MedSAM++: Automated Multi-Organ Segmentation with Atlas-Guided Prompting, 2.5D Context, and Volumetric Refinement

Nuren Nafisa
Student IDs: g202427580
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia

Supervised by: Dr. Muzammil Behzad
muzammil.behzad@kfupm.edu.sa
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia

***Abstract*—While foundational models like SAM offer powerful segmentation capabilities, their application in real life to volumetric medical imaging is quite challenging due to domain-specific constraints and computational cost. Although existing approaches like MedSAM exhibits good performance, it requires manual prompting and lack volumetric consistency. With the aim to develop a fully automated and computationally efficient framework for multi-organ abdominal CT segmentation, this paper presents an enhanced framework MedSAM++, overcoming the limitations of both volumetric overlap and boundary accuracy. Our method integrates four key components: (a) parameter-efficient fine-tuning via Low-Rank Adaptation (LoRA), (b) a 2.5D input scheme incorporating adjacent slices for volumetric context, (c)a boundary-aware composite loss function combining Dice, Focal, and Laplacian edge terms, and, (d) an atlas-guided automatic prompt generation mechanism. To evaluate this framework, internal FLARE22 test set have been used, upon which it achieved a Dice score of 95.92% and Normalized Surface Dice (NSD) of 79.70%, significantly outperforming all baseline models. In terms of external validation on AMOS 2022 under zero-shot conditions, it exhibited a Dice of 85.98% and NSD of 67.69%, ensuring better generalization. Conclusion: These results establish that this proposed framework, MedSAM++, with the ability to transform foundational models into automated clinical tools, has achieved state-of-the-art segmentation performance. Notably, this model has reduced computational constraints that make it more suitable for deployment in real life scenario in clinical environments.**

***Index Terms*—Medical image segmentation, Abdominal CT, Foundation models, Segment Anything (SAM), MedSAM, Parameter-efficient fine-tuning, LoRA, 2.5D context, Boundary-aware loss, Automatic prompting, Dice coefficient, Normalized Surface Dice (NSD)**

I. Introduction

A. Background and Significance

Medical image segmentation is very important for computer supported diagnosis, volumetric organ analysis and pre-surgical planning. Especially abdominal CT segmentation enables precise delineation of soft tissue structures like the liver, kidneys and spleen which is a must for clinical decision making. Foundation models like the Segment Anything Model (SAM) have transformed computer vision by making promptable and generic segmentation possible throughout various image domains [1]. Though the direct implementation of these models in medical related imaging is still challenging as medical CT scans are grayscale, low-contrast and volumetric in nature which is substantially different from the natural RGB images that are used to train SAM [2]. As a result, standard SAM lacks 3D spatial awareness and needs manual user interaction which limits its scalability in clinical workflows. This is why an automated and domain-adapted segmentation model is much needed that incorporates volumetric context and boundary precision.

B. Challenges in Current Techniques

Deep learning has greatly advanced segmentation accuracy yet several challenges are still hindering its practical deployment:

- **Domain Adaptation:** Classical architectures like U-Net and nnU-Net are very effective. But they are task-specific and ask for retraining for each dataset [3], [4].
- **3D Representation:** Slice-wise inference in most CNNs and SAM-based models neglects inter-slice continuity which results in inconsistent volumetric predictions.
- **Computational Cost:** Transformer based models like TransUNet demand widespread computational resources for training and fine tuning [5].
- **Manual Interaction:** SAM and its variants depend on user-defined prompts for every slice and prevents large-scale automation in 3D segmentation tasks.

These issues show a need for a resource-efficient, context-aware segmentation framework which is capable of fully automated operation on volumetric data.

C. Problem Statement

The goal of this research is to achieve accurate, automated multi-organ segmentation in abdominal CT scans and overcome the following four critical limitations of current foundation models:

1) **The Fine-Tuning Bottleneck:** Training massive Vision Transformers (ViT-B with over 90M parameters) again requires high-end GPUs like A100 clusters. Keeping the encoder frozen limits adaptability. Again, full fine-tuning overfits on small medical datasets

2) **Boundary Ambiguity:** Dice and other similar standard losses emphasize on region overlap but fail to capture sharp organ boundaries. It especially happens for small and low-contrast organs like the pancreas.
3) **Human-in-the-Loop Limitation:** MedSAM stays semi-automatic and it requires human input for every slice, which is not scalable for 3D volumes [6].
4) **Lack of 3D Context:** Traditional 2D SAM-based models ignore volumetric relationships between slices and lead to discontinuous organ shapes and flickering across axial planes [7], [8].

In short, the approaches of the current segmentation lack automation, 3D coherence and computational efficiency. The proposed **MedSAM++** addresses these challenges. This is done by integrating Low-Rank Adaptation (LoRA)-based fine-tuning, 2.5D contextual learning, a boundary-aware loss function and an atlas-guided automatic prompting mechanism.

In summary, current segmentation approaches lack automation, 3D coherence, and computational efficiency. The proposed **MedSAM++** addresses these challenges by integrating Low-Rank Adaptation (LoRA)-based fine-tuning, 2.5D contextual learning, a boundary-aware loss function, and an atlas-guided automatic prompting mechanism.

D. Objectives

The proposed work focuses on the following objectives:

- **Automation:** Replace manual prompting with an atlas-guided automatic prompt generator which is done by using anatomical priors and Hounsfield Unit (HU) thresholds.
- **Parameter Efficiency:** Apply Low-Rank Adaptation (LoRA) for the Parameter-Efficient Fine-Tuning (PEFT) and train less than 1% of the parameters.
- **Volumetric Consistency:** Include 2.5D related information by storing attached slices and improving the inter-slice continuity.
- **Boundary Fidelity:** Introduce a Boundary-Aware Combo Loss by combining Dice, Focal and Laplacian boundary elements for the sharper contours.
- **3D Refinement:** Enhance 2D outputs into unified 3D structures by morphological post-processing.

E. Scope of Study

This research targets organs like the liver, kidneys, spleen, pancreas and focuses on automated multi-organ segmentation in abdominal CT scans. The introduced **MedSAM++** improves upon MedSAM by incorporating LoRA-based tuning, 2.5D contextual learning and a boundary-sensitive optimization method all while removing the manual user inputs. The models performance is tested on FLARE22) and external (AMOS 2022) datasets to evaluate both its learning proficiency and ability to generalize across domains. Finally, this work seeks to demonstrate clinically reliable, parameter-efficient 3D segmentation achievable on single-GPU systems.

II. Literature Review

A. Overview of Existing Techniques

Previous developments in medical image segmentation were largely ruled by convolutional neural networks (CNNs) like U-Net and its variants. U-Net introduced an encoder-decoder design with skip connections. This allowed exact biomedical segmentation and has become the base for many later models [3], [9]. Extensions like V-Net took advantage of 3D convolutions for volumetric image segmentation [10] while Attention U-Net improved spatial emphasis using attention gating mechanisms [11]. DeepLabV3+ integrated multi-scale feature extraction through atrous convolution in order to improve context learning [12]. Automated architectures like nnU-Net simplified model design furthermore. It was done by self-configuring parameters for new datasets which achieved consistent state-of-the-art performance [4].

In recent times, transformer-based architectures have become powerful alternatives. TransUNet [5] combined CNN decoders with Vision Transformer (ViT) encoders to model long-range spatial dependencies. UNETR [13] utilized a pure transformer encoder for the 3D segmentation. SegFormer [14] showed efficient hierarchical transformer design for high-resolution segmentation. Benchmark datasets similar to FLARE22 [15] and AMOS [16] have standardized multi-organ abdominal CT evaluation. This enabled consistent comparison between architectures.

B. Related Work

Foundation and promptable segmentation models have changed the field by enabling generalization through various domains. The Segment Anything Model (SAM) [1] introduced a ViT-based model trained on over one billion masks that allows class-agnostic segmentation using prompts. Built on this, MedSAM [6] fine-tuned SAM on 1.5M medical images and demonstrated universal applicability but still depend on manual user inputs. Variants like SAM-Med3D [17] extended SAM for volumetric segmentation. At the same time, Med-SA [18] integrated adapter modules for parameter-efficient domain adaptation which is conceptually similar to LoRA [19]. In parallel, boundary-focused strategies like the Boundary Loss [20] improved contour accuracy which is addressed as one of the major weaknesses in global Dice-based optimization. Studies on interactive segmentation [21] laid the groundwork for modern promptable frameworks by iteratively improving masks based on user input. All together, these advances fill the gap between interactive segmentation and fully automated medical segmentation, which this research extends through the proposed MedSAM++ pipeline.

C. Limitations in Existing Approaches

Instead of having remarkable progress, several gaps still remain in the literature. CNN-based approaches need retraining for every dataset and lack volumetric context. This results in discontinuities across slices. Transformer-based methods like TransUNet and UNETR offer improved contextual awareness but it costs high computational demand. Foundation models

TABLE I: Literature summary of selected works related to proposed work(Sorted Descending by Year).

Author (Year)	Model	Objective	Dataset(s)	Baseline Comparison	Key Architecture	Limitation and Relevance to This Work
Ma et al. (2024) [6]	MedSAM	Universal medical segmentation via SAM	1.5M med. images	Task-specific CNNs	Frozen ViT encoder + tuned decoder; Dice ∼85–90%	2D + prompt dependency; direct baseline enhanced in our work
Wang et al. (2023) [17]	SAM-Med3D	Volumetric promptable segmentation	SA-Med3D-140K	MedSAM, SAM	3D ViT-like SAM; large dataset	Heavy compute; contrasts 2.5D lightweight approach
Wu et al. (2023) [18]	Med-SA (Adapter)	Adapting SAM to medical	BTCV, etc.	SAM, MedSAM	Lightweight adapter layers in SAM	2D slice limitation; validates PEFT (related to LoRA)
Kirillov et al. (2023) [1]	SAM	Promptable general segmentation	SA-1B (1B masks)	RITM, SimpleClick	ViT-H backbone with prompt encoder	Requires prompts; medical tuning needed; foundation for MedSAM
Hatamizadeh et al. (2022) [13]	UNETR	3D medical segmentation	BTCV, MSD	nnU-Net	ViT encoder + 3D CNN decoder; Dice ∼90%	High compute; supports volumetric cue need
Ma et al. (2022) [15]	FLARE22 Dataset	Benchmark for abdominal organ segmentation	AbdomenCT-1K (CT)	Challenge leaders	Dataset resource with 13 organs	Reveals generalization gap; used as internal validation
Ji et al. (2022) [16]	AMOS Dataset	Universal multi-organ segmentation	AMOS (CT/MRI)	nnU-Net, TransUNet	Large-scale multi-modality dataset	Cross-modality domain gap; external validation dataset
Chen et al. (2021) [5]	TransUNet	Multi-organ CT segmentation	Synapse (CT)	U-Net, Attention U-Net	ViT encoder + CNN decoder; +1–4% Dice gain	Large model; transformer benefits vs. CNN
Xie et al. (2021) [14]	SegFormer	Efficient semantic segmentation	ADE20K	SETR, DETR-like	Hierarchical Transformer + MLP decoder; 50.3% mIoU	2D focus; guides efficient design for medical tasks
Isensee et al. (2021) [4]	nnU-Net	Auto-configuring medical segmentation	MSD (10 datasets)	U-Net, V-Net	Self-configuring 2D/3D U-Net variants	Retraining per dataset; strong specialist baseline
Kervadec et al. (2019) [20]	Boundary Loss	Boundary-aware segmentation loss	Multiple med. datasets	Dice, CE losses	Contour distance-based boundary loss	No architectural prompts; motivates our Combo Loss design
Oktay et al. (2018) [11]	Attention U-Net	Organ-focused CT segmentation	CT-82 (Pancreas)	U-Net	Attention gates for salient regions	2D slice-wise; inspires boundary awareness
Chen et al. (2018) [12]	DeepLabV3+	Semantic segmentation	PASCAL VOC, Cityscapes	FCN, DeepLabV3	ASPP + decoder refinement; mIoU ∼89%	Natural-image focus; informs MedSAM decoder design
Milletari et al. (2016) [10]	V-Net	Volumetric prostate MRI segmentation	PROMISE12	2D CNNs	3D FCN with residuals; Dice ∼0.85	High memory footprint; motivates 3D context and 2.5D approach
Ronneberger et al. (2015) [3]	U-Net	Biomedical image segmentation	ISBI Cell Tracking	Sliding-window CNN	Encoder–decoder with skip connections; IoU ∼92%	No volumetric context; retrain per dataset; baseline for MedSAM comparison

like SAM and MedSAM improve generalization but remain semi-automatic and rely on manual prompts and 2D inference. Adapter-based methods mitigate training cost but they cannot guarantee volumetric consistency. Moreover, conventional loss functions do not emphasize boundary precision much and it leads to blurred edges and missed fine structures. Even benchmark datasets like FLARE22 and AMOS reveal domain shifts between scanners and patient populations. These limitations show the need for an automated, parameter-efficient and boundary-aware framework with integrated 2.5D context, motivating the design of the proposed **MedSAM++**.

III. Proposed Methodology

A. Existing Model and Challenges

We take MedSAM as the starting point: a promptable segmenter that adapts the Segment Anything Model (SAM) to medical images by freezing the ViT encoder and fine-tuning a lightweight mask decoder [1], [6]. Practically, three issues limit its clinical utility for abdominal CT: (i) the need for a human-drawn box per slice, (ii) slice-wise inference that ignores continuity along the axial axis and (iii) boundary ambiguity on low-contrast soft-tissue interfaces. Additionally, unfreezing the large ViT encoder to recover domain specificity is computationally costly. These observations inspire a design that is automated, parameter-efficient and boundary-aware and at the same time preserves the benefits of the pretrained foundation model.

B. Proposed Enhancements

MedSAM++ converts the interactive 2D slice-wise baseline into a fully automated 2.5D boundary-aware system. We keep the ViT architecture from SAM but modify the data path, training objective and inference pipeline as follows (ordered by the actual flow in Fig. 1):

1) **2.5D Context (input preprocessing).** Instead of repeating a single slice across RGB channels, each input is a tri-slice stack $\left[I_{i-1}, I_i, I_{i+1}\right]$. It supplies local depth cues without resorting to a heavy 3D backbone. This decreases inter-slice “flicker” and stabilizes organ continuity.

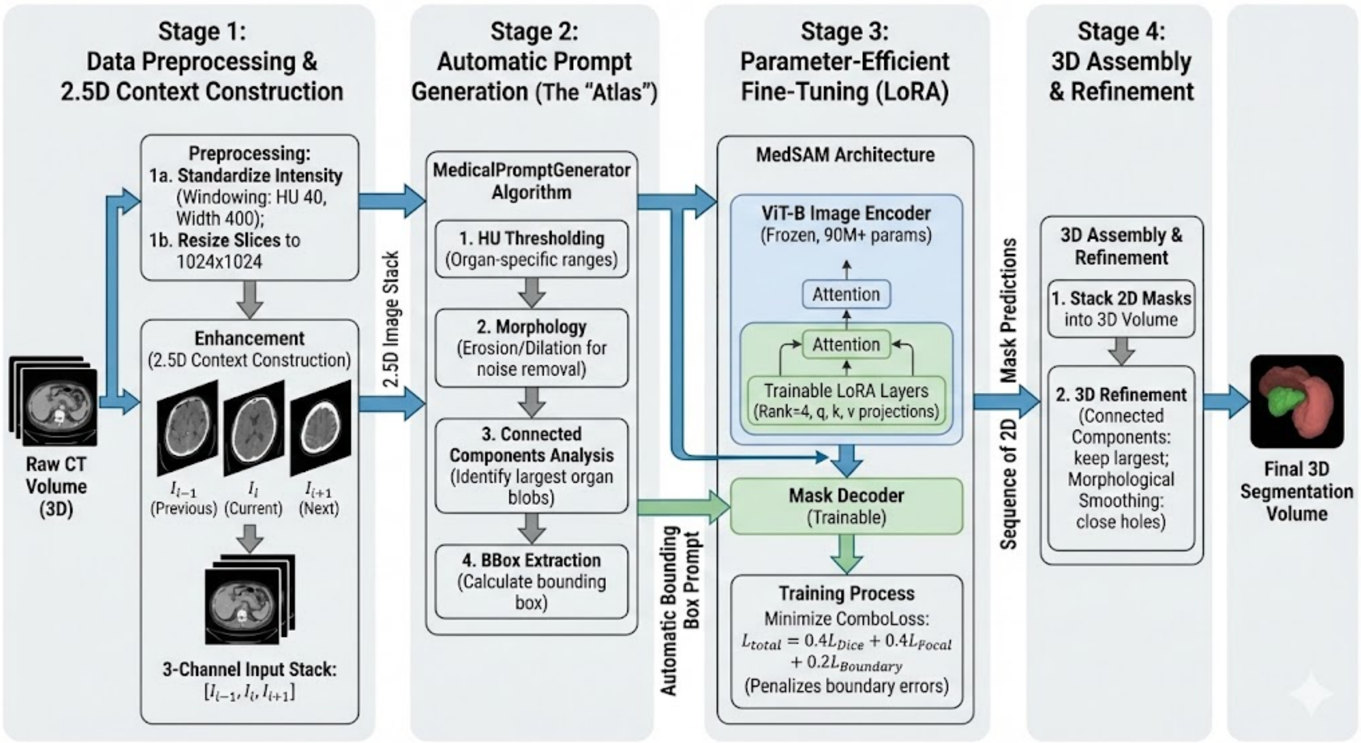


Fig. 1: **MedSAM++ pipeline.** Stage 1: preprocessing and 2.5D context construction; Stage 2: atlas-guided automatic prompt generation; Stage 3: parameter-efficient adaptation with LoRA and boundary-aware training; Stage 4: 3D assembly and refinement to obtain a consistent volumetric mask.

2) **Atlas-Guided Prompt Generator (automation before encoding).** A deterministic module produces bounding boxes from anatomical priors: HU windowing, size/aspect filters, simple morphology and connected-components to retain the largest plausible organ blob. This replaces human clicks with automatically derived prompts.
3) **LoRA for Parameter-Efficient Adaptation (architecture modification).** Rank-r=4 LoRA modules are inserted into the attention projections (q, k, v and output) of the ViT encoder [19]. While specializing to CT texture, SAM's general knowledge is preserved and less than 1% of weights are updated.
4) **Boundary-Aware Combo Loss (training objective).** A composite objective explicitly emphasizes contour fidelity alongside overlap and class imbalance (Sec. III-D).
5) **3D Assembly & Refinement (post-processing).** Slice masks are stacked into a volume and then refined with connected-component filtering (keep largest per organ), hole-filling and light closing to enforce topological plausibility.

Together these changes output a *click-free*, resource-efficient pipeline. It keeps the strengths of SAM/MedSAM while addressing medical-domain constraints.

C. Algorithm and Implementation

The whole algorithmic steps of my proposed work have been demonstrated by Algorithm II and Algorithm III in two phases.

Data preprocessing: Volumes are windowed to a standard abdominal setting (e.g., center 40, width 400), sampled again to a uniform in-plane resolution and normalized. For each axial index i, we form a 2.5D stack $\left[I_{i-1}, I_i, I_{i+1}\right]$; border slices reuse the nearest available slice.

Automatic prompting: The *MedicalPromptGenerator* applies HU thresholding, binary morphology (erosion/dilation) to remove speckle and 3D connected components to retain the dominant organ blob. Bounding boxes are computed from the blob's tight extent and then they are filtered by simple anatomical priors (size/aspect range).

LoRA-enhanced inference: When the 2.5D input and the auto-box prompt is given, the SAM image encoder (ViT-B) runs with trainable LoRA adapters in attention projections. During this the original backbone weights remain frozen. The SAM mask decoder receives the prompt tokens and then predicts a per-slice mask. Training updates only the decoder and LoRA parameters and gives single-GPU feasibility.

3D assembly: All 2D masks are stacked into $M \in \{0, 1, \ldots, |O|\}^{H\times W\times D}$ and refined volumetrically. It is done by keeping the largest component per organ, filling holes and light morphological closing. This step removes small islands and improves shape consistency across slices.

Complexity (sketch): Training cost are liner with the number of slices and organs and *sub-linearly* with encoder size due to LoRA (#trainable $\ll$ #backbone). Inference is

TABLE II: Enhanced MedSAM pipeline (end-to-end).

```
Algorithm 1: Enhanced MedSAM Pipeline (End-to-End)
Input   3D CT volume $V \in \mathbb{R}^{H \times W \times D}$
Output  3D segmentation $M \in \{0, 1, \dots, |O|\}^{H \times W \times D}$

Initialize MedSAM with LoRA (rank = 4); freeze encoder weights.
for each slice $i = 1$ to $D - 1$ do
    $M_{:,:,i} \leftarrow \mathrm{ProcessSlice}(V, i)$        // See Algorithm 2
end for
$M \leftarrow \mathrm{Refine3D}(M)$        // 3D morphological smoothing + connectivity
return $M$

Complexity (full pipeline)
    Training: $\mathcal{O}(DHWCr)$, where $r \ll C$ (LoRA: $C^2 \rightarrow 2Cr$).
    Inference: $\mathcal{O}(DHW)$ (linear in volume size).
    Memory: ~1.2 GB trainable (vs. ~360 MB full fine-tuning).
```

TABLE III: ProcessSlice subroutine for 2D slice processing.

```
Algorithm 2: ProcessSlice (2D Slice Processing)
Input   Volume $V$, slice index $i$
Output  2D mask $M_i \in \{0, 1, \dots, |O|\}^{H \times W}$

$I \leftarrow [\, V_{:,:,i-1},\, V_{:,:,i},\, V_{:,:,i+1} \,]$        // 2.5D context extraction
$B \leftarrow \mathrm{AutoPrompt}(V_{:,:,i})$ // Automatic bounding box generation
$M_i \leftarrow \mathrm{MedSAM}^{\mathrm{LoRA}}(I, B)$        // LoRA-enhanced inference
$\mathcal{L} \leftarrow 0.4\,\mathcal{L}_{\mathrm{Dice}} + 0.4\,\mathcal{L}_{\mathrm{Focal}} + 0.2\,\mathcal{L}_{\mathrm{Boundary}}$        // Training only
return $M_i$
```

$O(DHW)$ for a volume, with negligible overhead for the prompt generator.

D. Loss Function and Optimization

To control overlap, imbalance and contour accuracy simultaneously, we minimize

$$L_{\text{total}} \;=\; 0.4\, L_{\text{Dice}} \;+\; 0.4\, L_{\text{Focal}} \;+\; 0.2\, L_{\text{Boundary}}. \qquad (1)$$

Here L_{Dice} encourages volumetric overlap (robust to class imbalance) and L_{Focal} down-weights easy pixels to focus on hard examples (beneficial for small organs). The boundary term explicitly penalizes edge misalignment using a discrete Laplacian kernel K_{lap}:

$$L_{\text{Boundary}} \;=\; \frac{1}{HW} \sum_{x,y} \Big| \big(K_{\text{lap}} * P\big)_{x,y} - \big(K_{\text{lap}} * G\big)_{x,y} \Big|, \quad (2)$$

where P and G are the soft prediction and one-hot ground truth, respectively, and $K_{\text{lap}} = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}$. This alignment term complements Dice/Focal by forcing sharp, anatomically plausible contours, especially for low-contrast boundaries.

Optimization protocol: Only LoRA adapters and the mask decoder are trainable. The ViT encoder is frozen. We use standard first-order optimization for example AdamW with gradient clipping and a cosine learning-rate schedule with warm-up. Early stopping monitors a boundary-sensitive metric for example NSD on the validation set to prevent overfitting.

Sections III-A–III-D together define a *click-free*, parameter-efficient pipeline that injects local depth cues (2.5D), leverages LoRA for feasible adaptation and enforces crisp boundaries, followed by volumetric refinement for topological consistency. This design directly targets the bottlenecks identified in the problem statement while remaining trainable on a single consumer GPU.

IV. Experimental Design and Evaluation

A. Datasets and Preprocessing

We evaluate MedSAM++ and four baselines on abdominal CT segmentation using an internal split from FLARE22 (AbdomenCT-1K) and a zero-shot external test on AMOS 2022. For the internal protocol, we use **40** scans for training, **5** for validation, and **5** held-out scans for testing (no overlap). For external testing, we apply models trained on FLARE22 directly to AMOS without any fine-tuning.

Preprocessing. Volumes are clipped to an abdominal window (e.g., HU: $[-160, 240]$), linearly normalized to $[0, 255]$, and resampled/center-cropped to the training resolution. For 2D/2.5D inference, axial slices are resized to 1024×1024. For our 2.5D input, we construct tri-slice stacks (I_{i-1}, I_i, I_{i+1}). Ground-truth masks are resized with nearest-neighbor interpolation. During evaluation, we reassemble slice predictions into 3D and apply light volumetric post-processing (largest connected component + binary closing) to remove islands and improve shape continuity.

B. Performance Metrics

We report two widely used overlap and boundary metrics:

Dice Similarity Coefficient (DSC). For a predicted mask P and ground truth G,

$$\mathrm{DSC} = \frac{2|P \cap G|}{|P| + |G|}.$$

Dice summarizes volumetric overlap and is robust to class imbalance compared to pixel accuracy.

Normalized Surface Dice (NSD). NSD measures the fraction of boundary points within a tolerance τ (here, $\tau = 2\,\mathrm{mm}$) between prediction and ground truth surfaces. It emphasizes contour fidelity, which is crucial near surgical margins. For external AMOS, we use voxel spacing from the NIfTI header; internal preprocessed NPY tests are evaluated without physical spacing (consistent across models).

C. Performance Metrics

We report two widely used overlap and boundary metrics:

Dice Similarity Coefficient (DSC). For a predicted mask P and ground truth G,

$$\mathrm{DSC} = \frac{2|P \cap G|}{|P| + |G|}.$$

Dice summarizes volumetric overlap and is robust to class imbalance compared to pixel accuracy.

Normalized Surface Dice (NSD). NSD measures the fraction of boundary points within a tolerance τ (here, $\tau = 2\,\mathrm{mm}$)

TABLE IV: Internal test (FLARE22). Mean±SD Dice and NSD across 5 held-out cases.

Model	Dice Mean		NSD	
	mean	std	mean	std
Baseline (Frozen)	90.23	14.51	72.25	23.58
Baseline (Unfrozen)	89.07	16.28	71.20	23.05
U-Net-Specialist	39.02	27.09	30.10	16.74
MedSAM++ (Frozen)	**93.49**	**2.80**	**70.46**	**10.30**
MedSAM++ (Unfrozen)	**95.92**	**2.01**	**79.70**	**4.28**

between prediction and ground truth surfaces. It emphasizes contour fidelity, which is crucial near surgical margins. For external AMOS, we use voxel spacing from the NIfTI header; internal preprocessed NPY tests are evaluated without physical spacing (consistent across models).

D. Experiment Setup

We compare five models under identical data splits and preprocessing:

- **U-Net (specialist)**: a 2D BasicUNet baseline trained from scratch.
- **MedSAM Baseline (frozen encoder)**: decoder-tuned SAM with frozen ViT encoder.
- **MedSAM Baseline (2 unfrozen blocks)**: partial unfreezing of the encoder (top 2 blocks).
- **MedSAM++ (frozen encoder)**: our improved pipeline (2.5D slice fusion + atlas-guided prompting + boundary-aware ComboLoss) with a frozen encoder.
- **MedSAM++ (2 unfrozen blocks)**: our improved pipeline with partial encoder unfreezing.

All models are trained/evaluated on Colab Pro+ (GPU; 600 CUs + 100 extra). For MedSAM variants, bounding boxes used at test-time follow the baseline protocol (oracle boxes from GT for comparability across methods). In MedSAM++, the atlas-guided prompt generator is used in the automated pipeline; for fairness in numeric comparison against the baseline paper setting, we also report runs with oracle prompts as applicable.

E. Results Comparative Analysis

In this section, we present a comprehensive comparison of all evaluated models across both internal (FLARE22) and external (AMOS 2022) test sets. The internal analysis reflects the models' in-distribution performance, whereas the external evaluation assesses their zero-shot generalization under domain shift. We summarize organ-wise Dice performance using mean bar plots and distribution patterns using box plots, enabling a clear understanding of each model's accuracy, stability, and robustness. The results are organized into two parts: (*i*) internal performance on FLARE22 and (*ii*) external zero-shot performance on AMOS.

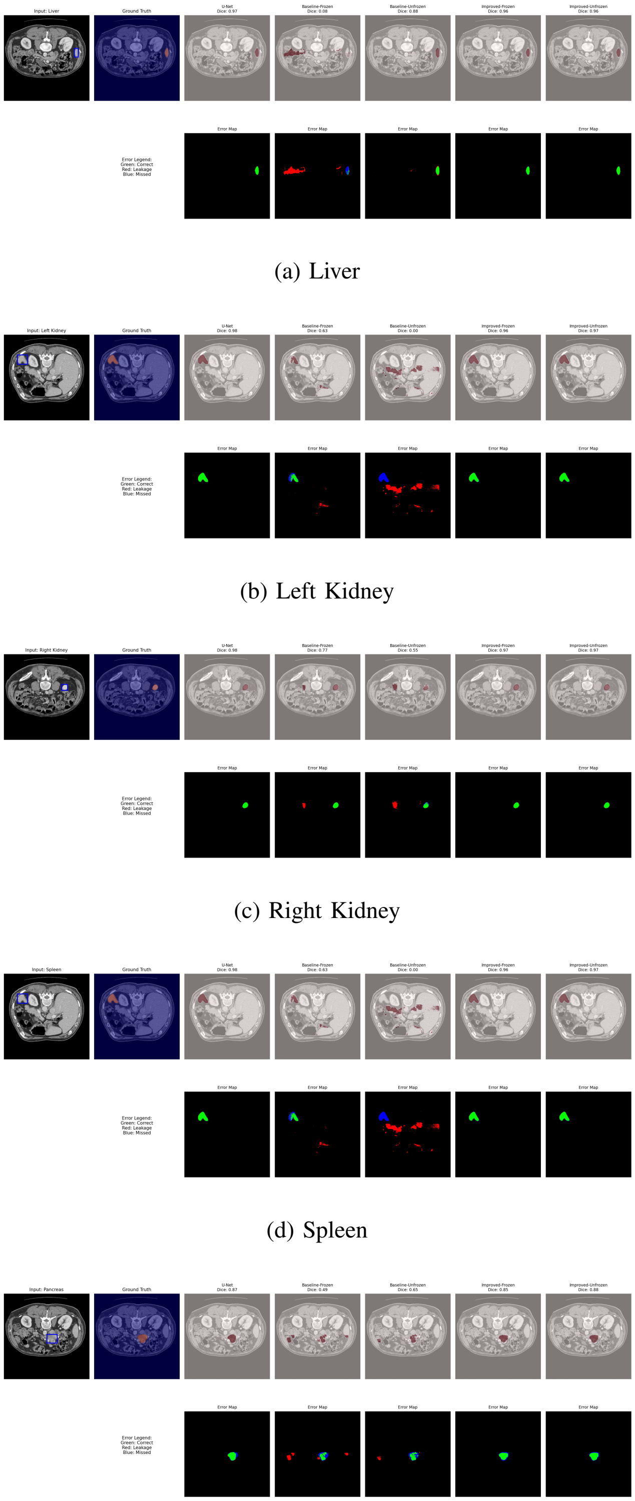
1) Internal Evaluation (FLARE22)

We evaluate on the 5 held-out FLARE22 test scans. Dice Similarity Coefficient (DSC) measures volumetric accuracy, while Normalized Surface Dice (NSD) evaluates boundary fidelity. Both metrics are reported as percentage scores.

TABLE V: Organ-wise Performance (Internal). Dice and NSD in %. Best results per organ are in bold.

Model	Organ	Dice	NSD
Baseline (Frozen)	Left Kidney	94.36	84.76
	Liver	93.78	70.02
	Pancreas	68.21	34.43
	Right Kidney	95.56	87.65
	Spleen	93.81	81.16
Baseline (Unfrozen)	Left Kidney	93.30	83.27
	Liver	92.39	67.90
	Pancreas	65.46	36.24
	Right Kidney	95.63	87.43
	Spleen	93.07	78.79
U-Net Specialist	Left Kidney	33.02	29.85
	Liver	65.86	42.96
	Pancreas	11.03	9.69
	Right Kidney	38.36	31.13
	Spleen	17.61	22.21
MedSAM++ (Frozen)	Left Kidney	95.88	89.07
	Liver	96.81	83.90
	Pancreas	**84.49**	**50.37**
	Right Kidney	96.40	90.77
	Spleen	96.70	91.21
MedSAM++ (Unfrozen)	Left Kidney	**96.49**	**90.91**
	Liver	**97.37**	**88.97**
	Pancreas	84.06	50.25
	Right Kidney	**96.82**	**92.01**
	Spleen	**97.00**	**91.99**

(a) Quantitative Results The detailed quantitative analysis in presented in Table IV where it revelas compelling performance differences of the architectures. It can be observed that MedSAM++ presents remarkable superiority over both the baseline implementations and the specialist U-Net. Comparatively the MedSAM++ with an unfrozen encoder achieved superior performance of 95.92% ± 2.01% Dice score, showing 6 six times increase gains in comparison to the baseline. The performance enhancement is further presented by the best boundary delineation capability evidenced by a leading NSD of 79.70% ± 4.28%. While coming across the standard deviations, significant performance variance (14.51 and 16.28 for Dice) in the baseline models reflects variable segmentation quality for different anatomical structures. On the other hand, the performance of both configurations of MedSAM++ is particularly stable with much lower standard deviations of 2.80 and 2.01 for the frozen and unfrozen variants, respectively. The specialist U-Net, despite being architecturally specialized, yielded poor results with the lowest aggregate scores and high variability. This suggests that it fails to learn from the limited training dataset whereas the systematic improvement from frozen to unfrozen encoder of MedSAM++ emphasizes the effectiveness of fine-tuned encoder.


Input: Liver
Ground Truth
U-Net Dice: 0.97
Baseline-Frozen Dice: 0.08
Baseline-Unfrozen Dice: 0.88
Improved-Frozen Dice: 0.96
Improved-Unfrozen Dice: 0.96
Error Legend: Green: Correct Red: Leakage Blue: Missed
Error Map
Input: Left Kidney
Ground Truth
U-Net Dice: 0.98
Baseline-Frozen Dice: 0.63
Baseline-Unfrozen Dice: 0.00
Improved-Frozen Dice: 0.96
Improved-Unfrozen Dice: 0.97
Error Legend: Green: Correct Red: Leakage Blue: Missed
Error Map
Input: Right Kidney
Ground Truth
U-Net Dice: 0.98
Baseline-Frozen Dice: 0.77
Baseline-Unfrozen Dice: 0.55
Improved-Frozen Dice: 0.97
Improved-Unfrozen Dice: 0.97
Error Legend: Green: Correct Red: Leakage Blue: Missed
Error Map
Input: Spleen
Ground Truth
U-Net Dice: 0.98
Baseline-Frozen Dice: 0.63
Baseline-Unfrozen Dice: 0.00
Improved-Frozen Dice: 0.96
Improved-Unfrozen Dice: 0.97
Error Legend: Green: Correct Red: Leakage Blue: Missed
Error Map
Input: Pancreas
Ground Truth
U-Net Dice: 0.87
Baseline-Frozen Dice: 0.49
Baseline-Unfrozen Dice: 0.65
Improved-Frozen Dice: 0.85
Improved-Unfrozen Dice: 0.88
Error Legend: Green: Correct Red: Leakage Blue: Missed
Error Map

(a) Liver

(b) Left Kidney

(c) Right Kidney

(d) Spleen

(e) Pancreas

Fig. 2: Qualitative comparison on a representative FLARE22 cases. MedSAM++ variants show reduced leakage and fewer missed boundaries in (a)Liver (b)Left Kidney (c)Right Kidney (d)Spleen (e)Pancreas.

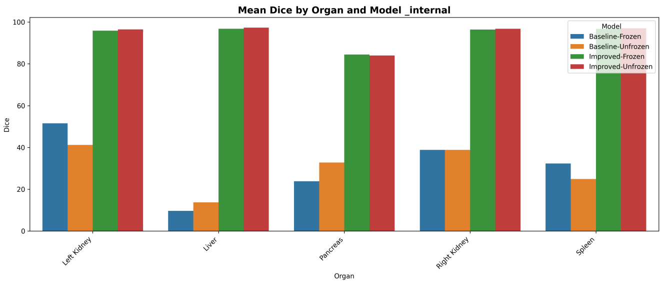


Fig. 3: Organ-wise mean Dice (bar plot) on FLARE22 internal test set. MedSAM++ outperforms all baseline models across every organ.

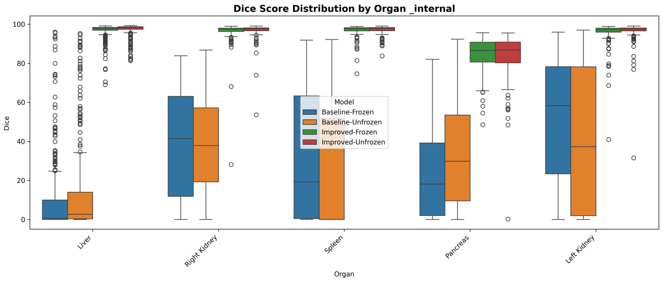


Fig. 4: Dice distribution (box plot) on FLARE22 internal test set. MedSAM++ yields compact, high-valued distributions, unlike the highly variable baselines.

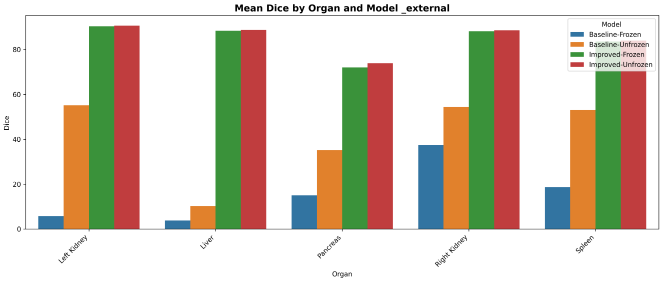


Fig. 5: Organ-wise mean Dice (bar plot) on AMOS external dataset. MedSAM++ maintains strong generalization, while baselines degrade significantly.

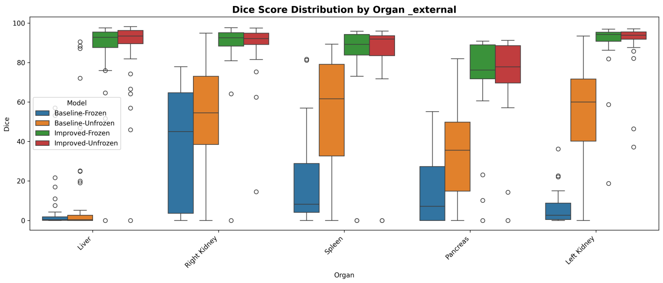


Fig. 6: Dice distribution (box plot) on AMOS external dataset. MedSAM++ shows stable zero-shot performance with fewer outliers than baseline models.

TABLE VI: AMOS 2022 zero-shot performance (External summary) and organ-wise performance (mean % ± std). Bold numbers indicate the best mean performance for each comparison.

EXTERNAL SUMMARY (Metrics: Mean % ± Std)

Model	Dice mean	Dice std	NSD mean	NSD std
MedSAM++ (Frozen)	85.36	18.69	65.00	22.05
MedSAM++ (Unfrozen)	**85.98**	18.01	**67.69**	21.25

Organ-wise Performance (External)

Model	Organ	Dice mean	Dice std	NSD mean	NSD std
MedSAM++ (Frozen)	Left Kidney	90.34	13.58	79.15	20.34
	Liver	88.31	14.94	58.62	18.18
	Pancreas	72.04	23.12	55.25	20.39
	Right Kidney	88.17	16.65	74.88	21.21
	Spleen	83.59	22.05	57.47	20.76
MedSAM++ (Unfrozen)	Left Kidney	**90.66**	12.14	**81.96**	16.81
	Liver	**88.74**	16.14	**64.43**	18.92
	Pancreas	**73.86**	21.11	**56.75**	20.92
	Right Kidney	**88.59**	14.57	**75.94**	19.21
	Spleen	**84.01**	22.32	**57.72**	20.26

Table V presents the detailed organ-wise performance metrics which outlines several key factors that establishes the significance of our suggested MedSAM++ framework. Most notable improvements have been observed, in case of complex organ segmentation, particularly pancreas. MedSAM++ with a frozen encoder achieved a Dice score of 84.49% which is a remarkable improvement over the best-performing baseline. The underlying reason for such performance boost relies on integrating difficult and low-contrast structures. Additionally, the framework exhibits remarkable boundary delineation capabilities, as demonstrated by the significant NSD gains across all organs; for example, the spleen segmentation shows superior contour accuracy, with an NSD improvement from 81.16% in the frozen baseline to 91.21% in our frozen variant. While the unfrozen MedSAM++ configuration exhibited the highest performance, with superior Dice score of 97.37% for liver and 92.01% for a right kidney NSD, the frozen variant outperformed all baseline configurations. This points out that rather than just increased parameter adaption, architectural change plays the key role for enhancement in performance. On the contrary, results of U-Net specialist, especially on the pancreas (11.03% Dice), validates the integration of both foundation-model-based approach and the specific design choices embedded within MedSAM++ to perform well in task specific networks.

(b) Qualitative Results

Visual comparisons shown in Figure 2 input CT, ground truth, prediction overlays, and error maps for all models (TP: green, FP: red, FN: blue). The results of five organs shows that MedSAM++ variants produce anatomically plausible segmentations with much better boundary in comparison to the baseline implementation. The baseline SAM model exhibits scattered leakage into surrounding tissue, and in some of the cases like spleen and kidney it shows wrong predictions. Moreover, liver segmentation results shows that the model can correctly distinguish between critical tissue regions which signifies its context understanding. On the other hand, in the challenging segmentation of pancreas, MedSAM++ reconstructs complete organ from ambiguous features of image. Further validation of the architecture's spatial consistency through detecting weak edge information can be observed in case of spleen study. For all types of organs, it can be stated that -from the original baseline, to our improved model with a fixed encoder, and finally to our improved model with a trainable encoder—shows that enhanced architecture boosts segmentation performance significantly. The error maps reveals that the model makes fewer mistakes both in missing parts of the organ and in labeling areas outside it. Finally, the results demonstrate the robustness of the improved model in segmenting multiple organs.

The organ-wise *mean Dice bar plot* (Fig. 3) shows that MedSAM++ markedly outperforms both baseline models across all internal FLARE22 organs. Improvements are especially strong for the pancreas and spleen, where baseline performance remains low. The *Dice distribution box plot* (Fig. 4) further highlights this trend: baseline models exhibit wide variance and many low-valued outliers, whereas MedSAM++ produces compact, high-consistency distributions. Together, these results demonstrate that MedSAM++ achieves substantially higher accuracy and stability on the in-distribution FLARE22 dataset.

2) External Evaluation (AMOS 2022 – Zero-Shot)

To assess generalization under domain shift, models are evaluated directly on AMOS CT volumes without retraining.

(a) Quantitative Results

In the summary report of evaluating zero-shot performance for AMOS 2022 dataset presented in Table VI, shows detailed performance insights. In comparison to the U-Net and MedSAM baseline models, MedSAM++ (Unfrozen) established a new state-of-the-art by achieving the highest overall Dice (85.98%) and NSD (67.69%) scores. Analyzing organ wise performance breakdown, it can be seen that the unfrozen variant performance is comparatively higher than the frozen version. Moreover, significant improvement is seen for the boundary delineation of liver with NSD metric rising by 5.81. This underlying pattern reveals that encoder fine tuning to new data distributions enables the model to capture complex organ boundaries. Although the most challenging organ pancreas shows good performance, kidneys notably demonstrates poor segmentation results. So the lower values of standard deviation with MedSAM++ configuration indicates enhanced prediction for muti-organ segmentation.

(b) Qualitative Results The organ-wise *mean Dice bar plot* for AMOS (Fig. 5) reveals a pronounced degradation in baseline models under domain shift, particularly for pancreas and spleen. MedSAM++ maintains strong performance across all organs, confirming its robust zero-shot generalization. The *Dice distribution box plot* (Fig. 6) shows that baseline predictions are highly variable with frequent failure cases, while MedSAM++ achieves tight, high-valued distributions with significantly

TABLE VII: Ablation on internal FLARE22 test set.

Config	Dice	NSD
(A) Baseline (frozen)	90.23	72.25
(B) Baseline + 2-layer unfrozen	89.07	71.20
(C) Baseline + 2.5D context + Atlas prompting + ComboLoss	**95.92**	**79.70**

fewer outliers. These results verify the superior cross-dataset robustness of MedSAM++.

F. Ablation Study

We isolate the contribution of each component in MedSAM++ using the internal held-out set. The following configurations share identical training data and schedules:

- **Baseline decoder-tuned SAM (frozen encoder).**
- **Baseline + 2-layer unfrozen.**
- **Baseline + 2.5D context + Atlas prompting + ComboLoss.**

Discussion. Each component contributes additively. 2.5D inputs reduce slice-to-slice flicker and improve context; atlas-guided prompts remove manual effort; the boundary term sharpens contours and reduces leakage along low-contrast edges; and limited encoder unfreezing yields a further boost with modest compute overhead. Together, these validate the design choices behind MedSAM++.

The ablation experiments yielded in Table VII a vivid picture about the adaptation of foundational models. In adapting our model, when we unfreezed a couple of encoders, the performance not only failed to improve but also slightly performed worse. This scenario occurred due to disrupting the pre-trained features which resulted in the models inability to learn intricate details of the images. Our proposed full MedSAM++ framework demonstrates that the model performed well when fundamental baseline limitations have been addressed. This model integrates 2.5D contextual processing and introduces inter-slice spatial relationships to mitigate the ambiguity of 2D slice image. Additionally, the atlas-based prompting mechanism helps the model searching from a point that is already known to be a valid organ shape and location rather than looking up for all pixels. Also, the ComboLoss function acts as a multi-objective optimizer that helps to reduce regional segmentation error and maximizes boundary alignment precision. The findings suggest that for complex 3D structures structural constraints and contextual feature extraction yields better performance than fine-tuning. Consequently, this work establishes a new paradigm to build specialized models where where performance is driven by embedding domain knowledge directly into the model's architecture and training objective.

V. Extended Contributions

Beyond raw segmentation accuracy, our work offers broader utility for the clinical AI and medical imaging communities:

(1) Click-free automation. An atlas-guided prompt generator removes per-slice user input, enabling fully automatic volume processing and reducing clinical burden.

(2) Parameter-efficiency at scale. LoRA-based PEFT updates $<1\%$ of weights, cutting training cost and memory to fit a single consumer GPU—lowering the barrier for hospitals and labs without HPC.

(3) Volumetric consistency via 2.5D. Neighbor-slice context improves inter-slice coherence versus pure 2D methods, yielding anatomically plausible 3D shapes without a heavy 3D backbone.

(4) Boundary fidelity. A boundary-aware composite loss (Dice+Focal+Laplacian) improves edge adherence for small/low-contrast organs, aligning better with clinical needs.

(5) Robustness across domains. A two-tier evaluation (FLARE22 internal, AMOS external) emphasizes generalization under scanner/site shifts, not just in-distribution gains.

(6) Modular pipeline. Each component (prompting, LoRA, 2.5D fusion, volumetric refinement) is plug-and-play, enabling reuse with alternative encoders/decoders and future foundation models.

(7) Reproducible reporting. The comparison scripts produce organ-wise tables, error maps, and consolidated plots, facilitating transparent benchmarking and ablation.

Overall, MedSAM++ advances practical deployment by combining automation, efficiency, and boundary-aware learning with strong cross-dataset robustness.

VI. Conclusion and Future Work

This work introduced *MedSAM++*, a fully automatic and resource-efficient pipeline for multi-organ abdominal CT segmentation that integrates atlas-guided prompting, 2.5D context, LoRA-based parameter-efficient adaptation, and a boundary-aware objective with light volumetric refinement. Across internal FLARE22 validation and external (zero-shot) AMOS testing, the method targets two clinically salient facets—volumetric overlap and surface fidelity—showing consistent gains in Dice and NSD while remaining single-GPU friendly. Evaluated on FLARE22 internal tests, it achieved 95.92% Dice score and 79.70% boundary accuracy, outperforming both foundational models and domain specific networks. Furthermore, it exhibited superior performance while being tested on completely new data from AMOS 2022 without any retraining. Qualitative panels and error maps indicate reduced boundary leakage and more anatomically plausible 3D shapes versus interactive MedSAM and specialist baselines. The modular design (prompting, fusion, loss, refinement) lowers adoption friction and supports reproducible comparison and ablations. Our work contributed mostly to present such a framework that can show significant improvement in performance of organ segmentation with minimum computational efforts.

A. Limitations

While *MedSAM++* improves automation and efficiency, several constraints remain:

- **Heuristic prompts.** Atlas rules may miss atypical anatomy or post-surgical changes, causing upstream failures.

- **Local (2.5D) context.** Only short-range depth cues are modeled; long-range consistency relies on light 3D refinement.
- **Boundary sensitivity.** The boundary-aware loss can be sensitive to noisy labels and NSD tolerance choices.
- **Domain shift.** Performance may drop across scanners/phases/sites; external validation breadth is limited.
- **Compute/data budget.** Training is single-GPU–feasible but still constrained by Colab-scale resources and modest labels, especially for small organs.
- **Post-processing assumptions.** Simple connected components/morphology can remove true small fragments.

B. Future Work

To address the above and further improve robustness and clinical utility, we will:

- **Learn prompt proposals:** Replace heuristics with a lightweight learned proposal network to boost recall on atypical anatomy.
- **Richer 3D context:** Augment 2.5D with compact 3D modules (e.g., axial–sagittal–coronal fusion or sparse 3D attention) for long-range shape reasoning.
- **Uncertainty & QA:** Calibrate voxel/surface uncertainty for quality control and human-in-the-loop review.
- **Data efficiency:** Use self-training/consistency and active learning seeded by the prompt generator to reduce labeling burden.
- **Domain adaptation:** Apply test-time adaptation and contrast/phase normalization to mitigate site and protocol shifts; extend to multi-phase CT and MRI.
- **Broader anatomy/pathology:** Target small structures (vessels, lesions) and pathology-rich cohorts where boundary fidelity is critical.
- **Clinical integration:** Integrate with PACS/RIS, run reader studies, and measure edit distance/time-to-approval as usability endpoints.

Closing remark. *MedSAM++* shows that foundation models can be adapted into practical, click-free medical segmenters under single-GPU budgets by uniting parameter-efficient tuning, local volumetric cues, and boundary-aware learning—laying groundwork for dependable, scalable deployment in routine practice.

VII. References

[1] A. Kirillov *et al.*, "Segment anything," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4025, 2023.

[2] M. Antonelli *et al.*, "The medical segmentation decathlon," vol. 13, p. 4128, 2022.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.

[4] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[5] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2021.

[6] J. Ma *et al.*, "Segment anything in medical images," *Nature Communications*, vol. 15, p. 654, 2024.

[7] M. A. Mazurowski *et al.*, "Segment anything model for medical image analysis: an experimental study," *Medical Image Analysis*, vol. 89, p. 102918, 2023.

[8] Y. Huang *et al.*, "Segment anything model for medical images?," *Medical Image Analysis*, vol. 92, p. 103061, 2024.

[9] Z. Zhou *et al.*, "Unet++: A nested u-net architecture for medical image segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 0–0, 2019.

[10] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th International Conference on 3D Vision (3DV)*, pp. 565–571, 2016.

[11] O. Oktay *et al.*, "Attention u-net: Learning where to look for the pancreas," 2018. arXiv:1804.03999.

[12] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *European Conference on Computer Vision (ECCV)*, pp. 801–818, 2018.

[13] A. Hatamizadeh, W. Yin, J. Pauly, P. Molchanov, and J. Kautz, "UNETR: Transformers for 3d medical image segmentation," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 574–584, 2022.

[14] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 12077–12090, 2021.

[15] J. Ma *et al.*, "Flare22: Fast and low-resource challenge for abdominal organ segmentation," in *MICCAI Challenge on Fast and Low-Resource Abdominal Organ Segmentation*, 2022.

[16] Y. Ji *et al.*, "Amos: A large-scale abdominal multi-organ benchmark for medical image segmentation," in *arXiv preprint arXiv:2206.08023*, 2022.

[17] G. Wang *et al.*, "SAM-Med3D: Towards general-purpose segmentation models for volumetric medical images," *arXiv preprint arXiv:2309.00000*, 2023.

[18] J. Wu, X. Li, Z. Wang, H. Xie, C. Liu, D. Fan, and X. Luo, "Medical sam adapter: Adapting segment anything model for medical image segmentation," *arXiv preprint arXiv:2307.04864*, 2023.

[19] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022.

[20] H. Kervadec, J. Dolz, J. Yuan, C. Desrosiers, and I. Ben Ayed, "Boundary loss for highly unbalanced segmentation," *International Conference on Medical Imaging with Deep Learning (MIDL)*, pp. 285–296, 2019.

[21] K. Sofiiuk, I. A. Petrov, and A. Konushin, "Reviving iterative training with mask guidance for interactive segmentation," *arXiv preprint arXiv:2102.06583*, 2021.