distance provides a strict measure by considering the maximum distance, it can be overly sensitive to outliers. To mitigate this, the HD95 metric is used, which considers only the 95th percentile of the distances instead of the maximum. This adjustment makes the HD95 less sensitive to outliers and provides a more robust measure for practical applications:

$$\text{HD95} = \text{95th percentile of } \{d(x, y) \mid x \in A, y \in B\}$$

This metric is particularly relevant in medical image analysis where precision in the segmentation of anatomical structures is critical and outliers can distort the evaluation of segmentation performance.

C. Data

C.1. Data Preprocessing

The original 3D datasets contain a variety of CT and MRI images stored in DICOM, NRRD, or MHD formats. To ensure uniformity and compatibility, all images, regardless of modality, were converted to the widely used NIfTI format. This conversion also included grayscale images, such as X-Ray and Ultrasound, while RGB images depicting endoscopy, dermoscopy, fundus, and pathology were converted into the PNG format. For tasks involving multiple segmentation targets, each target is treated as an individual task for predicting a binary segmentation mask. During the inference stage for predicting multiple targets, we predict a soft segmentation mask with a fixed threshold (averaging 0.5) to filter out uncertain predictions.

Notably, image intensities varied significantly across modalities. For instance, CT images ranged from -2000 to 2000, MRI values ranged from 0 to 800, endoscopy/ultrasound images from 0 to 255, and some modalities were already within the range 0 to 1. To harmonize this variability, intensity normalization was systematically conducted for each modality. The default normalization during training and inference involved normalizing each image independently by subtracting its mean and dividing by its standard deviation. For MRI, X-Ray, ultrasound, mammography, and Optical Coherence Tomography (OCT) images, intensity values were trimmed to fall between the 0.5th and 99.5th percentiles before normalization. If cropping resulted in a 25% or greater reduction in average size, a mask for central non-zero voxels was generated, and normalization was confined to this mask, disregarding surrounding zero voxels. For CT images, Hounsfield units were first normalized using window width and level values before applying standard normalization. Furthermore, since CT intensity values quantitatively reflect tissue properties, we applied a global normalization scheme to all images. Specifically, this involved clipping intensity values to the 0.5th and 99.5th percentiles of foreground voxels, followed by normalization using the global foreground mean and standard deviation.

To standardize image sizes, the provided samples were first cropped to their non-zero regions and then uniformly resized to 256×256. During resizing, we used bi-cubic interpolation for images and nearest-neighbor interpolation for masks, ensuring smooth standardization and compatibility across all images. For 3D images, we generally operated on the two axes with the highest resolution. If all three axes were isotropic, the two trailing axes were used for slice extraction. The channel was replicated threefold to ensure consistency during processing. For slice-based processing, no resampling along the out-of-plane axis was required.

Masks with multiple classes were processed into individual masks for each class. Masks containing multiple

connected components were dissected, while original masks were retained in cases with only one component. Additionally, masks where the target area was less than 0.153% of the total image (equivalent to areas smaller than 100 pixels in a resized 256×256 resolution) were excluded. This deliberate decision ensures the dataset only includes significant and well-defined target areas. The standardized preprocessing pipeline was consistently applied across all compared methods to ensure a fair and unbiased comparison.

C.2. Data Augmentation

During training, we utilize a range of data augmentation techniques, dynamically computed on the CPU. Spatial augmentations are applied, including rotations, scaling, Gaussian noise, Gaussian blur, intensity and contrast adjustments, low-resolution simulation, gamma correction, and flipping. To enhance image variability, most augmentations involve random parameter selection from predefined ranges. The application of these augmentations follows stochastic principles, adhering to predefined probabilities. Consistent augmentation parameters are maintained across datasets. Each augmentation is individually applied to both the template sample and the query sample.

Details of the augmentation techniques are as follows:

1. **Rotation:** Applied with a probability of 0.15 to all images. The rotation angle is uniformly sampled from the range $[-25, 25]$.
2. **Scaling:** Scaling is achieved by multiplying image coordinates with a scaling factor. Scale factors smaller than 1 result in a "zoom out" effect, while values larger than 1 create a "zoom in" effect. The scaling factor is uniformly sampled from $[0.7, 1.4]$, with a probability of 0.15.
3. **Gaussian Noise:** Zero-centered Gaussian noise is independently added to each sample with a probability of 0.15. The noise variance is sampled from $[0, 0.1]$, considering that normalized sample intensities are close to zero mean and unit variance.
4. **Gaussian Blur:** Blurring is applied with a probability of 0.15 per sample. For each task, it occurs with a probability of 0.5 per modality. The Gaussian kernel size is uniformly sampled from $[0.5, 1.5]$ for each modality.
5. **Intensity Adjustment:** Intensities are modified by multiplying them with a factor uniformly sampled from $[0.65, 1.2]$ with a probability of 0.15. Alternatively, intensities can be flipped using $1 -$ image. Intensity augmentation is not applied to labels. After multiplication, the values are clipped to the original intensity range.
6. **Low Resolution:** Applied with a probability of 0.25 per sample and 0.5 per associated modality. This augmentation downsamples the triggered modalities by a factor uniformly sampled from $[1, 2]$ using nearest neighbor interpolation, then resamples them back to the original size using cubic interpolation.
7. **Gamma Augmentation:** Applied with a probability of 0.15. Image intensities are first scaled to a range of 0 to 1, followed by a nonlinear intensity transformation defined as $x_{\text{new}} = x_{\text{old}}^{\gamma}$, where γ is uniformly sampled from $[0.7, 1.5]$. The intensities are then scaled back to their original range. This augmentation is applied after the intensity flip, also with a probability of 0.15.
8. **Spatial Flip:** Samples are flipped along all axes with a probability of 0.5.