RFMEDSAM 2: AUTOMATIC PROMPT REFINEMENT FOR MEDICAL IMAGE SEGMENTATION WITH SAM 2

Anonymous authors
Paper under double-blind review

ABSTRACT

The Segment Anything Model 2 (SAM 2) is a prompt-driven foundation model that generalizes SAM to both image and video segmentation, achieving strong zero-shot segmentation performance. However, despite its potential for medical image analysis, SAM 2 inherits key limitations from its predecessor, including binary mask outputs, inability to predict semantic labels, and reliance on precise prompts for target localization. Moreover, its direct application to medical image segmentation remains suboptimal. To investigate its upper performance bound, we fine-tune SAM 2 with targeted architectural modifications and oracle (ground-truth) prompts, reaching a Dice scores of 92.3% on BTCV, surpassing the state-of-the-art nnUNet by 9.8%; we stress that this is a diagnostic upper bound, not a comparable operating point, since it consumes ground-truth boxes and class labels. To eliminate the reliance on manual prompts, we develop an automatic prompt generator based on U-Net, which predicts coarse masks and bounding boxes that are subsequently refined through SAM 2's progressive dual-stage inference. Furthermore, we propose QRefine-Former to transfer domain-specific medical knowledge from the U-Net prompt generator to the SAM 2 foundation model, further improving segmentation performance. Extensive experiments on the AMOS22, ACDC, and BTCV datasets demonstrate the superiority of our method, achieving Dice scores of 91.0% and 88.6% on AMOS22 Task 1 and Task 2, 93.6% Dice on ACDC, and 86.8% Dice on BTCV.

1 INTRODUCTION

Medical image segmentation underpins downstream clinical workflows, including diagnosis, anomaly detection, and surgical planning. While deep learning has driven steady progress Ronneberger et al. (2015); Isensee et al. (2019); Zhou et al. (2021), medical datasets frequently lack large-scale, high-quality annotations, making it challenging to train highly expressive models effectively. As a result, architectures with strong inductive biases, particularly CNNs, have remained competitive when trained from scratch.

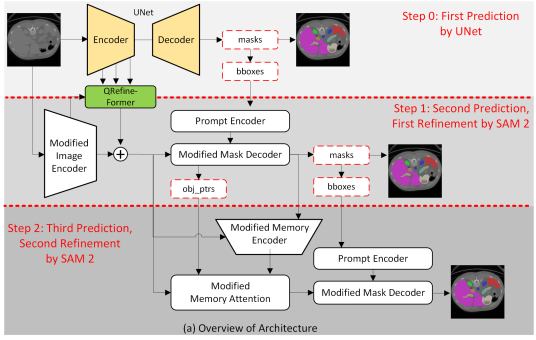


Figure 1: Overview of our proposed RFMedSAM 2.

Foundation models Devlin et al. (2018); He et al. (2022) shift the paradigm from task-specific training to "pre-train, then adapt," exhibiting strong zero-/few-shot transfer across modalities OpenAI (2023); Radford et al. (2021). The Segment Anything Model (SAM) Kirillov et al. (2023), trained on SA-1B, brought prompt-driven segmentation to natural images and has since inspired adaptations to medical imaging Ma et al. (2024); Xie et al. (2024; 2025); Deng et al. (2023); Bui et al. (2024). Building on SAM, SAM 2 generalizes to both images and videos, enabling real-time propagation from a single prompt. On BTCV Landman et al. (2015), SAM 2 improves over SAM (82.77% vs. 81.89% DSC; Tab. 1), indicating promise for medical segmentation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

However, SAM 2 inherits critical limitations for clinical use: it produces binary masks, cannot predict semantic labels, and depends on precise prompts for target localization. Applied naively, both SAM and SAM 2 remain suboptimal relative to specialized medical segmentation models.

The main contributions of this paper are three-fold: advancing the adaptation of foundation models, eliminating the reliance on manually annotated prompts and progressive dual-stage refinement, and transferring domain-specific medical knowledge to enhance foundation models.

- We propose RFMedSAM 2, a refinement-based framework for medical image segmentation that adapts SAM 2 through targeted architectural modifications, autonomous prompt generation, and query-guided medical knowledge transfer.
- We investigate the upper performance bound of adapting SAM 2 for medical image segmentation under ground-truth prompts, reaching 92.3% DSC on BTCV Landman et al. (2015) with oracle prompts, 9.8% above nnU-Net; this quantifies headroom rather than a deployable result.
- We develop an automatic U-Net-based prompt generator that predicts coarse masks and bounding boxes, followed by progressive dual-stage refinement in SAM 2 to improve segmentation accuracy.
- We propose QRefine-Former, a lightweight query-based module that transfers task-specific medical knowledge from the U-Net prompt generator to the SAM 2 foundation model, enabling SAM 2 to leverage both domain-specific medical priors and its strong generalization capability.
- Extensive experiments on AMOS22 Ji et al. (2022) task 1 and task 2, ACDC Bernard et al. (2018), and BTCV Landman et al. (2015) demonstrate that RFMedSAM 2 achieves consistent state-of-the-art performance, exceeding nnU-Net by 1.7%, 0.6%, 2.2%, and 4.3% DSC, respectively.

2 Related Work

2.1 Medical Image Segmentation

Deep learning has largely replaced traditional pipelines for medical image segmentation. U-Net Ronneberger et al. (2015) remains foundational due to its encoder–decoder design with skip connections that preserve fine spatial detail. Building on this, nnU-Net Isensee et al. (2019) introduced an automated configuration pipeline that adapts U-Net-style architectures to diverse datasets, establishing robust baselines. Subsequent convolutional methods, including 3D-UXNet Lee et al. (2022), MedNeXt Roy et al. (2023), and STU-Net Huang et al. (2023), further advanced volumetric modeling and efficiency. More recently, transformer-based models such as UNETR Hatamizadeh et al. (2022), SwinUNETR Hatamizadeh et al. (2021), and nnFormer Zhou et al. (2021) leverage self-attention to capture long-range dependencies. Although these architectures are purpose-built for medical segmentation and often trained from scratch, their strong inductive biases and data demands can limit adaptability across modalities, scanners, and label spaces.

2.2 SAM and SAM 2 for Medical Segmentation

The Segment Anything Model (SAM) Kirillov et al. (2023), pre-trained on over one billion masks from 11 million natural images, is a powerful prompt-based foundation model with strong zero-shot capability. Following the "pre-train then adapt" paradigm, SAM has been extended to medical imaging in works such as MedSAM Ma et al. (2024), MaskSAM Xie et al. (2024), and Self-Prompt SAM Xie et al. (2025), among others Zhang et al. (2024); Deng et al. (2023); Ma & Wang (2023); Wu et al. (2023); Li et al. (2023); Gong et al. (2023). These efforts demonstrate flexibility but also expose limitations of the original formulation: SAM outputs binary masks without semantic labels and is sensitive to prompt quality, which constrains fully automatic clinical use.

SAM 2 generalizes SAM to both images and videos, employing memory mechanisms for real-time propagation from a single prompt. Domain adaptations (e.g., SAM2-Adapter Chen et al. (2024), Polyp SAM 2 Mansoori et al. (2024); see also Yu et al. (2024); Liu et al. (2024)) investigate parameter-efficient tuning and task-specific refinements. Nevertheless, core challenges persist: (i) lack of semantic label inference, (ii) dependence on accurate prompts, and (iii) added training/inference complexity and memory cost due to temporal attention. Current research mitigates these issues with auxiliary models that generate reliable prompts and learning schemes that dynamically refine prompts during training. Addressing these challenges is key to unlocking the full potential of SAM 2 for automated medical segmentation and bridging the gap between foundation-model generality and practical, state-of-the-art performance.

Prompt	Bounding boxes as prompts							Central points as prompts						
Method (#Frames/Class)	SAM (All)	SAM 2 (All)		SAM 2 (Two)		SAM 2 (One)		SAM (All)	SAM 2 (All)		SAM 2 (Two)		SAM 2 (One)	
#Frames for Step-2	–	All	Unprompted	All	Unprompted	All	Unprompted	–	All	Unprompted	All	Unprompted	All	Unprompted
DSC (%)	81.89	81.17	82.77	68.75	68.03	45.00	44.07	8.86	3.81	4.90	2.11	3.43	2.53	4.59

Table 1: Performance comparison of SAM and SAM 2 under different prompt types on the BTCV dataset.

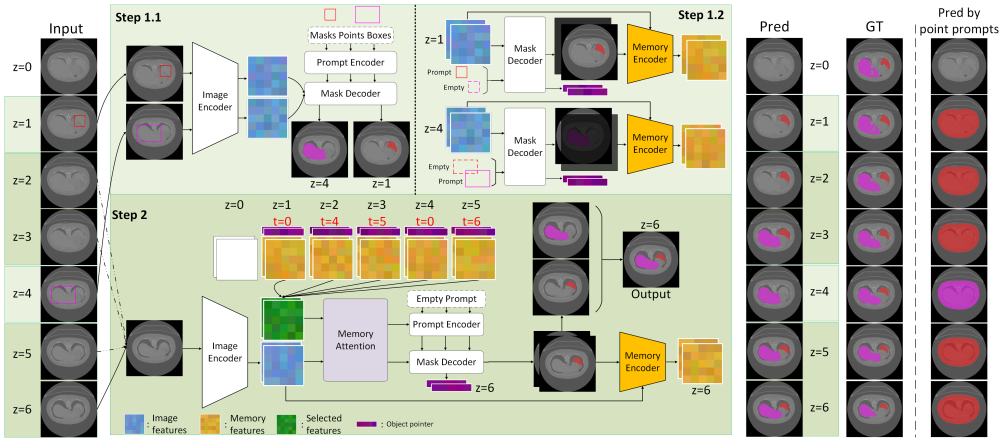


Figure 2: Overview of SAM 2. The pipeline includes steps for prompted and unprompted frames.

3 Analysis of SAM and SAM 2

Before introducing our proposed refinements, we analyze SAM and SAM 2 to understand their architectural design, limitations, and performance behavior in medical segmentation tasks. This analysis provides motivation and insight for developing RFMedSAM 2.

3.1 SAM and SAM 2 Overview

Model components. SAM and SAM 2 share an image encoder, a prompt encoder, and a mask decoder. The encoder produces dense visual embeddings, the prompt encoder embeds user inputs (points, boxes, or masks)and the decoder fuses both into binary masks. SAM 2 replaces SAM's plain ViT backbone with the hierarchical Hiera (Ryali et al., 2023), and adds *memory attention*, which conditions current-frame features on stored representations and object pointers, and a *memory encoder*, which fuses current-frame features with predicted masks into memory features.

Pipeline. As illustrated in Fig. 2, SAM 2 operates in two stages: (i) *Prompted-frame processing* (Steps 1.1 and 1.2), where frames with explicit user or ground-truth prompts are segmented independently. To ensure comprehensive segmentation, the number of segmented instances is determined by the number of explicit prompts. The resulting masks and object pointers are stored in the memory encoder to generate memory features for subsequent frames. (ii) *Unprompted-frame processing* (Step 2), where memory attention propagates temporal cues from previously segmented frames as implicit prompts to segment frames without explicit prompts. The memory attention module aggregates memory features from past prompted and unprompted frames to provide contextual cues to segment the current frame. Prompted frames are assigned a temporal position 0 and unprompted frames increasing positions up to 6, weighting recent frames more heavily. Pinning every prompted frame at position 0 distorts relative positions and can cause segmentation errors.

3.2 Analysis and Insights

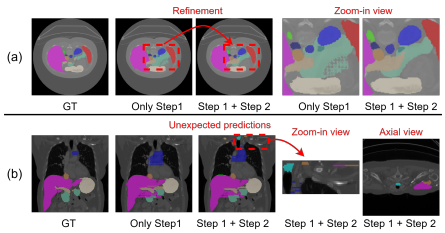


Figure 3: Benefits of refinement in Step 2.

Tab. 1 summarizes experiments on BTCV under different SAM and SAM 2 settings using ground-truth prompts without architectural modifications. For each frame containing an object, a single prompt is provided. Since the official SAM 2 supports up to two prompted frames per object during training, we evaluate one, two, and all prompted-frame settings for fair comparison with SAM, which requires prompts for every frame. For Step 2 in SAM 2, we compare (i) the official

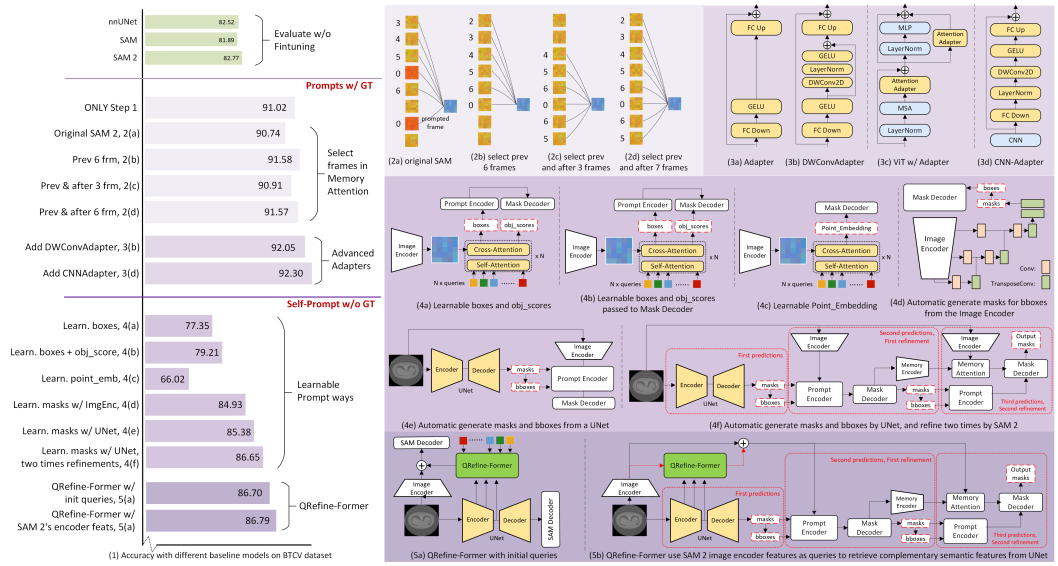


Figure 4: (1) Performance comparisons based on proposed methods. (2) Ablation studies on frame selection strategies. (3) Proposed Adapters. (4) Ablation studies on prompt generators.

protocol, which is applied only to unprompted frames, and (ii) a variant that applies Step 2 to all frames to evaluate its refinement capability.

- *(i) Bounding-box vs. center-point prompts.* As shown in Tab. 1, center-point prompts achieve less than 10% Dice for both SAM and SAM 2, indicating that they are insufficient for accurate segmentation. In contrast, bounding-box prompts substantially improve performance and are therefore adopted in all subsequent experiments.
- *(ii) Per-frame prompts.* SAM 2 achieves its best performance (82.77% Dice) when every frame contains a bounding box for each object, highlighting the importance of dense temporal prompting.
- *(iii) SAM vs. SAM 2.* Under identical per-frame prompting, SAM achieves 81.89% Dice, whereas SAM 2 improves to 82.77%, demonstrating the benefit of its memory-based architecture.
- *(iv) Step 2 refinement.* Applying Step 2 to all frames slightly decreases Dice (82.77% $\rightarrow$ 81.17%) but can refine local boundaries, as shown in Fig. 3(a). However, Fig. 3(b) reveals a limitation: assigning all prompted frames a temporal position of 0 may lead to inaccurate temporal positioning, causing false positives in unrelated frames. Correct temporal positioning can alleviate this issue.
- *(v) Streaming operation.* Except for memory attention, all SAM 2 modules process 2D frames independently, reducing memory consumption. Memory attention aggregates features from previous frames to establish temporal context, providing an effective trade-off between accuracy and efficiency. We therefore preserve this streaming design in RFMedSAM 2 (Fig. 4).

4 Proposed RFMedSAM 2 Architecture

The final model. RFMedSAM 2 (Fig. 1) runs three stages. *Step 0*: a 3D U-Net predicts coarse multi-class masks, from which per-slice bounding boxes are derived, while QRefine-Former reads U-Net features and injects query features into the SAM 2 image embeddings. *Step 1*: the modified SAM 2 encodes the image and those boxes, and its mask decoder emits masks and object pointers that yield refreshed boxes; the modified memory encoder turns masks and frame features into memory features. *Step 2*: modified memory attention combines memory from preceding frames with the current QRefine-Former-enhanced features, and the mask decoder produces the final mask. Two stacked convolutions map arbitrary-channel medical volumes into SAM 2's three-channel input; DWConvAdapters and CNN-Adapters are inserted throughout, and all 224.4M SAM 2 parameters stay frozen (69.5M trainable). Full details are in Appendix.

In this section, we present RFMedSAM 2 from three perspectives. i) First, we explore the upper performance bound of adapting SAM 2 for medical image segmentation by introducing targeted architectural modifications while preserving the pretrained SAM 2 backbone. Ground-truth (GT) annotations are used to generate prompts, including both bounding boxes and class labels. This setting

allows us to fully exploit the segmentation capability of the adapted SAM 2 architecture. ii) Second, we consider the practical setting where GT prompts are unavailable. To approach the upper bound established with GT prompts, we explore effective prompt generators that automatically produce high-quality prompts for SAM 2. iii) Third, we introduce QRefine-Former to bridge the U-Net-based prompt generator, a medical image segmentation expert, and the general-purpose foundation model (SAM 2). QRefine-Former extracts informative query features from U-Net representations and injects them into the SAM 2 image embeddings, enabling SAM 2 to leverage both its strong generalization capability and task-specific medical priors.

4.1 Upper-Bound Adaptation of SAM 2

To explore the upper performance bound of adapting SAM 2 for medical image segmentation, we first establish a baseline adaptation while preserving the pretrained SAM 2 backbone. To accommodate multi-channel medical images while remaining compatible with the RGB input expected by SAM 2, two stacked convolutional layers transform the input while preserving spatial details. In addition, the original adapters are inserted after the self-attention and cross-attention blocks and in parallel with the MLP layers (Fig. 4(3a, 3c)) in SAM 2. This baseline serves as the starting point for the subsequent architectural enhancements.

4.1.1 Frame Selection and Memory-Attention

SAM 2's memory attention plays a crucial role in propagating segmentation cues across frames, but its original temporal-positioning strategy assigns a temporal position of 0 to all prompted frames, which can introduce ambiguity and reduce accuracy. We therefore explore alternative schemes (Fig. 4(2)). The original strategy (Fig. 4(2a)) achieves 90.74% DSC, while our structured scheme (Fig. 4(2b)), which assigns position 0 to the current frame and progressively larger positions to up to six preceding frames, achieves 91.58% DSC by distinguishing current from past frames. Strategies incorporating forward/backward frame selection (Figs. 4(2c)–(2d)) either reduced performance or increased memory overhead, confirming the effectiveness of our design.

We preserve the streaming operation of SAM 2: most modules process frames independently to reduce memory usage, while memory attention is the sole component that integrates contextual information from previous frames, an efficient and effective design for sequential image segmentation.

4.1.2 Novel Adapters for Enhanced Fine-Tuning

To enable parameter-efficient fine-tuning while retaining SAM 2's strong zero-shot generalization, we introduce two complementary adapter mechanisms that enhance spatial and convolutional processing. i) Depth-Wise Convolutional Adapters (DWConvAdapters), Fig. 4(3b)) are integrated into the image encoder, memory attention, , and mask decoder. Depth-wise convolution applied independently per channel keeps cost low while preserving the fine spatial detail that anatomical boundaries require; it adds 0.47% DSC. ii) CNN-Adapters (Fig. 4(3d)) further refine feature adaptation within convolutional layers, particularly in the FPN module and memory encoder, enhancing multi-scale representation and feature fusion to achieve an additional 0.25% DSC gain.

We employ DWConvAdapters within image-embedding attention blocks, CNN-Adapters within convolutional layers, and retain the original adapters for point-embedding attention. Collectively, under oracle prompts these refinements reach 92.3% DSC on BTCV (Tab. 3), 5.4% above the strongest fully-supervised baseline in that table.

4.2 Advanced Prompt Generation

GT prompts are unavailable in practice, and annotating a box per organ per slice is impractical. To close the gap to the GT-prompt bound we develop an automatic prompt generator that refines both the prompts and the segmentation during training. We investigate six strategies (Figs. 4(4a)–(4f)) in two families — learnable point coordinates (4a–4c) and learnable masks (4d–4f) — and analyze both below; Fig. 4(1) reports their scores.

Dataset	Semantic labels	Prompts	Method	Spl.	R.Kd	L.Kd	GB	Eso.	Liver	Stom.	Aorta	IVC	Panc.	RAG	LAG	Duo.	Blad.	Pros.	Avg.
AMOS-CT (task 1)	✔	–	UNETR Hatamizadeh et al. (2022)	0.928	0.913	0.903	0.719	0.763	0.955	0.849	0.922	0.838	0.766	0.663	0.663	0.662	0.815	0.744	0.807
			nnFormer Zhou et al. (2021)	0.950	0.948	0.944	0.789	0.784	0.967	0.914	0.931	0.868	0.828	0.654	0.695	0.759	0.865	0.773	0.845
			SwinUNETR Hatamizadeh et al. (2021)	0.954	0.954	0.950	0.819	0.852	0.972	0.919	0.955	0.911	0.875	0.775	0.801	0.816	0.895	0.812	0.884
			SwinUNETRv2 He et al. (2023)	0.959	0.962	0.958	0.842	0.867	0.976	0.933	0.957	0.920	0.889	0.783	0.812	0.843	0.913	0.836	0.897
			3D UX-Net Lee et al. (2022)	0.955	0.956	0.953	0.826	0.858	0.972	0.922	0.955	0.915	0.881	0.781	0.809	0.820	0.902	0.823	0.889
			nn-UNet Isensee et al. (2019)	0.951	0.961	0.956	0.826	0.869	0.973	0.931	0.957	0.923	0.880	0.784	0.809	0.846	0.898	0.827	0.893
	✗	nnUNet	SAM Kirillov et al. (2023)	0.679	0.741	0.640	0.168	0.443	0.773	0.671	0.651	0.554	0.434	0.232	0.324	0.444	0.698	0.602	0.538
	✗	nnUNet	SAM 2 Ravi et al. (2024)	0.784	0.817	0.819	0.664	0.734	0.780	0.697	0.793	0.739	0.536	0.457	0.604	0.563	0.744	0.691	0.695
	✗	nnUNet	MedSAM Ma et al. (2024)	0.714	0.811	0.702	0.193	0.469	0.759	0.725	0.701	0.681	0.434	0.365	0.412	0.462	0.783	0.758	0.600
	✔	No needs	SAMed Zhang & Liu (2023)	0.849	0.857	0.830	0.573	0.733	0.894	0.816	0.855	0.784	0.727	0.622	0.683	0.701	0.844	0.819	0.772
	✔	No needs	SAM3D Bui et al. (2024)	0.796	0.863	0.871	0.428	0.711	0.908	0.833	0.878	0.749	0.699	0.564	0.607	0.635	0.884	0.840	0.751
	✔	No needs	RFMedSAM 2	**0.973**	**0.971**	**0.968**	**0.881**	**0.882**	**0.980**	**0.946**	**0.959**	**0.925**	**0.898**	**0.789**	**0.820**	**0.857**	**0.925**	**0.869**	**0.910**
AMOS-CTMRI (task 2)	✔	–	UNETR Hatamizadeh et al. (2022)	0.925	0.938	0.923	0.737	0.776	0.948	0.854	0.918	0.836	0.803	0.672	0.684	0.723	0.814	0.757	0.821
			nnFormer Zhou et al. (2021)	0.945	0.946	0.942	0.773	0.779	0.971	0.903	0.931	0.856	0.826	0.696	0.726	0.740	0.837	0.741	0.840
			SwinUNETR Hatamizadeh et al. (2021)	0.953	0.963	0.955	0.795	0.828	0.976	0.919	0.949	0.891	0.858	0.733	0.764	0.793	0.888	0.785	0.869
			SwinUNETRv2 He et al. (2023)	0.951	0.960	0.949	0.815	0.838	0.962	0.911	0.940	0.887	0.862	0.740	0.777	0.809	0.879	0.801	0.872
			3D UX-Net Lee et al. (2022)	0.948	0.950	0.945	0.814	0.832	0.972	0.915	0.945	0.887	0.869	0.738	0.772	0.802	0.873	0.806	0.870
			nn-UNet Isensee et al. (2019)	0.955	0.961	0.957	0.823	0.841	0.976	0.919	0.952	0.894	0.877	0.744	0.787	**0.819**	0.880	0.816	0.880
	✗	nnUNet	SAM Kirillov et al. (2023)	0.692	0.739	0.628	0.188	0.434	0.786	0.665	0.651	0.547	0.455	0.243	0.323	0.433	0.679	0.595	0.537
	✗	nnUNet	SAM 2 Ravi et al. (2024)	0.879	0.876	0.868	0.729	0.789	0.888	0.784	0.846	0.757	0.644	0.544	0.678	0.618	0.720	0.676	0.754
	✗	nnUNet	MedSAM Ma et al. (2024)	0.709	0.777	0.662	0.190	0.456	0.760	0.711	0.688	0.657	0.444	0.360	0.409	0.441	0.760	0.746	0.582
	✔	No needs	SAMed Zhang & Liu (2023)	0.845	0.849	0.828	0.549	0.714	0.888	0.796	0.827	0.773	0.713	0.604	0.628	0.675	0.833	0.795	0.754
	✔	No needs	SAM3D Bui et al. (2024)	0.788	0.860	0.869	0.415	0.693	0.903	0.812	0.876	0.716	0.682	0.553	0.582	0.618	0.881	0.824	0.738
	✔	No needs	RFMedSAM 2	**0.969**	**0.964**	**0.959**	**0.854**	**0.848**	**0.978**	**0.931**	**0.953**	**0.901**	**0.880**	**0.749**	**0.780**	0.817	**0.889**	**0.834**	**0.886**

Table 2: AMOS-CT and AMOS-CTMRI test results on official leaderboards using 5-fold cross-validation ensembling for each model. "Semantic labels" indicates whether the model outputs semantic classes; "Prompts" indicates the required prompt source (e.g., *BBox* for bounding boxes, or *None* when no prompts are needed).

4.2.1 Learnable Point Coordinate Representations

A straightforward approach is to learn point coordinates directly. The architecture in Fig. 4(4a) initializes object queries for each class and processes them via self- and cross-attention mechanisms interacting with image features. Multiple MLP layers adjust embedding dimensions to predict box coordinates and object scores. Unlike SAM, SAM 2 enforces stricter prompt labeling, classifying points as no object (-1), negative/positive points $(0/1)$, or box prompts $(2/3)$. Prior experiments with GT prompts included indicators for absent objects; here, object scores are trained to decide whether a frame should contain a prompt.

Despite these efforts, directly learning point coordinates proved challenging. The model in Fig. 4(4a) achieved only 77.35% DSC, revealing a large performance gap. Incorporating object scores from the mask decoder (Fig. 4(4b)) improved accuracy by 1.9%, but results remained suboptimal. A variant using a learnable point embedding block (Fig. 4(4c)), where coordinate and label representations are jointly learned, caused an 11% drop in DSC, showing that learning precise coordinates from scratch is unreliable. This limitation arises from the absence of explicit spatial context in image embeddings. Randomly initialized embeddings lack coordinate encoding, making spatial alignment difficult. Moreover, bounding boxes alone fail to convey sufficient semantic information for robust multi-class segmentation. These observations led us to shift toward mask-based prompting strategies.

4.2.2 Learnable Masks

Instead of predicting coordinates, we explored generating segmentation masks first and deriving bounding boxes from them. The architecture shown in Fig. 4(4d) employs a hierarchical design integrating convolutional layers with multi-scale features from the image encoder. It begins from coarse, low-resolution features and progressively refines them through convolutional upsampling, integrating higher-resolution details at each stage. The generated masks are supervised with auxiliary losses against GT labels, achieving a DSC of 84.93%, a clear improvement over coordinate-based methods. However, this design introduced optimization challenges. The auxiliary mask losses sometimes conflicted with SAM 2's segmentation losses, complicating joint training. Additionally, architectural disparity between the prompt generator and SAM 2 caused synchronization issues and hindered convergence.

To resolve these conflicts, we introduced an independent U-Net-based prompt generator (Fig. 4(4e)) operating alongside SAM 2. This separation ensured that SAM 2's parameters remained unaffected during prompt learning. The U-Net-generated masks were converted into bounding boxes and used as input prompts for SAM 2, improving DSC to 85.38%. Finally, to further enhance interaction between the U-Net and SAM 2, we employed a multi-stage refinement process(Fig. 4(4f)). Generated masks and bounding boxes were first passed to Step 1 of SAM 2 to obtain initial segmentations, then refined in Step 2 using memory attention. This dual-stage pipeline achieved a final DSC of 86.65%, validating the effectiveness of the proposed prompt generation framework.

Semantic labels	Prompts	Method	Spl.	R.Kd	L.Kd	GB	Eso.	Liv.	Stom.	Aorta	IVC	Veins	Panc.	AG	DSC
✔	–	TransUNet Chen et al. (2021)	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
		3D UX-Net Lee et al. (2022)	0.946	0.942	0.943	0.593	0.722	0.964	0.734	0.872	0.849	0.722	0.809	0.671	0.814
		UNETR Hatamizadeh et al. (2022)	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 Hatamizadeh et al. (2021)	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 Isensee et al. (2019)	0.949	0.932	0.933	0.514	0.797	0.967	0.852	0.880	0.856	0.777	0.769	0.673	0.825
		nnFormer Zhou et al. (2021)	0.935	0.949	0.950	0.641	0.795	0.968	0.901	0.897	0.859	0.778	0.856	0.739	0.856
✗	GT	SAM Kirillov et al. (2023)	0.933	0.922	0.927	0.805	0.831	0.899	0.808	0.890	0.894	0.492	0.728	0.708	0.819
✗	GT	SAM 2 Ravi et al. (2024)	0.946	0.923	0.924	0.859	0.888	0.928	0.893	0.852	0.884	0.434	0.694	0.705	0.828
✗	GT	MedSAM Ma et al. (2024)	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
✗	GT	SAM-U Deng et al. (2023)	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
✗	GT	SAM-Med2D Cheng et al. (2023)	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
✗	GT	RFMedSAM 2	0.961	0.943	0.945	0.909	0.918	0.965	0.945	0.954	0.942	0.968	0.883	0.843	0.923
✔	No Needs	SAMed_h Zhang & Liu (2023)	0.953	0.921	0.929	0.621	0.753	0.964	0.902	0.876	0.798	0.742	0.779	0.610	0.821
✔	No Needs	SAM3D Bui et al. (2024)	0.933	0.901	0.909	0.601	0.733	0.944	0.882	0.856	0.778	0.722	0.759	0.590	0.801
✔	No Needs	RFMedSAM 2	0.969	0.954	0.955	0.661	0.790	0.974	0.925	0.919	0.881	0.798	0.861	0.730	0.868

Table 3: Comparison of RFMedSAM 2 with state-of-the-art methods on the BTCV dataset. "Semantic labels" indicate the model's ability to infer labels, while "Prompts" specify the source of the prompt.

4.3 QREFINE-FORMER

In the previous subsection, we employed a U-Net-based prompt generator to produce prompts for SAM 2. Beyond prompt generation, we observe that U-Net, as a strong medical image segmentation expert, learns rich domain-specific representations that are highly effective for medical images. In contrast, SAM 2 is a general-purpose foundation model with excellent generalization capability but limited medical-specific priors. This motivates us to bridge these complementary strengths by transferring informative medical representations from U-Net to SAM 2.

To this end, we propose QRefine-Former, a lightweight two-layer Transformer, where each layer consists of a self-attention block, a cross-attention block, and an MLP. As illustrated in Fig. 4(5a), we first initialize a set of learnable query tokens to retrieve informative features from the U-Net representations through QRefine-Former. The extracted query features are then fused with the SAM 2 image embeddings, providing medical priors for subsequent segmentation. On BTCV, this design improves the Dice score by 0.05%.

Although effective, the learnable queries are independent of the input image and therefore cannot adapt to image-specific characteristics. To address this limitation, we further propose the data-driven variant shown in Fig. 4(5b), where the SAM 2 image embeddings themselves are used as query tokens to retrieve relevant information from U-Net. This adaptive query mechanism better aligns the transferred medical knowledge with the current image, leading to a further 0.09% improvement on BTCV. We adopt this design as the final QRefine-Former in RFMedSAM 2.

To verify that the improvement generalizes beyond BTCV, we further evaluate QRefine-Former on four datasets. As shown in Tab. 5 bottom, introducing QRefine-Former consistently improves or maintains the segmentation accuracy across all datasets, increasing the Dice score from 93.2% to 93.6% on ACDC, from 86.7% to 86.8% on BTCV, from 90.7% to 91.0% on AMOS-CT, from 88.4% to 88.6% on AMOS-CTMRI. These results demonstrate that the proposed feature refinement module generalizes well across different medical imaging tasks. Despite its effectiveness, QRefine-Former introduces only 4.5M trainable parameters. As shown in Tab. 9 (Appendix), together with the 46M-parameter U-Net and 19M adapters, our method trains only 69.5M parameters (23.6% of the total 293.9M), while the 224.4M SAM 2 backbone remains entirely frozen. This lightweight design enables efficient adaptation with state-of-the-art performance.

5 EXPERIMENTAL EVALUATION

We evaluate on three public datasets: AMOS22 Ji et al. (2022), BTCV Landman et al. (2015), and ACDC Bernard et al. (2018). AMOS22 has two tasks: CT (Task 1) with 300 training and 200 test CT scans, and CT+MRI (Task 2) adding 60 and 40 MRI scans for training and test, both covering 16 abdominal organs. BTCV contains 30 abdominal CT scans (24 train / 6 val) with 13 organs. ACDC includes 200 training and 100 test cardiac MRI scans with RV, MYO, and LV annotations. AMOS22 and ACDC both uses 5-fold cross-validation. Performance is measured by Dice Similarity Coefficient (DSC) on all datasets.

In Tables 2 and 3, the column "Semantic labels" indicates whether a model predicts semantic classes; "Prompts" specifies the prompt source. Because SAM, MedSAM, and related prompt-driven

Semantic labels	Prompts	Method	RV ↑	Myo ↑	LV ↑	DSC ↑
✔	–	UNETR Hatamizadeh et al. (2022)	88.51	88.35	92.78	89.88
		SwinUNETR Hatamizadeh et al. (2021)	89.97	89.07	93.12	90.72
		3D UX-Net Lee et al. (2022)	89.13	88.53	92.94	90.20
		nnUNet Isensee et al. (2019)	90.59	89.52	93.86	91.32
		nnFormer Zhou et al. (2021)	92.09	90.74	94.40	92.41
✗	GT	SAM Kirillov et al. (2023) bbox	74.31	40.32	80.41	65.01
✗	GT	SAM 2 Ma et al. (2024) bbox	90.83	79.16	82.73	84.24
✗	GT	MedSAM Ma et al. (2024) bbox	86.58	79.83	90.60	85.67
✔	No needs	SAMed Zhang & Liu (2023)	89.57	88.27	94.22	90.69
✔	No needs	SAM3D Bui et al. (2024)	89.14	86.51	92.95	89.53
✔	No needs	RFMedSAM 2 (Ours)	**93.05**	**91.96**	**95.67**	**93.56**

Table 4: Comparison of RFMedSAM 2 with SOTA methods on ACDC (DSC in %). For fairness, all results are based on 5-fold cross-validation without any ensembles. The best are in **bold**.

models do not infer semantics and depend on external prompts, we supply either ground-truth or nnU-Net–predicted masks to generate prompts, and use the corresponding labels as semantics.

5.1 Comparison with State-of-the-Art Methods

Across three benchmarks spanning abdominal CT, multimodal CT/MRI, and cardiac MRI, RFMedSAM 2 consistently establishes new state-of-the-art performance over strong CNN-, transformer-, and SAM-based methods. The consistent improvements across diverse anatomical structures and imaging modalities demonstrate the effectiveness of our prompt-free adaptation of SAM 2 for robust 3D medical image segmentation.

AMOS22. Tab. 2 reports quantitative results on AMOS22 Task 1 (CT) and Task 2 (CT+MRI). RFMedSAM 2 achieves the best average DSC of **91.0%** on Task 1 and **88.6%** on Task 2, outperforming nnU-Net by +1.7 and +0.6 points, respectively, while achieving the best performance on nearly all anatomical structures. In contrast, prompt-dependent SAM, SAM 2, and MedSAM exhibit substantial performance degradation when relying on nnU-Net-generated bounding-box prompts, whereas RFMedSAM 2 establishes a new state-of-the-art without requiring external prompts.

BTCV. As shown in Tab. 3, in prompt-based settings, RFMedSAM 2 achieves the best DSC of **92.3%**, outperforming the previous best result by +5.0 points and surpassing the strongest prompt-based SAM baseline (SAM-Med2D) by +7.6 points. In prompt-free settings, RFMedSAM 2 achieves the excellent DSC of **86.8%**, outperforming SAMed and SAM3D by +4.0 and +6.0 points, respectively, as well as nnU-Net and nnFormer by +4.3 and +1.2 points. Qualitative comparisons in Fig. 9 in Appendix highlight more accurate delineation of organs such as the stomach, spleen, and liver, demonstrating the robustness and precision of our method.

ACDC. Tab. 4 shows that RFMedSAM 2 achieves the best overall performance, outperforming nnU-Net, nnFormer, SwinUNETR, and SAM3D by +2.2, +1.1, +2.8, and +4.0 points, respectively. It consistently improves segmentation of all three cardiac structures (RV, MYO, and LV), demonstrating robust generalization for cardiac MRI segmentation.

5.2 Discussion

Refinement. Tab. 6 summarizes the experimental results for output predictions at different stages on the BTCV and AMOS datasets. The results show a consistent and progressive improvement in segmentation accuracy, beginning with the initial prediction at Step 0 (U-Net), followed by the first refinement at Step 1 (SAM 2), and reaching its highest performance after the second refinement at Step 2 (SAM 2). Fig. 5 visualizes this progression, demonstrating how segmentation gaps are gradually closed across refinement stages. These results validate the effectiveness of the multi-stage refinement design, which enables RFMedSAM 2 to iteratively improve mask quality and precision.

SAM 2 as a teacher. A standalone U-Net (Tab. 8) trained independently, corresponding to Step 0, reaches 82.5% DSC on BTCV, whereas the same backbone trained jointly with SAM 2 already achieves 86.3% at Step 0. Gradients from SAM 2's segmentation losses, propagated through the QRefine-Former and learnable masks, reshape the U-Net's representations during training. As a result, SAM 2 acts as a teacher rather than a post-hoc corrector, effectively distilling foundation-model knowledge into a compact volumetric expert.

	train with prompts	learnable bboxes	learnable masks
w/ obj_score	0.923	0.792	0.847
w/o obj_score	0.920	0.628	0.867

	ACDC	BTCV	AMOS-CT	AMOS-CTMRI
RFMedSAM 2 w/o QRefine-Former	0.922	0.867	0.907	0.884
RFMedSAM 2 w/ QRefine-Former	0.936	0.868	0.910	0.886

Table 5: The effect of object score (top), and QRefine-Former (bottom).

Dataset	Step 0 - UNet	Step 1 - SAM	Step 2 - SAM
ACDC	0.933	0.935	0.936
BTCV	0.863	0.867	0.868
AMOS-CT	0.901	0.904	0.910
AMOS-CTMRI	0.884	0.886	0.886

Table 6: Performance of output predictions across different steps. Two refinements.

	(2, 1024, 1024)	(8, 512, 512)	(32, 256, 256)
DSC	0.751	0.827	0.867

Table 7: Performance comparison of different patch sizes on the BTCV dataset.

	3D UNet	2D UNet	2D UNet + Attention	3D UNet + Attention
DSC	0.825	0.807	0.805	0.815

Table 8: Performance comparison of different UNet models on the BTCV dataset.

Object scores. Tab. 5 (top) tests object-score prediction under three prompt regimes: fine-tuning SAM 2 with external prompts, using learnable bounding boxes as prompt generators, and using learnable masks as prompt generators. Each configuration is tested with and without object score prediction. The results reveal several insights. First, learning object scores alongside explicit prompts does not yield significant performance improvements. This is expected, as the presence of a prompt already implies that the target object exists in the corresponding frame. Second, models employing learnable bounding boxes benefit from object score learning, since bounding box predictions are less spatially accurate and thus profit from the additional supervisory signal. Third, models using learnable masks experience a decrease in performance when object scores are included. This likely occurs because predicted masks already encode probabilistic spatial distributions, whereas object scores introduce an additional single probability term that may interfere with the richer mask-level information.

Patch size and UNet variants. We also analyze the influence of input patch size and U-Net architecture on segmentation performance. Tab. 7 presents results with a fixed pixel budget, deeper patches help monotonically (2×1024^2: 75.1%; 8×512^2: 82.7%; 32×256^2: 86.7% DSC), so volumetric context matters more here than in-plane resolution. Tab. 8 compares multiple U-Net architectures, showing that 3D U-Net (82.5%) beats 2D U-Net (80.7%). Incorporating attention modules within the bottleneck does not lead to further gains, likely because conventional medical segmentation tasks already benefit from strong spatial priors and inductive biases that reduce the need for additional attention mechanisms.

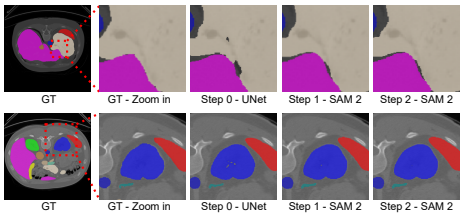


Figure 5: Comparison of Step 0, 1, and 2.

Overall, these findings highlight the complementary effects of refinement, prompt generation, and architectural design in RFMedSAM 2. The progressive refinement process improves segmentation quality, the teacher-student synergy between SAM 2 and U-Net enhances representation learning, and the volumetric U-Net backbone ensures robust feature extraction across depth dimensions. Together, these components enable RFMedSAM 2 to achieve consistent and generalizable state-of-the-art performance across diverse medical segmentation benchmarks.

6 Conclusion

In this paper, we presented RFMedSAM 2, a refinement-based framework that adapts SAM 2 for volumetric medical image segmentation through architectural adaptations, automatic prompt generation, and query-guided medical knowledge transfer.

First, we showed that, under ground-truth prompts, the adapted SAM 2 achieves 92.3% DSC on BTCV, outperforming nnU-Net by 4.3%. Second, we proposed a U-Net-based prompt generator with dual-stage SAM 2 refinement to eliminate the need for manual prompts. Finally, we introduced QRefine-Former to transfer medical priors from U-Net to SAM 2, improving segmentation accuracy while preserving the foundation model's strong generalization capability.

Extensive experiments on AMOS22, ACDC, and BTCV demonstrate that RFMedSAM 2 achieves state-of-the-art performance across multiple benchmarks. In future work, we will extend the framework to additional imaging modalities, such as ultrasound, and explore real-time clinical deployment.

Reproducibility Statement

All datasets used in this work (AMOS22, BTCV, ACDC) are publicly available. Training and inference hyper-parameters, the deep-supervision objective, the data augmentation pipeline and the hardware configuration are given in Appendix H; the architecture of every modified SAM 2 module is specified in Appendix D. Code will be released upon acceptance.

Use of Large Language Models

Large language models were used only for language editing of author-written text. They were not used for research idea, experimental design, implementation, data analysis, or the generation of any reported result. The authors take full responsibility for all content in this paper.

References

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

Nhat-Tan Bui, Dinh-Hieu Hoang, Minh-Triet Tran, and Ngan Le. Sam3d: Segment anything model in volumetric medical images. *arXiv preprint arXiv:2309.03493*, 2023.

Nhat-Tan Bui, Dinh-Hieu Hoang, Minh-Triet Tran, Gianfranco Doretto, Donald Adjeroh, Brijesh Patel, Arabinda Choudhary, and Ngan Le. Sam3d: Segment anything model in volumetric medical images. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–4. IEEE, 2024.

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. *arXiv preprint arXiv:2105.05537*, 2021.

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.

Tianrun Chen, Ankang Lu, Lanyun Zhu, Chaotao Ding, Chunan Yu, Deyi Ji, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam2-adapter: Evaluating & adapting segment anything 2 in downstream tasks: Camouflage, shadow, medical image segmentation, and more. *arXiv preprint arXiv:2408.04579*, 2024.

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.

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. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 368–377. Springer, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation. *arXiv preprint arXiv:2306.13465*, 2023.

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*, pp. 272–284. Springer, 2021.

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*, pp. 574–584, 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Yufan He, Vishwesh Nath, Dong Yang, Yucheng Tang, Andriy Myronenko, and Daguang Xu. Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 416–426. Springer, 2023.

Ziyan Huang, Haoyu Wang, Zhongying Deng, Jin Ye, Yanzhou Su, Hui Sun, Junjun He, Yun Gu, Lixu Gu, Shaoting Zhang, et al. Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. *arXiv preprint arXiv:2304.06716*, 2023.

Fabian Isensee, Paul F Jäger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. Automated design of deep learning methods for biomedical image segmentation. *arXiv preprint arXiv:1904.08128*, 2019.

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.

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.

B Landman, Z Xu, J Eugenio Igelsias, M Styner, T Langerak, and A Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI: Multi-Atlas Labeling Beyond Cranial Vault-Workshop Challenge*, 2015.

Ho Hin Lee, Shunxing Bao, Yuankai Huo, and Bennett A Landman. 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. *arXiv preprint arXiv:2209.15076*, 2022.

Chengyin Li, Prashant Khanduri, Yao Qiang, Rafi Ibn Sultan, Indrin Chetty, and Dongxiao Zhu. Auto-prompting sam for mobile friendly 3d medical image segmentation. *arXiv preprint arXiv:2308.14936*, 2023.

Haofeng Liu, Erli Zhang, Junde Wu, Mingxuan Hong, and Yueming Jin. Surgical sam 2: Real-time segment anything in surgical video by efficient frame pruning. *arXiv preprint arXiv:2408.07931*, 2024.

Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023.

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

Mobina Mansoori, Sajjad Shahabodini, Jamshid Abouei, Konstantinos N Plataniotis, and Arash Mohammadi. Polyp sam 2: Advancing zero shot polyp segmentation in colorectal cancer detection. *arXiv preprint arXiv:2408.05892*, 2024.

OpenAI. GPT-4 technical report, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F Jaeger, and Klaus H Maier-Hein. Mednext: transformer-driven scaling of convnets for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 405–415. Springer, 2023.

Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International Conference on Machine Learning*, pp. 29441–29454. PMLR, 2023.

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.

Bin Xie, Hao Tang, Bin Duan, Dawen Cai, and Yan Yan. Masksam: Towards auto-prompt sam with mask classification for medical image segmentation. *arXiv preprint arXiv:2403.14103*, 2024.

Bin Xie, Hao Tang, Dawen Cai, Yan Yan, and Gady Agam. Self-prompt sam: Medical image segmentation via automatic prompt sam adaptation. *arXiv preprint arXiv:2502.00630*, 2025.

Jieming Yu, An Wang, Wenzhen Dong, Mengya Xu, Mobarakol Islam, Jie Wang, Long Bai, and Hongliang Ren. Sam 2 in robotic surgery: An empirical evaluation for robustness and generalization in surgical video segmentation. *arXiv preprint arXiv:2408.04593*, 2024.

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

Yichi Zhang, Zhenrong Shen, and Rushi Jiao. Segment anything model for medical image segmentation: Current applications and future directions. *Computers in Biology and Medicine*, pp. 108238, 2024.

Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021.

A Intrinsic issues of SAM2

Fig. 2 illustrates the whole pipeline of SAM 2, highlighting several intrinsic issues for medical image segmentation.

i) **Omission to predict the first few frames:** The first frame in two objects is the second frame, therefore, SAM 2 begins processing from the second frame, disregarding the first frame, even though it contains both objects.

ii) **Empty prompt affecting object prediction:** When no prompt is provided for an object, but the object is still present, the empty prompt restricts prediction for that object. For instance, at frames $z = 1$ and $z = 4$, the purple and red objects, respectively, are omitted from the predictions.

iii) **Confusion of temporal positions:** All prompted frames are assigned a temporal position of 0. While this approach increases attention to the prompted frames, it loses the relative temporal positioning of all the prompt frames. Moreover, since SAM 2 skips over prompted frames, the relative temporal positions of the unprompted frames are distorted. For example, the real relative temporal position of the frame $z = 3$ with respect to the current frame $z = 6$ should be 3, but due to the prompted frame at $z = 4$, the relative temporal position is incorrectly assigned as 2.

B Potentials to force Step 2 for all frames.

When we provide prompts at each frame for each class, SAM 2 does not process Step 2 and does not leverage the capabilities of Memory Attention, which can build relations with previous frames and prompted frames. To explore this functionality, we force Step 2 for all frames after processing Step 1. Although the results decreased slightly from 82.77% to 81.17% Dice, we find a potential

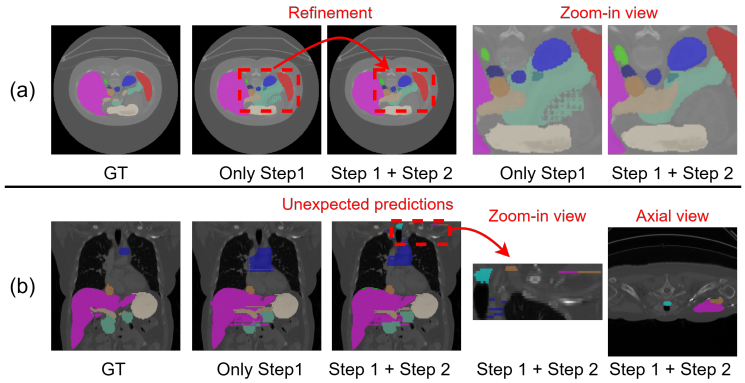


Figure 6: Benefits for refinement by Step 2.

refinement benefit illustrated in Fig. 6(a). Through Step 2, the green area is refined and becomes more accurate, demonstrating the significant potential of the refinement process. As a result, we plan to incorporate this approach into our method. However, the refinement introduced by Step 2 also has some drawbacks. In Fig. 6(b), we show orthogonal planes in relation to the axial plane (a sequence of the axial plane images is fed to SAM 2). The top portion of Fig. 6(b) presents unexpected predictions. Since SAM 2 assigns a temporal position of 0 to the prompted frames, which are always involved in memory attention, the incorrect relative temporal positioning leads to these unexpected and incorrect predictions. We will address this issue in the next section.

C Motivation Behind the Designed Adapters.

Since the image encoder, the memory attention, and the mask decoder contain attention blocks for image embedding, which includes significant spatial information. Therefore, we design the depth-wise convolutional adaption (DWConvAdapter) illustrated in Fig. 4(3b) to learn spatial information. After using DWConvAdapters for the attention blocks with image embedding, the performance increases by 0.47%. The motivation behind the DWConvAdapter design is to extend the original adapter by incorporating a depth-wise convolution layer, followed by layer normalization and a

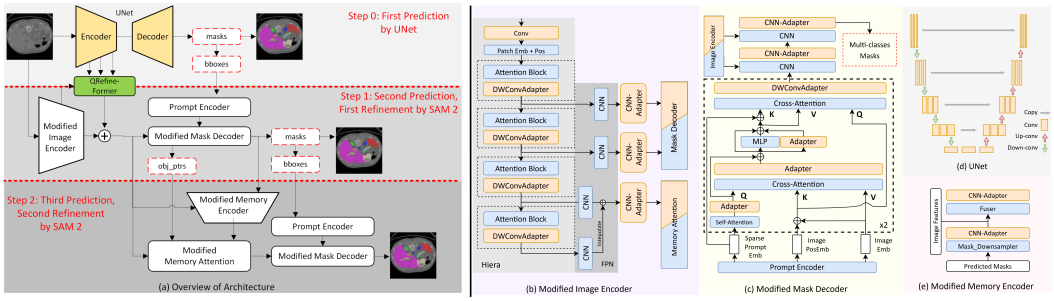


Figure 7: Details of the whole architecture of RFMedSAM 2.

GeLU activation function, to effectively learn spatial information. A parallel skip connection is included to preserve the original structure. In the worst case, where the depth-wise layer learns nothing (*i.e.,* its output is zero), the skip connection ensures that all original information is retained. Building on this concept, we designed the CNN-Adapter for adapting convolutional layers since more convolutional layers are involved at SAM 2 compared to SAM. The CNN-Adapter uses a point-wise convolutional layer to downsample the channel dimension, reducing complexity, followed by a depth-wise convolutional layer to capture spatial dimensions. Finally, a point-wise convolutional layer recovers the channel dimension to its original size. Inspired by ConvNext, we use only layer normalization and a GeLU activation function in this block. The bottleneck structure helps reduce complexity, and a parallel skip connection ensures that the output from the convolutional layers in SAM 2 is preserved. In the worst case, where the depth-wise layer learns nothing (*i.e.,* its output is zero), the skip connection still retains all relevant information.

D ARCHITECTURE OF RFMEDSAM 2

RFMedSAM 2 is a refined adaptation of SAM 2 tailored to improve segmentation performance in medical imaging. Fig. 1 depicts the overall system, comprising three sequential stages: an initial prediction stage, a preliminary segmentation stage, and a refinement stage.

In the initial prediction stage, a U-Net serves as a medical image segmentation expert to generate coarse multi-class masks, which are converted into bounding boxes as prompts for subsequent stages. Meanwhile, the proposed QRefine-Former bridges the medical expert (U-Net) and the general-purpose foundation model (SAM 2) by extracting informative query features from U-Net representations and injecting them into the SAM 2 image embeddings. This enables SAM 2 to exploit both its strong generalization capability and task-specific medical priors.

The preliminary segmentation stage integrates a modified SAM 2 to perform a second prediction. The image encoder extracts image embeddings from the input image, while the prompt encoder converts the bounding boxes into prompt embeddings. The extracted query features from QRefine-Former are fused with the image embeddings before the mask decoder, enabling SAM 2 to leverage both general visual representations and medical-specific prior knowledge. The mask decoder then predicts segmentation masks and object pointers, which are then converted into updated bounding-box prompts. A modified memory encoder further processes these masks together with frame features to produce memory features.

The final refinement stage enhances accuracy via a modified memory-attention module, which combines memory features from previous frames with the current-frame image features enhanced by QRefine-Former query features. The mask decoder consumes these integrated features, together with new prompt embeddings from the prompt encoder, to produce the final refined segmentation.

To efficiently adapt SAM 2 for medical image segmentation, RFMedSAM 2 incorporates several architectural modifications. Two convolutional layers are employed to transform multi-channel medical images into the RGB input space required by SAM 2. Lightweight DWConvAdapters and CNN-Adapters are inserted into the image encoder, mask decoder, memory encoder, and memory-attention modules for parameter-efficient fine-tuning. Combined with the U-Net-based prompt generator, QRefine-Former, and the proposed memory-attention strategy, these components form the complete RFMedSAM 2 framework. Additional implementation details are provided in the Appendix.

Fig. 7(b)-(e) illustrates each component of RFMedSAM 2, described as follows.

D.1 Modified Image Encoder

Fig. 7(b) illustrates the redesigned image encoder. i) SAM works on natural images that have 3 channels for RGB while medical images have varied modalities as channels. There are gaps between the varied modalities of medical images and the RGB channels of natural images. Therefore, we design a sequence of two stacked convolutional layers to an invert-bottleneck architecture to learn the adaption from the varied modalities with any size to 3 channels. ii) SAM 2 employs Hiera Ryali et al. (2023) that is hierarchical with multiscale output features as its image encoder backbone and a FPN module. Hiera consists of four stages with different feature resolutions and every stage contains various number of attention blocks. We insert our designed DWConvAdapter blocks into each attention block in Hiera. The output of each stage will be connected with one convolution in the FPN module. The latest output feature is up-sampled and summed with the second latest output feature as the image embedding. The third and fourth latest output feature are as skip connections to to incorporate high-resolution embeddings for the mask decoding. To adapt these convolutional layers, we insert our designed CNN-Adapters for the output features from the FPN module.

D.2 Modified Mask Encoder

Fig. 7(c) illustrates the redesigned mask encoder. The mask encoder contains two subsequent transformers and two following convolutional layers. i) Each transformer first applies self-attention to the prompt embedding. We insert an adapter behind the self-attention. Then, a cross-attention block is adopted for tokens attending to image embedding. We insert an adapter behind the cross-attention. Next, we insert a adapter parallel to an MLP block. Finally, a cross-attention block is utilized for image embedding attending to tokens. We insert a DWConvAdapter behind the cross-attention. In this way, our model can learn the spatial information for the image embedding and adapt information for the prompt embedding. ii) We inserted a CNN-Adapter behind the two following convolution layers to adapt convolutional layers from natural images to medical images.

D.3 UNet, Modified Memory Encoder and Modified Memory Attention.

Fig. 7(d) and (e) illustrate the UNet and the redesigned memory encoder, respectively. i) UNet is designed with a symmetrical encoder-decoder structure with skip connections. The encoder consists of several stages, each formed by a sequence of convolutional layers followed by down-sampling layers, progressively increasing the number of channels while reducing the spatial resolution to capture different deep-level features. The decoder upsamples the feature maps using transposed convolutions to restore spatial resolution and refine predictions. Skip connections between corresponding encoder and decoder layers enable the network to retain fine-grained spatial details, enhancing localization accuracy. ii) The memory encoder comprises two modules: the mask downsampler, which processes predicted masks, and the fuser, which integrates image features and mask features. To adapt these CNN-based modules to medical images, a CNN-Adapter is inserted after each module. iii) The memory attention module stacks several transformer blocks, the first one taking the image encoding from the current frame as input. Each block performs self-attention, followed by cross-attention to memory features. Therefore, we inserted our designed DWConvAdapter blocks into each attention block since the transformer blocks process the image embedding with the spatial dimension.

E Impact of Auxiliary Losses on Image Encoder Parameter Updates if Prompt Generator Built with Image Encoder

Fig. 4(4d) illustrates a hierarchical structure with convolutional layers combined with multi-level features from the image encoder. The features with a lower resolution gradually increase the resolution by convolution layers and then combined with higher resolution features. Auxiliary loss functions are employed to supervise between the predicted masks and the ground truth. Although this approach achieves a DSC of 84.93%, the result is not competitive. During training, both the auxiliary losses from the generated masks and the final output losses from SAM 2 influence the update of the image encoder parameters, which constitute a significant portion of the model. However, these two types of

losses, due to their distinct architectural differences, are challenging to optimize simultaneously and achieve a balanced update for the image encoder parameters.



Figure 8: Oscillated losses if prompt generator built with image encoder.

We conduct experiments to validate the insights presented in Fig. 8. The training process is divided into two phases: one phase updates the parameters based solely on the auxiliary losses supervised by the auxiliary loss function, while the other phase updates all parameters based on both the auxiliary loss function and the final output loss function. The results indicate that after the second phase begins, the validation loss oscillates and is in an unstable state shown in the red line. The dice of the auxiliary masks present an unstable state since the final output losses affect the update of the image encoder and then affect the accuracy of the auxiliary masks.

In conclusion, using a prompt generator built with the image encoder creates a challenge in balancing the update of the image encoder's parameters. As a result, we abandon this approach and instead employ an independent U-Net to generate masks and subsequently produce the corresponding bounding boxes.

F Analysis of parameters

Method	UNet	Adapters	QRefine-Former	Total Trainable Params	SAM 2	Total Params
RFMedSAM 2	46M	19M	4.5M	46M+19M+4.5M=69.5M	224M	69.5M+224.4M=293.9M

Table 9: The parameters of each components for our RFMedSAM 2.

Tab. 9 illustrates the number of parameters for each components. Our method is based on the SAM 2 large model, which contains 224.4M parameters in total. These parameters are entirely frozen during training. On top of SAM 2, we introduce two trainable components: i) A UNet (step 0) contains 46M parameters. ii) Our designed adapters, inserted into SAM 2, contains 19M parameters. iii) Our designed QRefine-Former contains 4.5M parameters. Together, these components result in 69.5M trainable parameters, which is approximately 23.6% of the total model size (69.5M out of 224.4M + 69.5M = 293.9M, shown at the table below). Despite training only a small fraction of the overall parameters, our method achieves efficient adaptation and delivers state-of-the-art performance.

G Qualitative comparison on BTCV

Fig. 9 highlight more accurate delineation of key organs such as the stomach, spleen, and liver, illustrating the robustness and precision of our method.

H Implementation Details

We utilize some data augmentations such as rotation, scaling, Gaussian noise, Gaussian blur, brightness, and contrast adjustment, simulation of low resolution, gamma augmentation, and mirroring. We

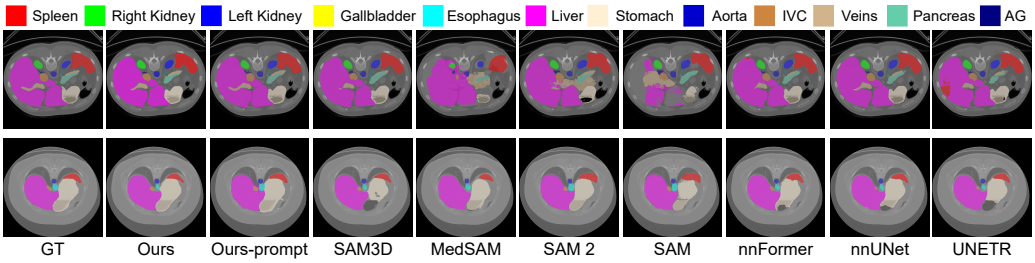


Figure 9: Qualitative comparison on BTCV. RFMedSAM 2 achieves the most precise segmentation.

set the initial learning rate to 0.001 and employ a "poly" decay strategy in Eq. equation 1.

$$lr(e) = init_lr \times (1 - \frac{e}{\text{MAX_EPOCH}})^{0.9}, \tag{1}$$

where e means the number of epochs, MAX_EPOCH means the maximum of epochs, set it to 1000 and each epoch includes 250 iterations. We utilize SGD as our optimizer and set the momentum to 0.99. The weighted decay is set to 3e-5. We utilize both cross-entropy loss and dice loss by simply summing them up as the loss function. We utilize instance normalization as our normalization layer. we employ the deep supervision loss for the supervision of the U-Net. All experiments are conducted using two NVIDIA RTX A6000 GPUs with 48GB memory.

Deep Supervision. The U-Net network is trained with deep supervision. For each deep supervision output, we downsample the ground truth segmentation mask for the loss computation with each deep supervision output. The final training objective is the sum of all resolutions loss:

$$\mathcal{L} = w_1 \cdot \mathcal{L}_1 + w_2 \cdot \mathcal{L}_2 + w_3 \cdot \mathcal{L}_3 + \cdots w_n \cdot \mathcal{L}_n \tag{2}$$

where the weights halve with each decrease in resolution (*i.e.,* $w_2 = \frac{1}{2} \cdot w_1; w_3 = \frac{1}{4} \cdot w_1$, etc), and all weight are normalized to sum to 1. Meanwhile, the resolution of $\mathcal{L}_1$ is equal to $2 \cdot \mathcal{L}_2$ and $4 \cdot \mathcal{L}_3$.