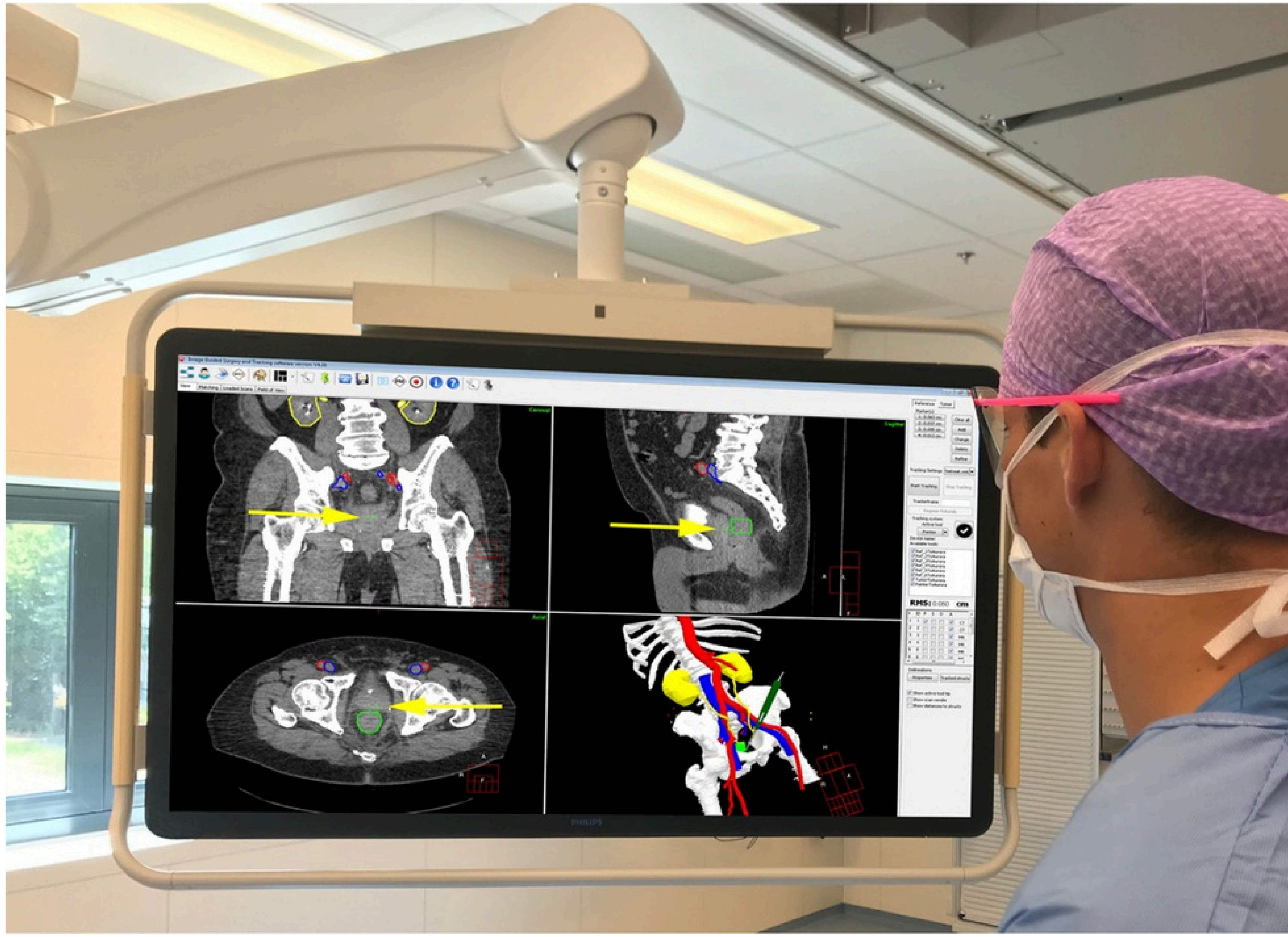
MedSAM++: Automated Multi-Organ Segmentation with Atlas-Guided Prompting, 2.5D Context and Volumetric Refinement



Presented By:

Nuren Nafisa

ID: g202427580

Supervised By:

Dr. Muzammil Behzad

King Fahd University of Petroleum and Minerals

The Problem & Clinical Motivation

- Medical image segmentation is crucial for diagnosis, surgery, and analysis.
- Foundation Models (like SAM) are powerful but built for natural images, not medical CTs.
- The Gap: Direct application fails due to:
    - Grayscale & Low-Contast nature of CT.
    - Lack of 3D Awareness: Treats each slice independently.
    - Manual Prompting: Requires a human to click on every slice—impossible for 3D volumes.

Research Gap

- Lack of Automation
- Ignored Volumetric Context
- Computational Inefficiency

Challenges

1. Manual Prompting
Requires human clicks per slice, preventing full automation for 3D volumes.
2. 2D Slice Inference
Lacks 3D context, causing inconsistent organ shapes across slices.
3. High Fine-Tuning Cost
Massive model size makes adaptation computationally expensive.

Our Solution: The MedSAM++ Pipeline

We transform the interactive MedSAM into a fully automated, efficient, and accurate 3D segmenter.

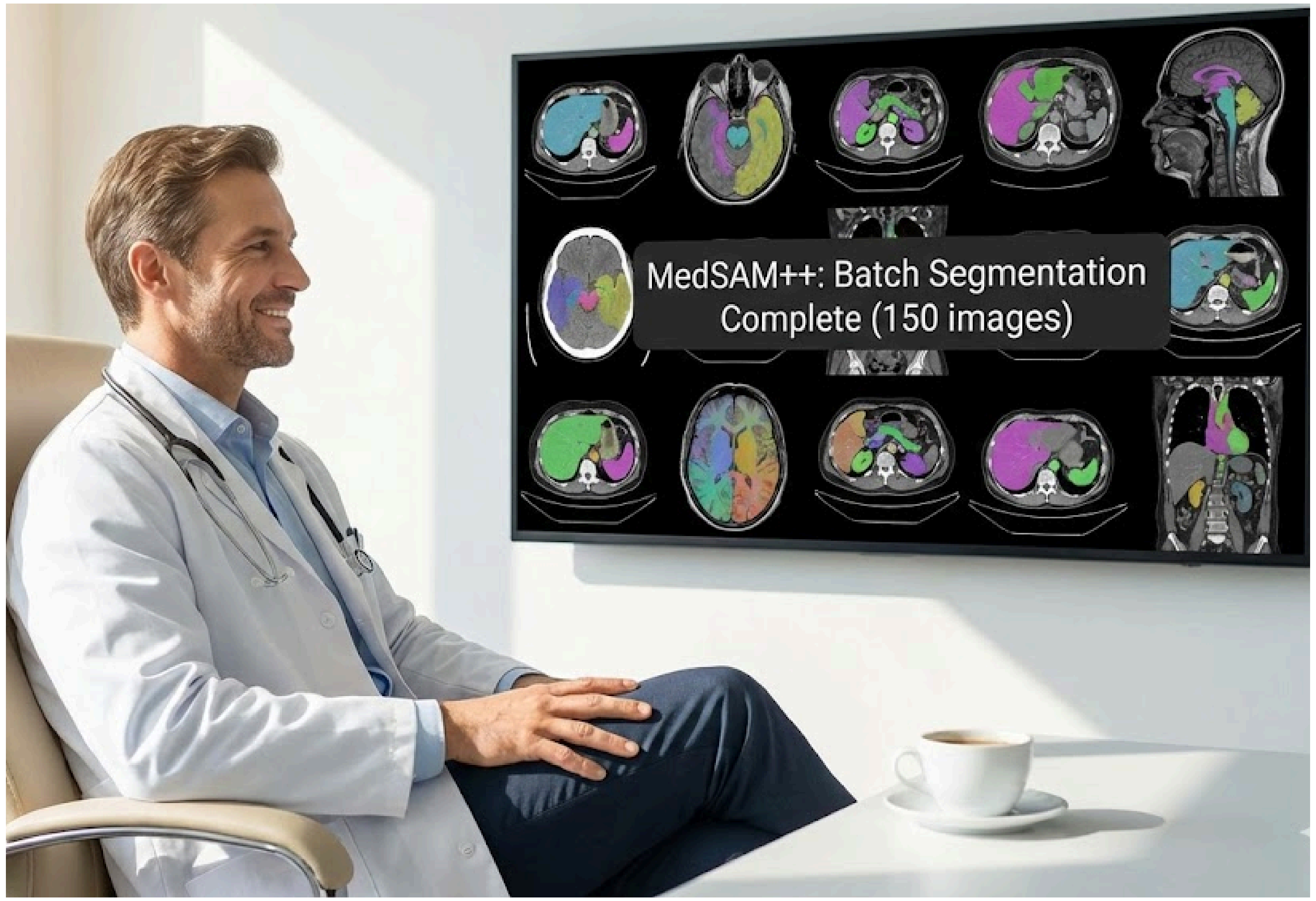
The Four Key Innovations:

01 AUTOMATION: ATLAS-GUIDED PROMPT GENERATOR (NO MORE MANUAL CLICKS).

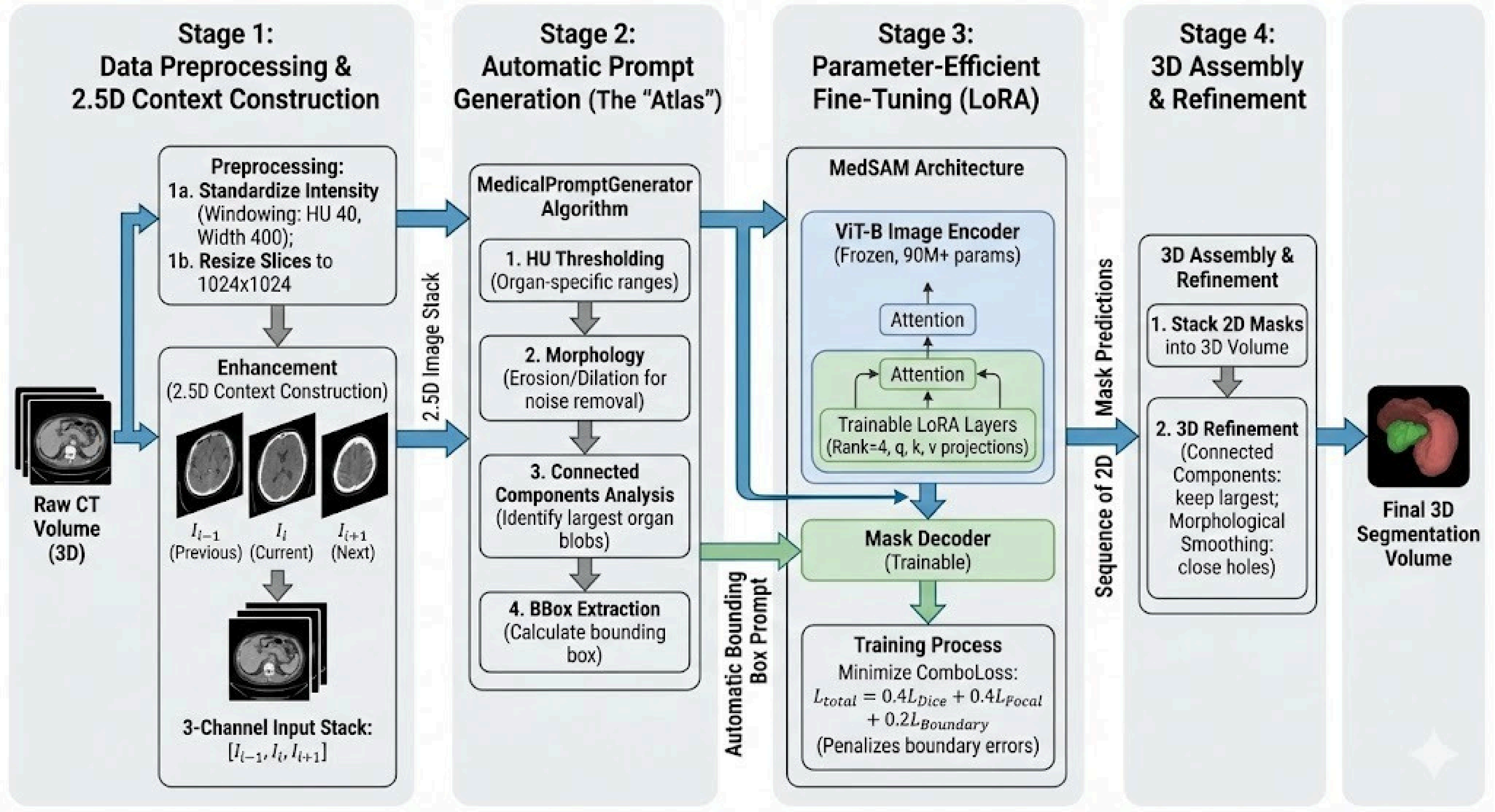
02 EFFICIENCY: LORA FOR PARAMETER-EFFICIENT FINE-TUNING (<1% PARAMETERS TRAINED).

03 CONTEXT: 2.5D SLICE INPUTS FOR VOLUMETRIC COHERENCE.

04 PRECISION: BOUNDARY-AWARE LOSS FOR SHARPER ORGAN EDGES.



Proposed Methodology



Technical Deep Dive 1: Automation & Context

Atlas-Guided Automatic Prompting

- Replaces the human with simple, robust heuristics:
- Process: HU Thresholding → Morphological Cleaning → Largest Connected Component → Bounding Box.

2.5D Context

- Problem: 2D models cause "flickering" between slices.
- Solution: Input = [Slice_i-1, Slice_i, Slice_i+1].
- Benefit: Provides local depth context without the heavy cost of a full 3D model.

Technical Deep Dive 2: Efficiency & Precision

LoRA for Parameter-Efficient Fine-Tuning:

- Problem: Full fine-tuning of the giant ViT encoder is computationally expensive and leads to overfitting.
- Solution: LoRA (Low-Rank Adaptation): Inject small, trainable rank-decomposition matrices into the attention layers. Freezes the original weights.
- Benefit: We train <1% of the parameters! Makes single-GPU training feasible.

Boundary-Aware Combo Loss:

- Problem: Standard Dice loss ignores boundary sharpness.
- Solution: $L_{Total} = \lambda_1 * L_{Dice} + \lambda_2 * L_{Focal} + \lambda_3 * L_{Boundary}$
- The Laplacian Boundary Loss explicitly penalizes blurry edges.

Experimental Setup

Datasets:

- Internal (Training/Validation): FLARE22 (Abdominal CT)
- External (Generalization Test): AMOS 2022 (Zero-shot)

Baselines:

- U-Net (Specialist)
- MedSAM (Frozen Encoder)
- MedSAM (Partially Unfrozen)
- Our Models: MedSAM++ (Frozen & Unfrozen)

Metrics:

- Dice (DSC): Volumetric Overlap.
- Normalized Surface Dice (NSD): Boundary Accuracy (Very Important!).

Key Result 1: Quantitative Performance (Internal Test)

We achieve significant gains in both volumetric overlap and boundary precision.

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**

Key Result 2: Quantitative Performance (organ wise Metrics Internal)

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**

Key Result 2: Zero-Shot-Generalization

External Test (AMOS)

- Show a similar table for AMOS, demonstrating that your gains hold even on unseen data from a different source. This proves robustness.

TABLE V: 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)		Dice		NSD	
Model		mean	std	mean	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)					
		Dice		NSD	
Model	Organ	mean	std	mean	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

Key Result 3: Qualitative Results

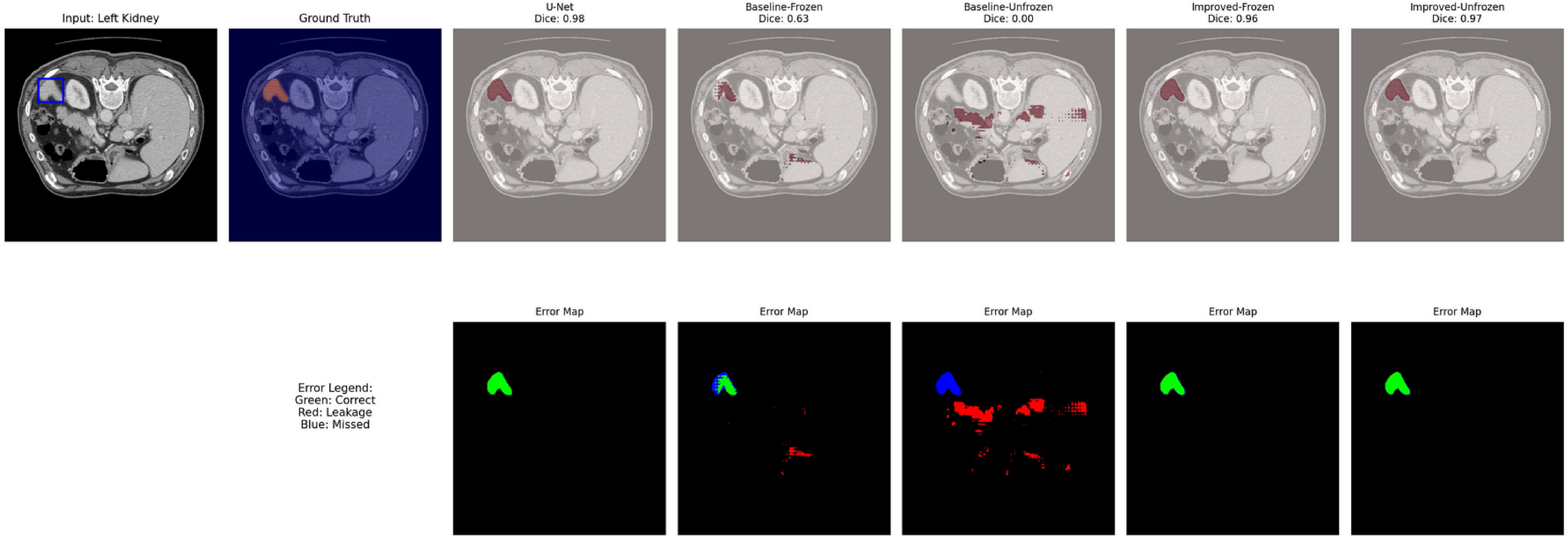


Figure-1: (a) Qualitative comparison on a representative FLARE22 cases. MedSAM++ variants show reduced leakage and fewer missed boundaries in **Left kidney**

Key Result 3: Qualitative Results

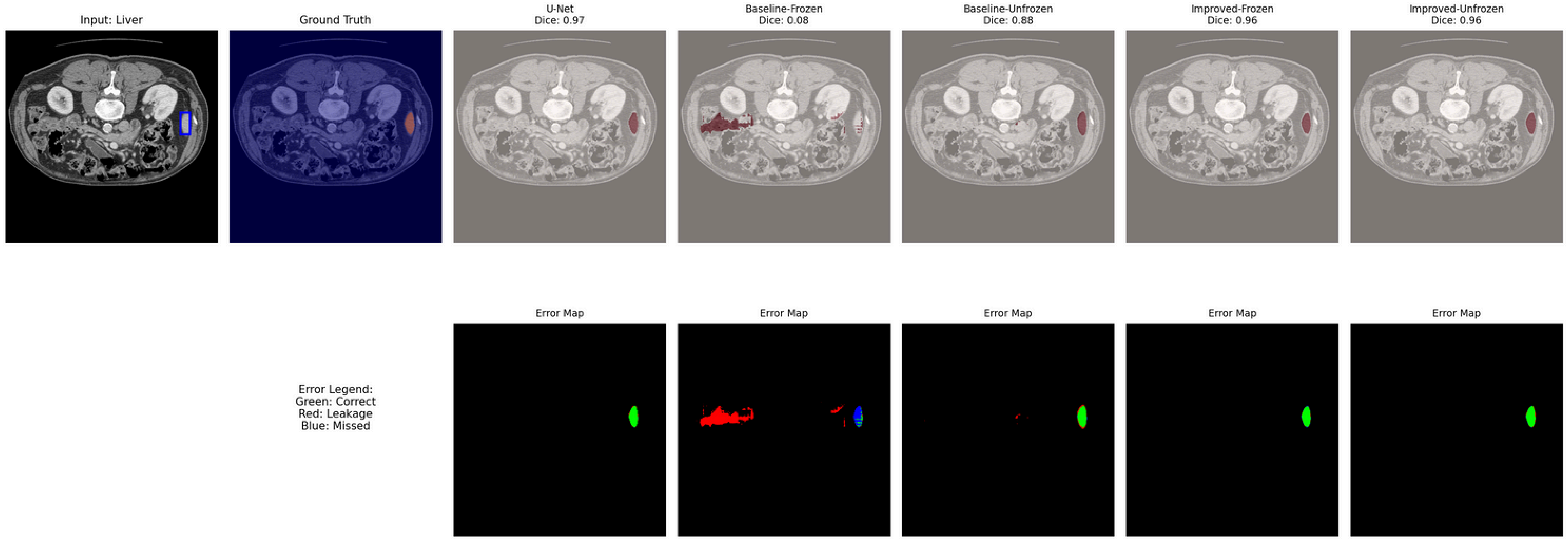


Figure-1: (b) Qualitative comparison on a representative FLARE22 cases. MedSAM++ variants show reduced leakage and fewer missed boundaries in **Liver**

Key Result 3: Qualitative Results

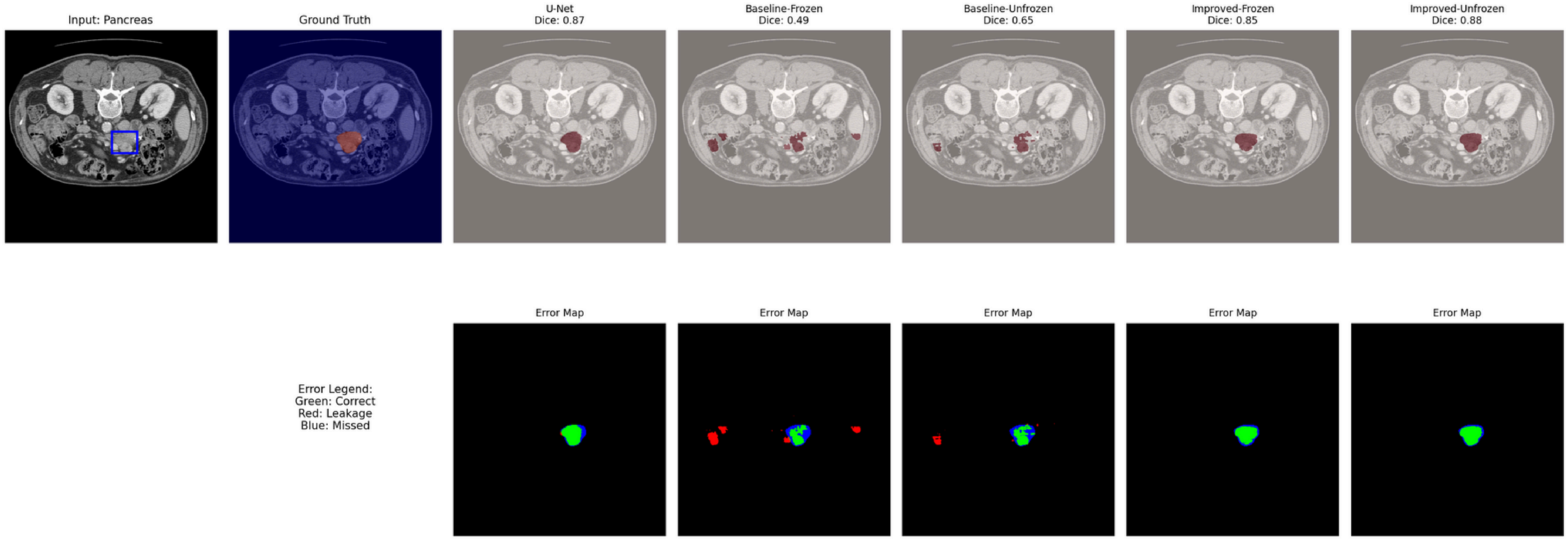


Figure-1: (c) Qualitative comparison on a representative FLARE22 cases. MedSAM++ variants show reduced leakage and fewer missed boundaries in **Pancreas**

Key Result 3: Qualitative Results

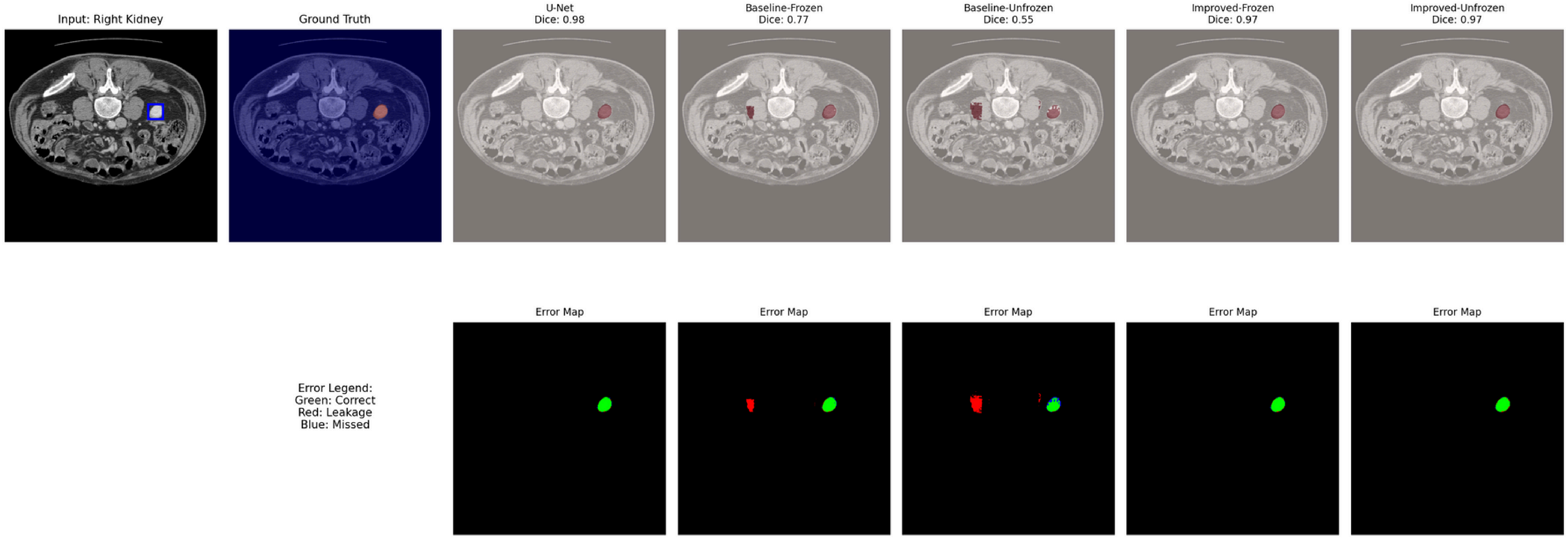


Figure-1: (d) Qualitative comparison on a representative FLARE22 cases. MedSAM++ variants show reduced leakage and fewer missed boundaries in **Right Kidney**

Key Result 3: Qualitative Results

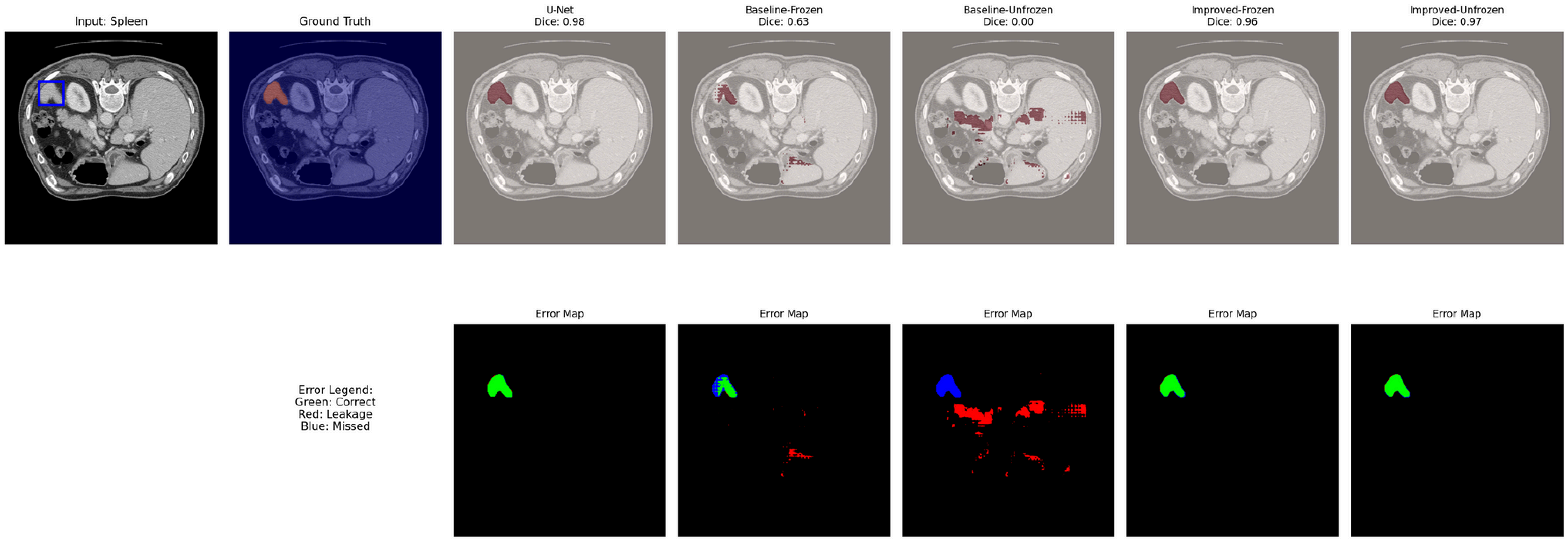
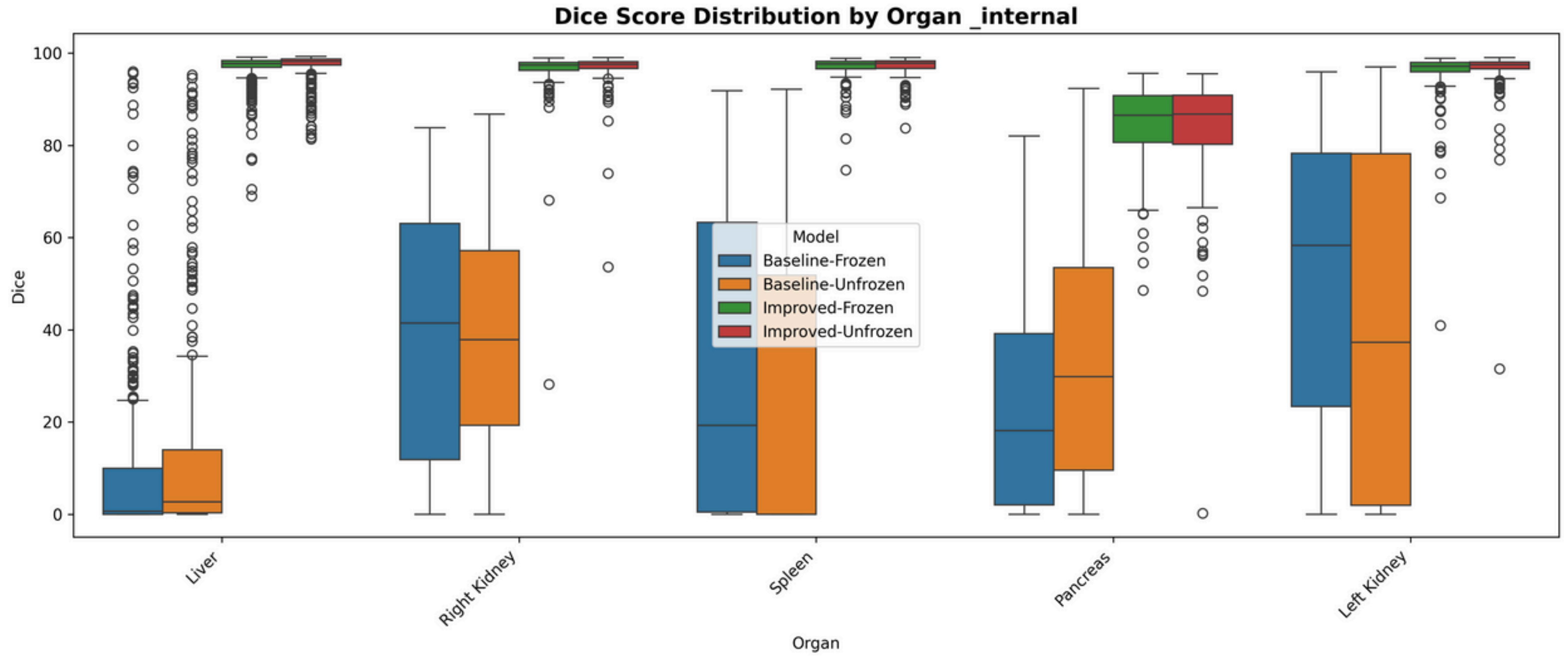
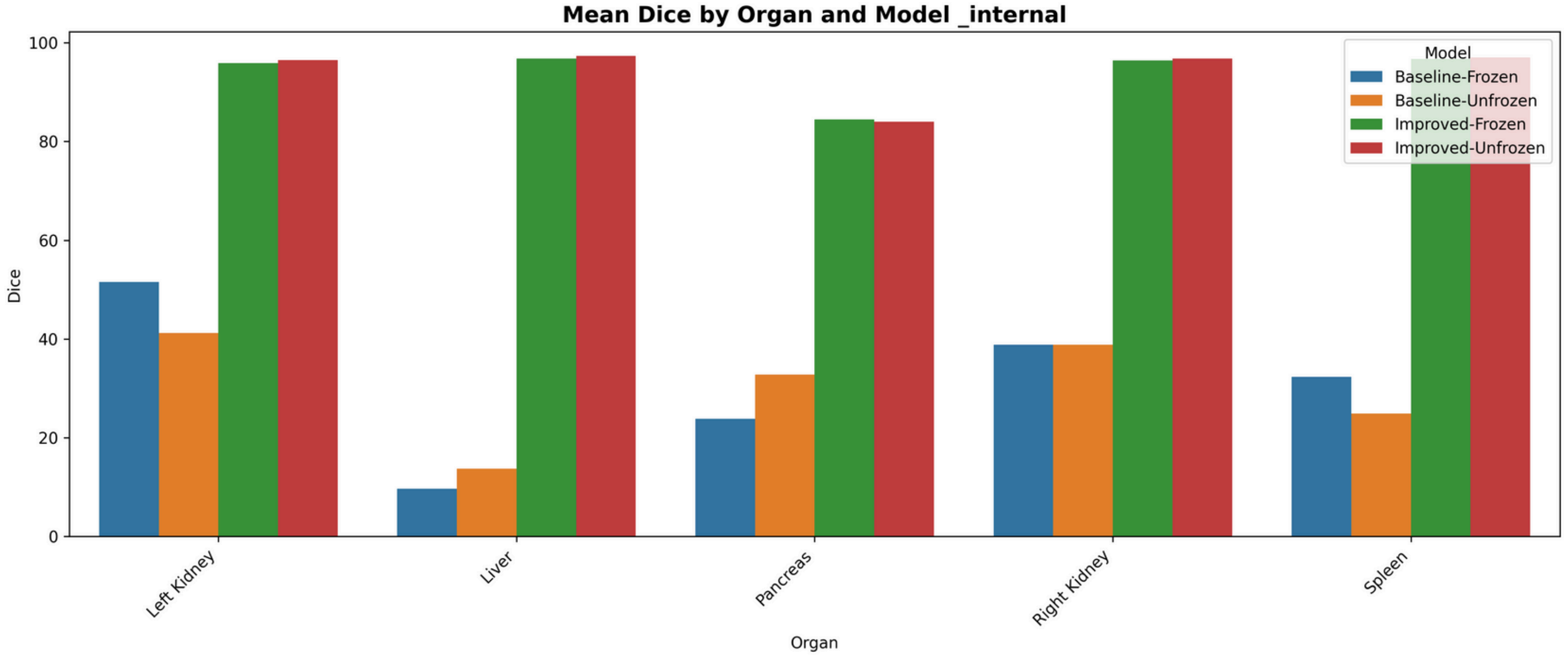


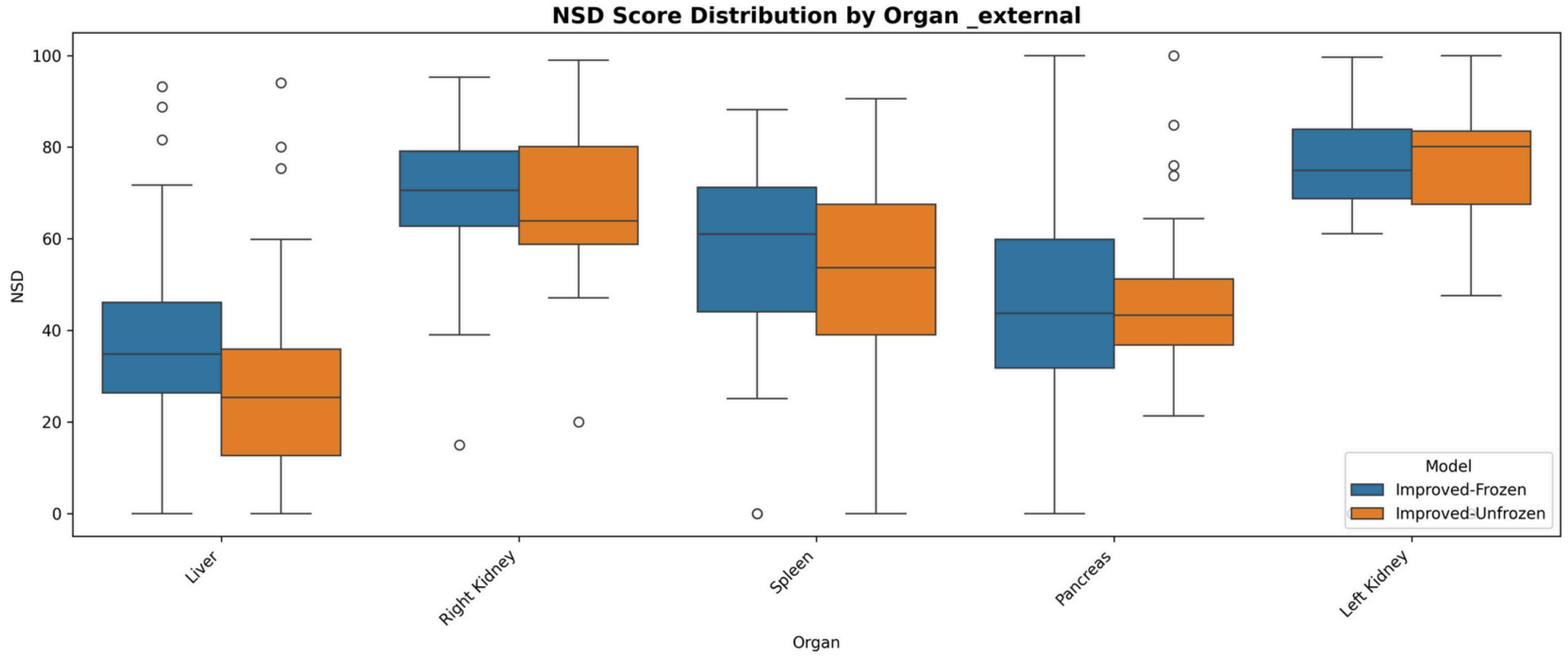
Figure-1: (e) Qualitative comparison on a representative FLARE22 cases. MedSAM++ variants show reduced leakage and fewer missed boundaries in **Spleen**

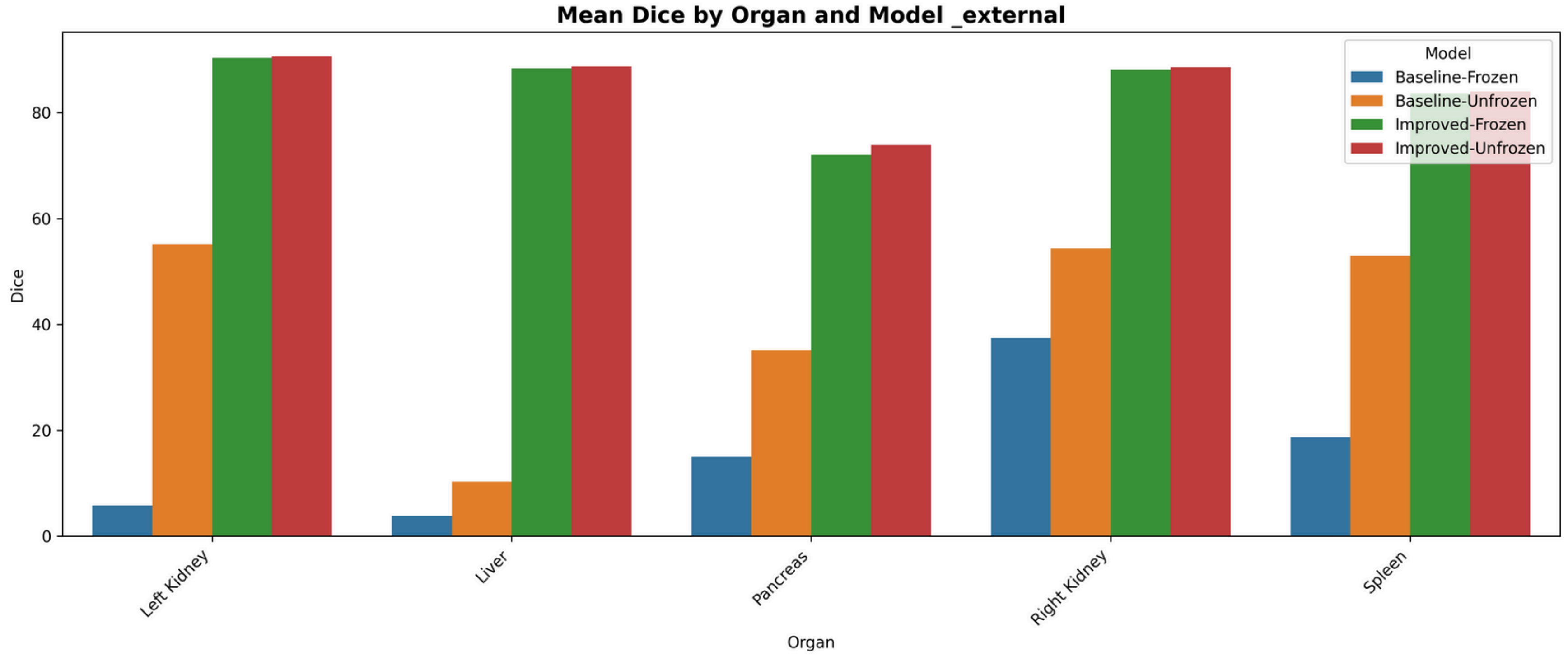
Key Result 3: Qualitative Results





Key Result 2: Qualitative Results





Ablation Study

- This is critical to prove that each of our ideas matters.

- Visual: A horizontal bar chart showing the progressive improvement from the baseline (A) to the full model (E), as in your Table VI.

- A: Baseline → B: +2.5D → C: +Prompting → D: +Boundary Loss → E: Full Model.
- Clearly label the Dice score on each bar.

TABLE VI: 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**

Conclusion and Summary

MedSAM++ successfully converts a prompt-based 2D model into an automated, efficient, and accurate 3D medical segmenter.

List of Core Contributions:

- Click-Free Automation via atlas-guided prompts.
- Parameter-Efficiency via LoRA, enabling single-GPU training.
- Volumetric Consistency via 2.5D context.
- Boundary Fidelity via a novel composite loss.
- Proven Robustness via internal and external validation.

Future Work

- Smarter Prompts: Replace heuristics with a lightweight neural proposal network.
- Beyond 2.5D: Explore efficient 3D transformers or multi-planar fusion.
- Handling Uncertainty: Integrate confidence measures and quality control flags.
- Clinical Deployment: Integrate with hospital systems (PACS) and conduct user studies.

Thank You

Assala-mu-alaikum