Deep Learning-Based MR Image Reconstruction and Segmentation

Mingjian Li
New York University
ml8347@nyu.edu

Shihang Wei
New York University
sw5672@nyu.edu

Haoyang Pei (Mentor)
New York University
hp2173@nyu.edu

***Abstract*—Magnetic Resonance Imaging (MRI) provides exceptional diagnostic value but is limited by inherently long acquisition times and the need for reliable downstream analysis. This project aims to build a unified deep learning framework that accelerates MRI reconstruction from undersampled k-space data while enabling high-quality segmentation for clinical interpretation. Our approach combines physics-informed reconstruction models with data-driven priors, leveraging recent advances in variational networks, score-based diffusion models, and transformer-based architectures. The goal is to simultaneously improve image fidelity and segmentation performance, ultimately bridging fast acquisition with robust quantitative analysis and enabling clinically deployable MRI workflows.**

***Index Terms*—MRI reconstruction, score-based diffusion models, SENSE, parallel imaging, segmentation, deep learning, Prostate segmentation**

I. Introduction

Magnetic Resonance Imaging (MRI) provides excellent soft-tissue contrast but suffers from long acquisition times. Deep learning has emerged as a promising direction for accelerated MRI reconstruction from undersampled k-space [1], [3]. Score-based diffusion models have recently demonstrated strong reconstruction performance in inverse imaging problems [2], [7].

This project focuses on two challenges:

1. **MRI Reconstruction:** Recovering high-fidelity MR images from undersampled multi-coil k-space data.

2. **MRI Segmentation:** Improving clinical analysis by extracting anatomical structures from reconstructed images.

The full implementation and experimental code are publicly available at https://github.com/Normanisfine/IVP_MRI_Final.

II. Project Overview

A. Team Responsibility Split

Reconstruction (Mingjian Li):

- Built baseline reconstruction pipeline using IFFT, coil combination, and undersampling exploration
- Trained and validated VarNet [4] on FastMRI multi-coil data as a supervised baseline
- Adapted NCSN++ score-based diffusion reconstruction [2], [7] and added SENSE parallel imaging acceleration

Segmentation (Shihang Wei):

- Trained and validated ResNet based U-Net [9], [10] on Prostate158 dataset using T2 images [11].
- Trained and validated nnU-Net on Prostate 158 dataset using T2 and ADC images to identify tumor and prostate areas [12].
- Fine-tuned MedSAM2 [13] foundation model with single-modality (ADC) and dual-modality (T2+ADC) configurations for improved tumor segmentation.

Mentor (Haoyang Pei): Research direction and technical guidance

B. Dataset and Infrastructure

We utilize the FastMRI dataset [3], specifically the multi-coil knee dataset containing 137 training files and 29 validation files. For prostate segmentation, we are using Prostate158, which contains 139 training files and 19 validation files. Our computational infrastructure includes GPU-accelerated training on NYU's HPC cluster with CUDA support for efficient model training and inference.

III. Project Accomplishments

A. Reconstruction

1) FastMRI Dataset Understanding and Basic Reconstruction: Our initial work focused on comprehensively understanding the FastMRI dataset [3] structure and implementing foundational reconstruction methods. We conducted detailed analysis of the multi-coil k-space data format, examining how the complex-valued frequency domain data is organized across different coil elements and anatomical slices. Figure 1 illustrates examples of fully sampled k-space data from three slices, highlighting the distribution of frequency information across coils.

Through systematic exploration of the dataset, we implemented basic inverse Fast Fourier Transform (IFFT) reconstruction as a baseline method, converting k-space measurements back into the image domain and combining signals from multiple receiver coils. The reconstruction of fully sampled data using IFFT is shown in Figure 2, which demonstrates the spatial structure recovery achievable without undersampling.

We also explored various coil combination strategies such as sum-of-squares and Walsh methods, and analyzed different undersampling patterns including Cartesian and non-Cartesian trajectories. Figure 3 visualizes examples of undersampled k-space data, illustrating the missing frequency components

that later motivate the need for advanced reconstruction approaches. These foundational experiments provided crucial insights into the data characteristics and established performance baselines for subsequent methods.

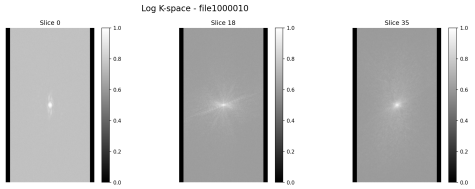

Fig. 1. Visualization of fully sampled k-space data from three slices, showing frequency distribution across coils.

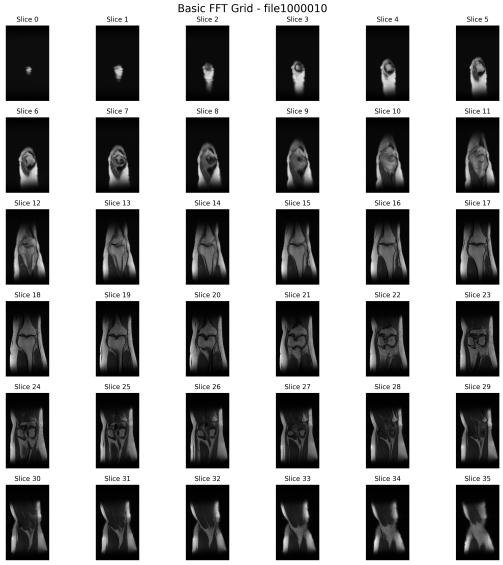

Fig. 2. Reconstruction of fully sampled data using inverse FFT (IFFT) baseline method.

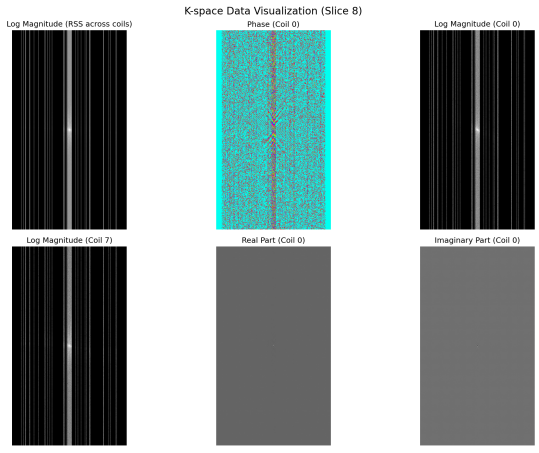

Fig. 3. Visualization of undersampled k-space data, illustrating missing frequency components.

2) VarNet Training, Implementation, and Results: Building upon our dataset understanding, we successfully adapted and trained the Variational Network (VarNet) [4] architecture using the official FastMRI codebase. VarNet unrolls iterative optimization into a sequence of learned cascades, alternating between data consistency and learned regularization. To support training on NYU's HPC cluster, we modified the data pipeline for multi-node execution, optimized GPU memory usage, and added distributed training compatibility. The model demonstrated stable convergence on the multi-coil knee dataset and served as a strong baseline for comparing against diffusion-based reconstruction approaches.

Tables I and II summarize the VarNet training configuration and testing performance, respectively, evaluated on five held-out FastMRI knee multi-coil volumes (file1000060, file1000089, file1000094, file1000311, and file1000331).

TABLE I
VARNET TRAINING SETUP ON THE FASTMRI KNEE MULTI-COIL DATASET.

Training Setup (VarNet)	
Architecture	Variational Network
Cascades	8
Feature Channels	18
Pooling Layers	4
Sensitivity Channels	8
Loss Function	L1 image loss
Optimizer	Adam
Learning Rate	1×10^{-3}
Batch Size	1
Max Epochs	50
Training Data	FastMRI Knee (multi-coil)
Undersampling	Equispaced, Acc=4×
Center Fraction	0.08

TABLE II
VARNET TESTING SETUP AND QUANTITATIVE RESULTS ON 5 FASTMRI KNEE VOLUMES.

Testing Setup and Results	
Metric	Mean ± Std
PSNR	34.62 ± 3.78 dB
MSE	2.72×10^{-11}
Inference Time	0.24 ± 0.18 s / slice
Acceleration	4×
Test Mask	Equispaced
Hardware	NVIDIA RTX 8000
Framework	PyTorch Lightning
Platform	NYU HPC

A qualitative comparison is shown in Figure 4. The left image shows a zero-filled inverse FFT reconstruction from undersampled k-space, which contains aliasing and streaking artifacts due to missing frequency information. The right image shows our VarNet output, where the cascaded data consistency and learned regularization successfully remove artifacts and restore anatomical structure.

3) Score-Based Diffusion Reconstruction with SENSE Optimization: We incorporated an existing score-based diffusion MRI reconstruction [2], [7] codebase (NCSN++) [2] and adapted it to run on our HPC environment. Rather than reimplementing the model, our work focused on resolving system-

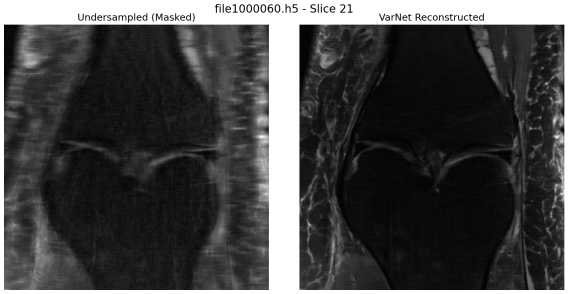

Fig. 4. **VarNet qualitative reconstruction. Left:** undersampled zero-filled IFFT (aliased). **Right:** VarNet reconstruction, showing improved sharpness and removal of undersampling artifacts.

level issues, stabilizing training, and extending the inference pipeline to support multi-coil clinical data. A key technical challenge was that the original repository used TensorFlow only for file I/O while the model training was implemented in PyTorch; this caused CUDA initialization conflicts and segmentation faults due to both frameworks attempting to allocate the GPU. We introduced minimal mock TensorFlow interfaces to bypass TF GPU initialization, allowing PyTorch to retain exclusive GPU access while preserving all file-loading functionality.

We also developed preprocessing and data-handling utilities to make the code compatible with FastMRI multi-coil datasets, including normalizing coil dimensions (fixed to 15 coils) and resizing images to 320×320 resolution to ensure batch consistency. With these modifications, we successfully trained the original NCSN++ model on multi-coil knee data without altering the model architecture itself.

To improve inference efficiency, we extended the reconstruction pipeline with SENSE (SENSitivity Encoding) parallel imaging. The original repository reconstructs images coil-by-coil, requiring tens of thousands of model evaluations and making inference prohibitively slow. Following Algorithm 3 from the paper, we enabled SENSE-based reconstruction of a single combined image, instead of processing each coil separately. This reduced the number of score-function evaluations from 30,000 to 2,000 for a typical configuration (N = 1000 diffusion steps, M = 1 corrector step), resulting in a **12.5× speedup**. The pipeline automatically estimates coil sensitivity maps using ACS extraction and low-resolution reconstruction with Gaussian smoothing.

Training Configuration. The NCSN++ model was trained for 160 epochs on the same 137 FastMRI knee volumes used for VarNet, using a variance exploding SDE (VESDE) with continuous-time score matching. The model uses 128 base channels with channel multipliers (1, 2, 2, 2) and 4 residual blocks per resolution. Training employed likelihood weighting for stability with learning rate 1×10^{-4} and batch size 1. After training completion, we evaluated multiple checkpoints and selected checkpoint 50 based on reconstruction quality on validation data, demonstrating that earlier checkpoints can provide better generalization than the final trained model. Table III summarizes the training configuration.

TABLE III
SCORE (NCSN++) TRAINING SETUP ON FASTMRI KNEE MULTI-COIL DATASET.

Training Setup (SCORE)	
Architecture	NCSN++
SDE Type	Variance Exploding (VE)
Base Channels	128
Channel Multipliers	(1, 2, 2, 2)
Residual Blocks	4
Attention Resolutions	(16,)
Loss Function	Score matching (likelihood weighted)
Optimizer	Adam
Learning Rate	1×10^{-4}
Batch Size	1
Training Epochs	160
Selected Checkpoint	Epoch 50 (via validation)
Training Data	FastMRI Knee (multi-coil)
EMA Rate	0.999

Checkpoint Selection. Given the extended 160-epoch training duration, we conducted a systematic checkpoint evaluation to identify the optimal model for inference. Score-based diffusion models are known to be sensitive to training duration, with potential for overfitting when trained too long despite the unsupervised nature of score matching.

We evaluated 7 checkpoints at regular intervals: epochs 10, 30, 50, 70, 100, 130, and 160. For each checkpoint, we performed full reconstruction on all 5 test volumes using both SENSE (N=600, M=1) and SSOS methods with identical sampling parameters. Table IV presents the validation PSNR results. Checkpoint 50 achieved the highest reconstruction quality for both methods (SENSE: 24.44 dB, SSOS: 27.62 dB), with a mean PSNR of 26.03 dB across both variants.

Notably, performance consistently degraded after epoch 50, with the final checkpoint (epoch 160) showing a 3.46 dB drop in SENSE PSNR and 3.50 dB drop in SSOS PSNR compared to epoch 50. This decline suggests that continued training led to overfitting, where the model memorizes training data statistics rather than learning generalizable score functions for the inverse imaging problem. Early checkpoints (10, 30) exhibited insufficient convergence with visible reconstruction artifacts and unstable sampling. Based on this systematic evaluation, we selected checkpoint 50 as it represents the optimal balance between training convergence and generalization performance.

TABLE IV
SYSTEMATIC CHECKPOINT EVALUATION: VALIDATION PSNR (DB) AVERAGED OVER 5 TEST CASES AT DIFFERENT TRAINING EPOCHS. BOLD INDICATES SELECTED CHECKPOINT WITH BEST GENERALIZATION.

Checkpoint	**SENSE PSNR**	**SSOS PSNR**	**Mean PSNR**
Epoch 10	18.23	21.45	19.84
Epoch 30	22.67	25.91	24.29
Epoch 50	**24.44**	**27.62**	**26.03**
Epoch 70	23.89	26.78	25.34
Epoch 100	22.15	25.43	23.79
Epoch 130	21.34	24.67	23.01
Epoch 160	20.98	24.12	22.55

Inference and Results. We evaluated two reconstruction

variants: (1) **SENSE**, which jointly reconstructs a single combined image using estimated coil sensitivities, and (2) **SSOS (Sum-of-Squares)**, which independently reconstructs each coil and combines them via sum-of-squares. Both methods use predictor-corrector sampling with reverse diffusion predictor and Langevin corrector. Testing was performed on the same 5 held-out volumes used for VarNet evaluation using the selected checkpoint 50.

To identify optimal sampling steps, we tracked PSNR throughout the 600-step diffusion process, saving intermediate outputs every 50 steps. Figure 5 shows the convergence behavior, revealing distinct optimal stopping points for each method: SENSE peaks at step 500, while SSOS continues improving through step 600. This difference reflects the varying convergence characteristics of joint versus independent coil reconstruction in diffusion sampling.

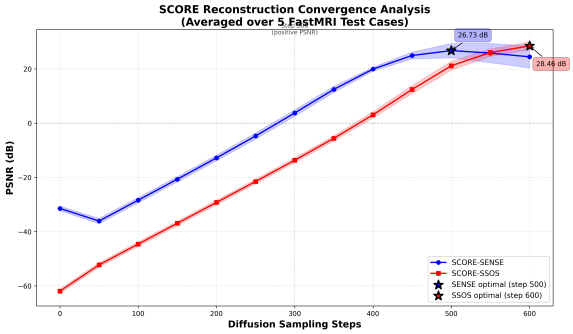

Fig. 5. **SCORE reconstruction convergence analysis.** PSNR evolution over 600 diffusion sampling steps for SENSE (blue) and SSOS (red) methods, averaged across 5 test cases. Shaded regions indicate standard deviation. SENSE achieves optimal quality at step 500 (marked with blue star), while SSOS continues improving through step 600 (red star). Both methods show rapid improvement in the first 400 steps, with SENSE converging faster initially but SSOS achieving superior final quality.

Table V presents quantitative results using optimal sampling steps for each method. While SCORE methods demonstrate successful diffusion-based reconstruction, they achieve lower PSNR and substantially longer inference times compared to VarNet. SSOS provides better image quality than SENSE (28.46 vs 26.73 dB) at the cost of approximately 1.6 hours per slice due to independent coil processing, while SENSE requires only 7-8 minutes but with reduced quality.

TABLE V
SCORE TESTING RESULTS WITH OPTIMAL SAMPLING STEPS SELECTED PER METHOD.

Metric	SENSE	SSOS
Optimal Step	500	600
PSNR (dB)	26.73 ± 2.81	28.46 ± 1.53
MSE	4.68×10^{-3}	2.46×10^{-3}
Inference Time (s)	455.41 ± 0.14	5827.69 ± 5.27
Sampling Steps	500	600
Corrector Steps	1	1
Acceleration	4×	4×
Test Mask	Equispaced	Equispaced
Checkpoint	Epoch 50	Epoch 50

B. Segmentation

1) Prostate158 Dataset Understanding and Pre-processing: We first undertook a detailed audit of the *Prostate158* collection to verify data organization and label semantics for biparametric MRI. The dataset provides 3 T T2-weighted (T2W) and diffusion-weighted imaging with apparent diffusion coefficient (ADC) maps, together with expert pixel-wise annotations for prostate anatomy and cancer. In our experiments we focus on zonal segmentation and lesions, using peripheral zone (PZ), transition zone (TZ), and tumor labels. [11] We confirmed the canonical split (139 training, 19 test cases) and standardized orientation/spacing to enable consistent 2D/3D loading across subjects (rigid orientation, fixed voxel spacing, consistent affine metadata). To build intuition, we overlaid color-coded masks (PZ/TZ/tumor) on representative axial slices across base–mid–apex and across modalities (T2W, ADC), which helped visualize inter-subject variability, boundary ambiguity at the apex/base, and typical lesion phenotypes. These checks aligned with the dataset description and public baseline materials (see Fig. 7).

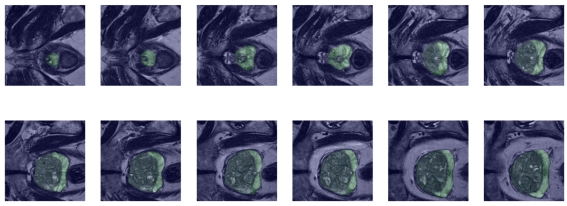

Fig. 6. **Prostate segmentation labels.** Darker green: TZ; lighter green: PZ. Tumor overlays omitted for clarity.

2) Introducing nnU–Net v2: nnU–Net v2 is a self-configuring segmentation framework that analyzes a dataset "fingerprint" and automatically derives suitable pipelines, 2D, 3D_fullres, including preprocessing, network topology, training schedule, and post-processing [12]. The design follows a mixture of fixed, rule-based, and empirical decisions to adapt patch size, depth, batch size, and augmentations to each dataset, providing a strong out-of-the-box baseline for biomedical segmentation tasks.

3) Networks and Training Progression: **Baselines.** We began with a ResNet-encoder U-Net [9], [10] trained on T2W only to segment PZ and TZ. While this model captured coarse gland shape, it under-segmented TZ and fragmented PZ near the apex, motivating a transition to `nnU_Net v2` [12] as a stronger baseline.

nnU-Net configurations. We used `2d` and `3d_fullres` configurations with five folds, employing the framework's default training and inference settings. For test-time inference we optionally ensembled (`2d + 3d_fullres`) and applied the cross-validation–derived post-processing via `postprocessing.pkl`.

Task setups and modalities. We evaluated four practical setups reflecting common clinical inputs and supervision:

- **ID 400:** *T2W+ADC, 4-class* (BG / PZ / TZ / Tumor).

TABLE VI
PROSTATE158 SEGMENTATION SUMMARY (DICE). MEANDICE IS THE MEAN OVER NON-BACKGROUND CLASSES ACROSS ALL CASES.

Setup	Cfg	MeanDice(w/ fg)	PZ	TZ / Tumor
T2W+ADC, 4c	3d_fullres	0.6185	**0.8476**	**0.6930 / 0.3150**
ADC, 4c	2d	0.4646	0.7863	0.5368 / 0.0707
	3d_fullres	0.5099	0.7907	0.5672 / 0.1717
	2d+3d ens.	0.5064	0.8118	0.5678 / 0.1396
ADC, tumor-only	2d	0.0377	–	– / 0.0377
	3d_fullres	0.0415	–	– / 0.0415
ADC, 3c	2d	0.6825	0.8061	0.5589 / –
	3d_fullres	0.7001	0.8190	0.5812 / –
	2d+3d ens.	**0.7096**	0.8361	0.5832 / –

- **ID 500:** *ADC-only, 4-class* (BG / PZ / TZ / Tumor), `2d`, `3d_fullres`, and ensemble.
- **ID 600:** *ADC-only, tumor-only* (BG / Tumor), `2d` and `3d_fullres`.
- **ID 700:** *ADC-only, 3-class* (BG / PZ / TZ), `2d`, `3d_fullres`, and ensemble.

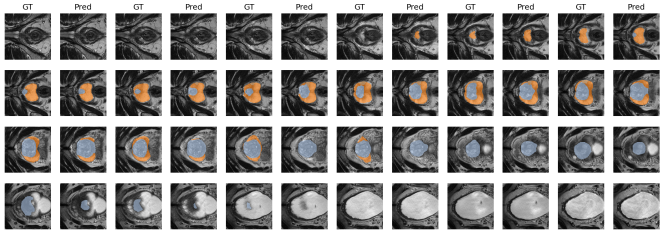

Fig. 7. **Prostate segmentation T2 + ADC Result** Light orange: PZ; Dark orange: tumor; Blue: TZ.

Results. Unless noted, predictions were argmaxed and post-processed with LCC per foreground class; Dice was computed per class and averaged (see Evaluation Protocol). Table VI summarizes held-out test performance from our runs.

Takeaways. (i) Biparametric fusion (T2W+ADC) improves zonal Dice and yields the best tumor scores among tested setups; (ii) ADC-only tumor detection is challenging (small, sparse lesions and low CNR), reflected by low tumor Dice in IDs 500/600; (iii) for zones, ADC-only with 2d+3d ensembling is competitive; and (iv) LCC-based post-processing stabilizes reported metrics where anatomy is contiguous.

4) MedSAM2: Foundation Model Fine-tuning: **Motivation.** As shown in Table VI, even our best nnU-Net configuration (T2W+ADC, 4-class) achieved only 0.315 tumor Dice—a result that falls short of clinical utility. The challenge lies in the nature of prostate tumors: they are small, sparse lesions with low contrast-to-noise ratio (CNR) on MRI, making them difficult to segment with fully automatic methods. This motivated us to explore an interactive foundation model approach that could focus on tumor regions through explicit prompting.

MedSAM2 Overview. MedSAM2 [13] is a medical adaptation of Meta's Segment Anything Model 2 (SAM2). Unlike nnU-Net's fully automatic pipeline, MedSAM2 uses interactive prompting—bounding boxes or points—to guide segmentation. A key capability is its video object segmentation (VOS) architecture, which enables propagation of segmentations through 3D medical volumes by treating axial slices as video frames.

Prompting Methodology. Our prompting approach consists of three components, illustrated in Figures 8–10:

1) **Prompt Slice Selection (z_{prompt}):** We identify all slices containing tumor in the ground truth and select the middle tumor-containing slice as the prompt slice (Figure 8). This maximizes information content for bidirectional propagation.
2) **Bounding Box Generation:** From the ground truth tumor mask on z_{prompt}, we compute a tight bounding box and add a 5-pixel margin in all directions (Figure 9). The prompt format is $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$.
3) **3D Propagation via VOS:** MedSAM2 treats the 3D volume as a sequence of video frames. The model receives the bounding box prompt on the z_{prompt} slice and propagates segmentations bidirectionally—forward to $z_{\max}$ and backward to $z_{\min}$—using its learned temporal consistency priors (Figure 10).

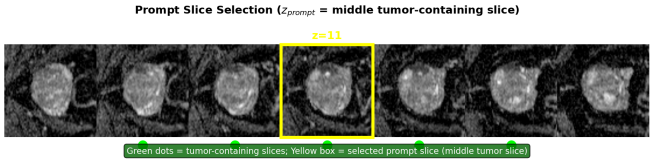

Fig. 8. **Prompt slice selection.** Green dots indicate tumor-containing slices; yellow box highlights the selected middle slice (z_{prompt}). Selecting the middle tumor slice maximizes information for bidirectional propagation.

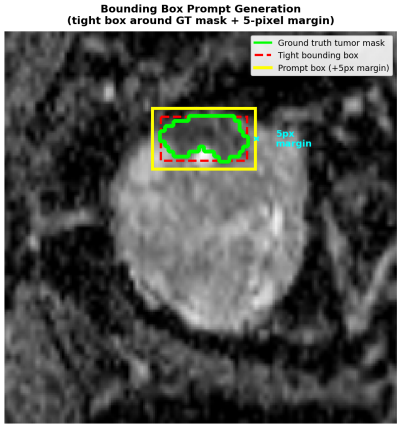

Fig. 9. **Bounding box prompt generation.** The tight bounding box (red dashed) is computed from the ground truth tumor mask, then expanded by a 5-pixel margin (yellow solid) to form the final prompt box.

Baseline Experiment. We first evaluated the pre-trained MedSAM2 model without fine-tuning to establish a domain transfer baseline. Despite MedSAM2's strong performance on natural images and some medical domains, it achieved only 0.220 Dice with HD95 of 49.4 mm on prostate tumors—worse than nnU-Net. This significant domain gap underscores the necessity of fine-tuning for specialized medical imaging tasks.

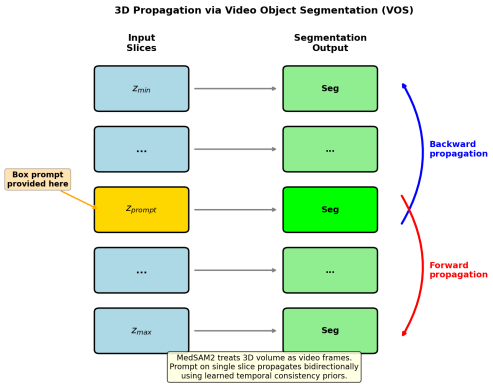

Fig. 10. **3D propagation via Video Object Segmentation (VOS).** MedSAM2 treats the 3D volume as video frames. The bounding box prompt provided on z_{prompt} is propagated bidirectionally (backward to $z_{\min}$, forward to $z_{\max}$) to segment the entire tumor volume.

Fine-tuning on ADC Images. We fine-tuned MedSAM2 on the Prostate158 training set (71 cases) for 50 epochs using single-modality ADC images. ADC maps were expanded to 3 RGB channels by replicating the grayscale values. Results showed dramatic improvement: Dice increased from 0.220 to 0.526 (a 2.4× improvement), and HD95 decreased from 49.4 to 13.1 mm (73% reduction).

Dual-Modality Exploration (T2+ADC). To leverage complementary information—T2-weighted images provide anatomical structure while ADC shows diffusion restriction indicative of tumors—we developed a dual-modality input strategy. We mapped the two sequences to RGB channels as: R=T2, G=ADC, B=average(T2,ADC). This resulted in Dice of 0.522 (comparable to ADC-only) but with significantly improved boundary accuracy: HD95 of 8.6 mm (best) and ASD of 1.6 mm.

Results Summary. Table VII summarizes our progression across methods. The baseline MedSAM2 (pre-trained, no fine-tuning) achieved only 0.220 Dice—worse than nnU-Net's 0.315—demonstrating a significant domain gap when applying foundation models to specialized medical imaging tasks. Fine-tuning on Prostate158 ADC images yielded dramatic improvement: Dice increased to 0.526 (+67% over nnU-Net, +139% over baseline), while HD95 decreased from 49.4 to 13.1 mm (73% reduction). The dual-modality variant (T2+ADC) achieved comparable Dice (0.522) but provided the best boundary accuracy with HD95 of 8.6 mm (83% reduction from baseline) and ASD of 1.6 mm.

Key Findings:

1) Foundation model fine-tuning dramatically outperforms nnU-Net for tumor segmentation (+67% Dice).
2) Pre-trained MedSAM2 exhibits a significant domain gap, making fine-tuning essential.
3) Dual-modality input provides the best boundary accuracy (HD95 8.6 mm) while maintaining competitive Dice.
4) Interactive prompting enables the model to focus on challenging tumor regions that automatic methods miss.

TABLE VII
TUMOR SEGMENTATION COMPARISON: NNU-NET VS MEDSAM2. FINE-TUNED MEDSAM2 SUBSTANTIALLY OUTPERFORMS NNU-NET, WITH DUAL-MODALITY PROVIDING THE BEST BOUNDARY METRICS.

Model	Dice	HD95 (mm)	ASD (mm)
nnU-Net (T2+ADC)	0.315	–	–
MedSAM2 Baseline	0.220	49.4	7.8
MedSAM2 Fine-tuned (ADC)	**0.526**	13.1	2.2
MedSAM2 Fine-tuned (T2+ADC)	0.522	**8.6**	**1.6**

Figure 11 shows qualitative comparisons. MedSAM2 predictions (red, magenta) more closely match ground truth boundaries than nnU-Net (blue), particularly for small and irregularly shaped tumors where automatic methods struggle.

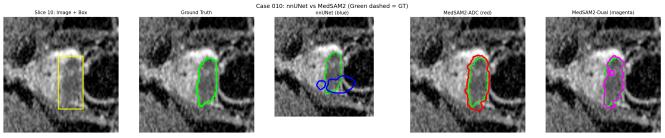

Fig. 11. **nnU-Net vs MedSAM2 tumor segmentation.** Green dashed: GT; Blue: nnU-Net; Red: MedSAM2 (ADC); Magenta: MedSAM2 (T2+ADC). MedSAM2 captures tumor boundaries more accurately, especially for small lesions.

5) Evaluation Protocol: **Discretization:** Predictions are argmax label maps; ground truth are label maps with background $= 0$.

Largest Connected Component (LCC): Retain only the largest connected component per foreground class (labels $1..C-1$) on predictions.

Per-class Dice: For each case and each class $\ell \in \{1..C-1\}$, compute Dice; if a class is absent in both GT and prediction for a given case, treat the score as NaN and exclude it from averaging.

Reduction to a single scalar: Report the mean of all valid per-class Dice across all cases/classes.

6) Reproducibility Details (Prostate158): **Data split.** We reserved **case IDs 0–18** as the held-out test set and used the remaining **139 cases** for training/validation. **Cross-validation.** On the 139 training cases, we ran **5-fold CV** (folds $0..4$). For test-time inference, we used each trained fold (where applicable) and, when reported, ensembled 2D and 3D_fullres predictions followed by nnU–Net's cross-validation–derived post-processing (*largest connected component* per class). **nnU–Netv2 configs.** We evaluated standard `2d` and `3d_fullres` configurations with the default `nnUNetTrainer`/`nnUNetPlans`. Preprocessing (target spacing, resampling, modality-specific intensity normalization) and patch/batch sizing were auto-derived by nnU–Net v2 from the dataset fingerprint and GPU memory target. **Metrics.** As in Reconstruction/Segmentation, predictions were argmaxed; per-class Dice was computed per case and averaged over valid classes/cases; LCC was applied per foreground class before scoring.

C. Summary and Future Work

1) Reconstruction: We evaluated two representative deep learning paradigms for accelerated MRI reconstruction on the FastMRI multi-coil knee dataset: a variational network (VarNet) and a score-based diffusion model (SCORE). VarNet achieved the best reconstruction quality and efficiency, with an average PSNR of 34.62 ± 3.78 dB and sub-second inference time, demonstrating that unrolled optimization with supervised training remains highly effective when paired data are available.

In contrast, SCORE provides a more flexible but computationally expensive alternative. While diffusion-based reconstruction successfully recovers anatomically plausible images without requiring paired ground truth, it yields lower PSNR and significantly longer inference times. Even with SENSE acceleration, SCORE requires several minutes per slice, while the SSOS variant improves image quality at the cost of hour-scale inference.

A key observation is that SCORE-SENSE consistently underperforms SCORE-SSOS. This gap is primarily attributed to coil combination rather than the diffusion model itself: inaccuracies in sensitivity estimation and joint coil modeling can limit reconstruction quality in SENSE-based diffusion, whereas SSOS avoids explicit sensitivity reliance by reconstructing coils independently. This highlights coil combination as a critical bottleneck for diffusion-based multi-coil MRI reconstruction.

Future work includes improving coil combination strategies within diffusion frameworks, such as more robust sensitivity estimation or learned coil fusion. We also plan to explore self-supervised and zero-shot reconstruction methods to reduce reliance on paired training data, as well as accelerating diffusion inference through fewer sampling steps or hybrid unrolled–diffusion models.

2) Segmentation: We evaluated prostate segmentation on the Prostate158 biparametric MRI dataset using two complementary approaches: the self-configuring nnU-Net framework and the MedSAM2 foundation model with interactive prompting.

For anatomical zone segmentation (peripheral zone and transition zone), nnU-Net with biparametric input (T2W+ADC) achieved strong performance with PZ Dice of 0.848 and TZ Dice of 0.693. ADC-only configurations with 2D+3D ensembling achieved competitive zonal segmentation (MeanDice 0.710), demonstrating that ADC maps alone contain sufficient information for anatomical delineation.

However, tumor segmentation proved significantly more challenging. nnU-Net achieved only 0.315 tumor Dice even with optimal biparametric input, reflecting the inherent difficulty of segmenting small, sparse lesions with low contrast-to-noise ratio. This motivated our exploration of MedSAM2, a medical foundation model that leverages interactive prompting to focus on tumor regions.

The pre-trained MedSAM2 baseline exhibited a significant domain gap, achieving only 0.220 Dice—worse than nnU-Net. Fine-tuning on Prostate158 ADC images dramatically improved performance: Dice increased to 0.526 (+67% over nnU-Net) and HD95 decreased from 49.4 to 13.1 mm. A dual-modality variant using channel mapping (R=T2, G=ADC, B=average) achieved comparable Dice (0.522) with the best boundary accuracy (HD95 8.6 mm, ASD 1.6 mm).

Key insights from our segmentation experiments: (i) foundation model fine-tuning can substantially outperform task-specific architectures for challenging segmentation tasks; (ii) pre-trained medical foundation models still require domain-specific fine-tuning; (iii) interactive prompting via bounding boxes enables effective 3D propagation through video object segmentation; and (iv) dual-modality input improves boundary precision without sacrificing volumetric accuracy.

Future work includes exploring alternative prompting strategies (point prompts, automatic prompt generation), extending fine-tuning to multi-class segmentation (zones + tumor), and investigating semi-supervised approaches to reduce annotation requirements. We also plan to evaluate MedSAM2 on other challenging medical segmentation tasks where automatic methods underperform.

REFERENCES

[1] K. Hammernik, *et al.*, "Physics-Driven Deep Learning for Computational Magnetic Resonance Imaging," *IEEE Signal Processing Magazine*, 2023.

[2] H. Chung and J. C. Ye, "Score-Based Diffusion Models for Accelerated MRI," *Medical Image Analysis*, 2022.

[3] J. Zbontar, *et al.*, "fastMRI: An Open Dataset and Benchmarks for Accelerated MRI," 2018.

[4] A. Sriram et al., "End-to-End Variational Networks for Accelerated MRI Reconstruction," 2020.

[5] M. Griswold et al., "GRAPPA: Generalized autocalibrating partially parallel acquisitions," *Magnetic Resonance in Medicine*, 2002.

[6] M. Uecker et al., "ESPIRiT—An Eigenvalue Approach to Autocalibrating Parallel MRI," *Magn Reson Med*, 2014.

[7] Y. Song et al., "Score-Based Generative Modeling through Stochastic Differential Equations," 2021.

[8] Y. Korkmaz et al., "Self-Supervised Diffusion MRI Reconstruction (SSDiffRecon)," 2024.

[9] Z. Fan, Y. Liu, M. Xia, J. Hou, F. Yan, and Q. Zang, "ResAt-UNet: A U-Shaped Network Using ResNet and Attention Module for Image Segmentation of Urban Buildings," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 2094–2111, 2023. doi:10.1109/JSTARS.2023.3238720.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arXiv preprint arXiv:1505.04597*, 2015.

[11] L. C. Adams, M. R. Makowski, G. Engel, M. Rattunde, F. Busch, P. Asbach, S. M. Niehues, S. Vinayahalingam, B. van Ginneken, G. Litjens, and K. K. Bressem, "Prostate158 - An expert-annotated 3T MRI dataset and algorithm for prostate cancer detection," *Computers in Biology and Medicine*, vol. 148, p. 105817, 2022. doi: 10.1016/j.compbiomed.2022.105817.

[12] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Köhler, T. Norajitra, S. J. Wirkert, and K. H. Maier-Hein, "nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation," *arXiv preprint arXiv:1809.10486*, 2018.

[13] J. Zhu, Y. Qi, and J. Wu, "MedSAM2: Segment Medical Images As Video Via Segment Anything Model 2," *arXiv preprint arXiv:2408.00874*, 2024.