From Minimal Clinical Prompts to 3D: Spacing-Aware Prompt Propagation for Multimodal Prostate Lesion Segmentation in bpMRI

Jiacheng Wang[1,2], Heinrich von Busch[3], Robert Grimm[3], Ipek Oguz[1], Dorin Comaniciu[2], Ali Kamen[2], and Bin Lou[2⋆]

[1] College of Connected Computing, Vanderbilt University, Nashville, TN, USA
{jiacheng.wang.1, ipek.oguz}@vanderbilt.edu
[2] Digital Technology and Innovation, Siemens Healthineers, Princeton, NJ, USA
{dorin.comaniciu, ali.kamen, bin.lou}@siemens-healthineers.com
[3] Diagnostic Imaging, Siemens Healthineers AG, Erlangen, Bavaria, Germany
{heinrich.von_busch, robertgrimm}@siemens-healthineers.com

Abstract. Accurate 3D delineation of prostate cancer lesions in biparametric MRI (bpMRI) supports targeted biopsy and therapy planning, yet clinical reports typically provide only sparse 2D measurements or a single contour. We present SAPP (**S**pacing-**A**ware **P**rompt **P**ropagation), a prompt-to-volume adaptation of SAM 2 that converts one clinical prompt on a single slice (diameter, box, circle, or contour) into a coherent 3D lesion mask from multimodal bpMRI. SAPP couples adaptive multimodal fusion for accurate *image-level* prompted segmentation with a spacing-aware *volume-level* prompt propagation module. The propagation uses spacing-decayed, confidence-gated streaming-memory attention and a learned stop rule to prevent leakage in anisotropic volumes. Trained on 3,256 scans from 7 institutions, SAPP generalizes to 14 held-out cohorts (3,504 fully annotated scans plus 516 routine-care weak-label studies), achieving 0.84/0.81 DSC on the prompted slice (box/diameter) and 0.86/0.78/0.76 DSC for volumetric masks (oracle-mask/box/diameter), consistently outperforming baselines while reducing over-propagation. Beyond segmentation, SAPP can accelerate annotation workflows and enable scalable generation of high-quality volumetric lesion masks, supporting the development of robust clinical AI models.

Keywords: Prostate MRI · Promptable foundation models · Multimodal fusion · Spacing-aware propagation · Weak supervision

1 Introduction

Prostate cancer (PCa) is a major global health burden and a leading cause of cancer morbidity [23]. Prostate MRI has become central to clinical workflows

⋆ Corresponding author.

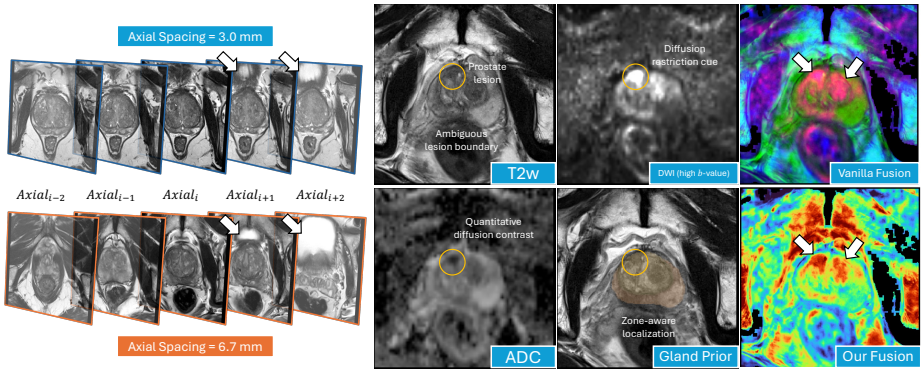

Fig. 1: Left: bpMRI volumes are often anisotropic; through-plane spacing varies across protocols (3.0 vs. 6.7 mm), so uniform slice-to-slice propagation can leak into lesion-absent slices. Right: lesion evidence is complementary across modalities (T2w, diffusion cues, gland priors), motivating multimodal fusion. As shown, vanilla fusion loses lesion specificity, whereas our approach yields a more lesion-focused fused representation.

because it improves lesion localization and enables targeted biopsy and treatment planning [2, 24]. Delineating the dominant lesion in 3D further supports biopsy targeting, focal therapy planning, and longitudinal assessment. However, accurate 3D PCa lesion segmentation remains challenging: lesion boundaries are often subtle or partially indistinct, and reliable interpretation typically requires simultaneously considering complementary cues across multiple MRI modalities rather than relying on any single modality [24]. Figure 1 highlights two practical obstacles that frequently break naive adaptations of promptable models to bpMRI: lesion evidence is distributed across modalities, and axial slice spacing varies substantially across protocols.

Full volumetric lesion annotation is therefore expensive and subjective: manual contouring is time-consuming and exhibits substantial inter-rater variability [4, 12]. As a result, large clinical archives often contain partial supervision, such as a single-slice diameter, a 2D contour on the most suspicious slice, or a coarse region-of-interest, instead of complete 3D masks. A practical system should leverage these routine annotations to recover coherent 3D masks with minimal effort; such masks can in turn supervise downstream models, e.g., lesion-mask priors improve PI-RADS scoring on external cohorts [34].

Promptable segmentation foundation models offer a natural mechanism to convert partial labels into dense masks. Segment Anything Model (SAM) produces strong 2D segmentation from simple prompts and generalizes across many natural image distributions [13], and SAM 2 extends this interface to images and sequences via streaming memory and prompt-based mask propagation [21]. Recent medical adaptations improve domain transfer and interaction efficiency [14, 18, 29], but most remain primarily image-centric or do not explicitly address volumetric anisotropy in clinical MRI. Applying promptable models to prostate bpMRI raises three intertwined challenges.

Challenge 1: multimodal fusion under heterogeneity. bpMRI evidence is distributed across modalities (T2w, ADC, high-b DWI) with different contrast mechanisms and failure modes. Previous work on incomplete multimodal learning and missing-modality robustness [9, 17, 25, 33], probabilistic distillation for missing modalities [5], and multimodal medical pre-training [22] highlights the value of cross-modality correlations. However, these fusion strategies are typically designed for task-specific architectures and do not directly yield SAM-compatible patch tokens under prompt conditioning, motivating a lightweight fusion mechanism tailored to SAM 2.

Challenge 2: lifting sparse 2D prompts to volumetric masks. Routine supervision is often a diameter/ROI or a single contour rather than full 3D masks. Interactive/prompt-based models can reduce annotation burden [15, 28, 30], and recent efforts adapt SAM to 3D medical data (e.g., SD-Trans in Med-SA [29] or treating volumes as sequences via SAM 2 propagation [21]). Yet reliable prompt-to-volume completion remains challenging at lesion boundaries and under domain shift, especially across diverse scanners and protocols. Prior promptable 3D methods either propagate 2D mask/box prompts on generic volumes (e.g., PAM [6], Sli2Vol+ [3]) or rely on broad CT-oriented spatial/class prompts (e.g., SegVol [8], VISTA3D [10]), and do not target the heterogeneous sparse 2D clinical prompts and multimodal anisotropic geometry of prostate bpMRI.

Challenge 3: anisotropic geometry breaks uniform propagation assumptions. Memory-based propagation is a powerful paradigm in video object segmentation (e.g., AOT [31], DeAOT [32], XMem [7]) and is core to SAM 2's streaming memory design [21]. Unlike videos, bpMRI volumes are often anisotropic with variable through-plane spacing, where adjacent slice indices can correspond to large physical gaps. Naively propagating over slice index can therefore leak into lesion-absent tissue, motivating propagation that explicitly accounts for physical distance and learns when to stop.

Our approach. We present SAPP, a spacing-aware prompt propagation framework for multimodal prostate lesion segmentation in bpMRI. At inference, SAPP follows a two-step pipeline: (i) *image-level* prompted segmentation on the prompted slice using SAM 2 in image mode, and (ii) bidirectional *volume-level* prompt propagation over axial slices using SAM 2 streaming memory. The method is built around two reusable components: an adaptive multimodal fusion module (Sec. 2.3) used in both steps, and a spacing-aware propagation module (Sec. 2.4) that modulates memory attention by physical spacing, applies confidence gating, and predicts a stop signal.

We train on 3,256 scans from 7 institutions and evaluate on 14 held-out cohorts, including 11 cohorts with full 3D masks (3,504 scans) and 3 routine-care cohorts with weak clinical annotations (516 scans), demonstrating robust generalization across scanners, protocols, and annotation formats. Our contributions are threefold:

- We formulate prompt-to-volume PCa lesion segmentation in bpMRI and propose SAPP, combining SAM2-image prompted segmentation with SAM 2 streaming-memory volume completion under a unified prompt interface.

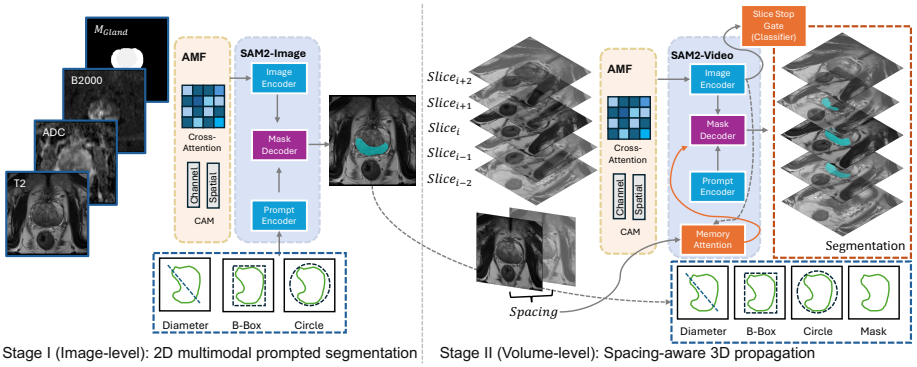

Fig. 2: Overview of SAPP. Image-level: given a single-slice clinical prompt (diameter/box/circle), SAPP predicts the prompted-slice mask using SAM 2 in image mode. **Volume-level:** the mask is then lifted to a coherent 3D lesion segmentation via bidirectional prompt propagation along axial slices using SAM 2 streaming memory. **Module 1 AMF** (light orange) performs adaptive multimodal fusion and is reused in both steps; **Module 2** (dark orange) augments propagation with spacing-aware memory modulation and a slice stop gate.

- We introduce two reusable modules—adaptive multimodal fusion and spacing-aware propagation—that extend promptable foundation models to multimodal MRI and reduce leakage in anisotropic volumes.
- We provide large-scale multi-center evaluation on 14 held-out cohorts (11 full-mask + 3 weak-label), showing consistent gains over strong specialist and promptable baselines under clinically realistic prompts.

2 Methods

2.1 Problem formulation

Each case contains a co-registered multimodal volume $\mathbf{X} = \{\mathbf{X}^{(m)}\}_{m\in\mathcal{M}}$ with $\mathbf{X}^{(m)} \in \mathbb{R}^{H\times W\times N}$, where $\mathcal{M}$ denotes the available modality (T2w, ADC, high-b DWI). If unavailable from the clinical site, the gland prior is predicted from the T2W image [27]. Given a prompt p on a single slice s, our goal is to predict a volumetric lesion mask $\hat{\mathbf{Y}} \in \{0,1\}^{H\times W\times N}$. When multiple lesions are present, each is treated as an independent object.

We support four clinically common prompt types that reflect different effort–accuracy trade-offs in routine reporting: **(i) diameter** — two endpoints marking the longest lesion axis, **(ii) box** — an axis-aligned bounding box, **(iii) circle** — a circular region-of-interest, and **(iv) single-slice mask** — a full 2D contour on one slice. Each prompt type is mapped to the SAM 2 prompt encoder Π: diameter endpoints are densified into evenly spaced positive points along the segment (acting as a scribble-like cue, inspired by [28]), circles are rasterized into binary masks, boxes are passed directly, and single-slice masks serve as

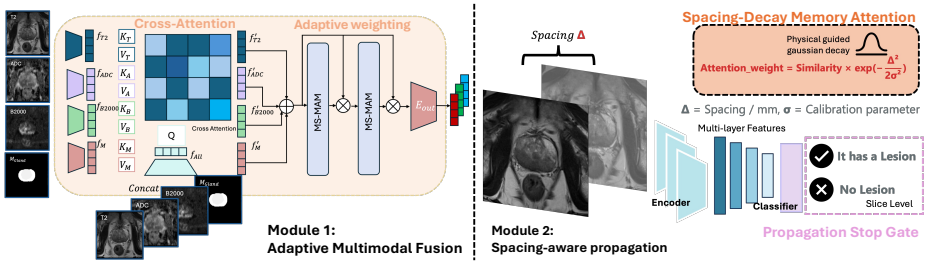

Fig. 3: Key components of SAPP. Module 1 (Adaptive multimodal fusion): modality tokens (T2w/ADC/high-b DWI, optionally a gland prior [27]) query a global context via cross-attention to produce SAM 2-compatible fused tokens for prompted segmentation. Module 2 (Spacing-aware propagation): streaming-memory attention is modulated by physical inter-slice distance Δz via a Gaussian decay (length-scale σ_z) and a slice-level stop classifier to reduce over-propagation.

dense mask prompts. Additional prompt construction details are provided in Sec. S2 (Fig. S1). The resulting embedding $\mathbf{e}_p = \Pi(p)$ conditions the image-level prediction on slice s, which is then propagated volumetrically.

2.2 Overview

Figure 2 summarizes SAPP. We build on SAM 2's prompted segmentation interface and its streaming-memory propagation mechanism [21]. Our method follows a two-stage design: (1) we predict slice-level mask on the prompted slice using SAM 2 image mode, and (2) we lift this prediction to a coherent 3D volume-level lesion mask by bidirectional volume-level propagation using SAM 2 video mode applied to the axial slice modalities. Crucially, the two key medical-specific modules are illustrated in Fig. 3. We detail the training objective and parameter-efficient adaptation in Sec. 2.5.

2.3 Module 1: Adaptive multimodal fusion (AMF) for bpMRI

Motivation. SAM 2 offers a strong promptable segmentation backbone pretrained at scale on RGB images/videos [21], but bpMRI violates its 3-channel input assumptions. Naively squeezing multimodal bpMRI into an RGB interface (e.g., stacking/averaging) can entangle modality-specific cues and effectively caps the number of usable inputs [34]. We implement Module 1 (Fig. 3, left) as a fusion block that maps the available modalities into SAM 2, enabling us to reuse the frozen pretrained encoder for both image-level and volume-level segmentation.

Tokenization and global context. For each slice n, we encode every available modality through a lightweight CNN stem ϕ_m consisting of a strided convolution (kernel size matching the encoder patch size), layer normalization, and GELU activation $\mathbf{T}_n^{(m)} = \phi_m\big(\mathbf{x}_n^{(m)}\big) \in \mathbb{R}^{P\times C}$, where P is the number of spatial tokens (matching the ViT patch grid) and C is the token dimension. We simultaneously compute a global context representation from the channel-concatenated

multimodal input using a shared stem ϕ_{ctx}:

$$\mathbf{T}_n^{\text{ctx}} = \phi_{\text{ctx}}\big(\text{concat}_{m\in\mathcal{M}}\,\mathbf{x}_n^{(m)}\big) \in \mathbb{R}^{P\times C}. \tag{1}$$

Context-injected fusion tokens. We enrich each modality with information from the global context via multi-head cross-attention (each modality queries the global context):

$$\tilde{\mathbf{T}}_n^{(m)} = \mathbf{T}_n^{(m)} + \text{MHCA}\big(\mathbf{Q}(\mathbf{T}_n^{(m)}), \mathbf{K}(\mathbf{T}_n^{\text{ctx}}), \mathbf{V}(\mathbf{T}_n^{\text{ctx}})\big), \tag{2}$$

and compute modality weights through a channel-wise gating function,

$$\boldsymbol{\omega}_n^{(m)} = \text{sigmoid}\left(\mathbf{W}_\omega\,\text{GAP}\big(\tilde{\mathbf{T}}_n^{(m)}\big) + \mathbf{b}_\omega\right) \in (0,1)^C, \tag{3}$$

where GAP is global average pooling over tokens and $\mathbf{W}_\omega \in \mathbb{R}^{C\times C}, \mathbf{b}_\omega \in \mathbb{R}^C$ maps pooled tokens to the weight vector ω. The fused token representation is $\mathbf{T}_n^{\text{fuse}} = \sum_{m\in\mathcal{M}} \boldsymbol{\omega}_n^{(m)} \odot \tilde{\mathbf{T}}_n^{(m)}$.

Rather than replacing SAM 2's pretrained patch projection, we map $\mathbf{T}_s^{\text{fuse}}$ to a three-channel spatial feature map $\mathbf{I}_{\text{fuse}} \in \mathbb{R}^{3\times H\times W}$ via a lightweight projection head; this map is passed through SAM 2's native patch embedding convolution, preserving the pretrained positional encodings and early attention patterns. AMF is applied to every slice and shared across stage I (Image-level) and stage II (Volume-level). To handle missing or corrupted modalities, we train it with modality dropout (Sec. 2.5).

2.4 Module 2: Spacing-aware propagation for volumetric completion

Propagation as a memory-conditioned sequence model. We treat the axial stack as an ordered sequence of slices with physical positions $\{z_n\}_{n=1}^N$ obtained from DICOM metadata. Starting from the prompted slice s, we perform forward and backward propagation. At slice t, SAM 2 predicts a mask $\hat{\mathbf{m}}_t$ conditioned on image features and a memory bank $\mathcal{B}_t$ of key-value embeddings computed from previously segmented slices. We augment memory attention leveraging physical slice spacing. (Fig. 3 right).

Spacing-aware attention decay. Standard video propagation weights memory entries largely by appearance similarity. In MRI, slice index difference is a poor proxy for geometric proximity; we therefore modulate memory utilization by physical distance $\Delta z_{t,j} = |z_t - z_j|$ between the current slice t and a memory slice j. We use a Gaussian kernel with a global length-scale σ_z (mm):

$$w_{\text{geo}}(t,j) = \exp\left(-\frac{\Delta z_{t,j}^2}{2\sigma_z^2}\right), \qquad \sigma_z = 3.3 \text{ mm}. \tag{4}$$

We set σ_z to the median through-plane spacing in our cohort and report a sensitivity analysis over σ_z in Sec. S5 (Tab. S4).

We additionally down-weight unreliable memory entries using a confidence gate based on SAM 2 confidence (predicted IoU $\hat{u}_j$ and optional objectness o_j)

and our slice gate g_j: $w_{\text{conf}}(j) = \text{sigmoid}(\gamma \hat{u}_j + \beta) \cdot \text{sigmoid}(o_j) \cdot g_j$, where γ, β are learnable scalars trained in Stage II (shared across layers/heads). We combine them as $w(t, j) = w_{\text{geo}}(t, j)\, w_{\text{conf}}(j)$.

We implement logit-level gating by adding a slice-wise bias to memory cross-attention. For a current token u and a memory token v from memory slice j,

$$\ell_{t,u,(j,v)} = \frac{q_{t,u}^{\top} k_{j,v}}{\sqrt{d_k}} + \log w(t, j), \qquad a_{t,u,(j,v)} = \text{softmax}_{(j,v)}\big(\ell_{t,u,(j,v)}\big), \tag{5}$$

where d_k is the key/query dimension.

Soft-to-hard stop gate. Even with spacing-aware attention, propagation can drift into healthy tissue once the lesion disappears. We introduce a slice-level lesion gate that predicts whether slice t contains any lesion content. Let $\{\mathbf{F}_t^{(\ell)}\}_{\ell \in \mathcal{L}}$ denote hierarchical encoder features from several depths. We upsample them to a common resolution, concatenate, and apply a lightweight classifier:

$$g_t = \text{sigmoid}\left(C\big(\text{GAP}(\text{concat}_{\ell \in \mathcal{L}}\, \text{Up}(\mathbf{F}_t^{(\ell)}))\big)\right) \in [0, 1]. \tag{6}$$

During inference, if $g_t < \tau$ we set $\hat{\mathbf{m}}_t = \mathbf{0}$, skip writing this slice into memory, and terminate propagation in that direction.

2.5 Training objectives and optimization

Parameter-efficient adaptation (LoRA). We keep SAM 2's pretrained image encoder frozen and adapt the model by training only lightweight components: the prompt encoder, mask decoder, memory-attention layers (video mode), and our proposed modules. In addition, we inject low-rank adapters (LoRA) into selected attention/MLP projections of the frozen encoder. For a linear projection with frozen weight $\mathbf{W}_0$, LoRA learns a rank-r update $\Delta\mathbf{W} = \frac{\alpha}{r}\mathbf{B}\mathbf{A}$ (only $\mathbf{A}, \mathbf{B}$ are trainable), yielding $\mathbf{h} = (\mathbf{W}_0 + \Delta\mathbf{W})\mathbf{x}$, where $\mathbf{A} \in \mathbb{R}^{r \times k}$ and $\mathbf{B} \in \mathbb{R}^{d \times r}$, with $r \ll \min(d, k)$.

Losses. Let $\hat{\mathbf{m}}_t \in [0, 1]^{H \times W}$ and $\mathbf{m}_t \in \{0, 1\}^{H \times W}$ denote the predicted and ground-truth masks for slice t. Define a binary label $y_t = \mathbb{I}(\|\mathbf{m}_t\|_1 > 0)$ and the positive/negative slice sets $\mathcal{T}^+ = \{t \in \mathcal{T} : y_t = 1\}$ and $\mathcal{T}^- = \{t \in \mathcal{T} : y_t = 0\}$. We optimize Dice+BCE on positive slices,

$$\mathcal{L}_{\text{seg}} = \frac{1}{|\mathcal{T}^+|} \sum_{t \in \mathcal{T}^+} \Big(\mathcal{L}_{\text{Dice}}(\hat{\mathbf{m}}_t, \mathbf{m}_t) + \lambda_{\text{BCE}} \mathcal{L}_{\text{BCE}}(\hat{\mathbf{m}}_t, \mathbf{m}_t)\Big), \tag{7}$$

We train the stop gate with BCE over all slices and suppress leakage on negative slices using a soft-gated foreground penalty:

$$\mathcal{L}_{\text{gate}} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \mathcal{L}_{\text{BCE}}(g_t, y_t), \qquad \mathcal{L}_{\text{neg}} = \frac{1}{|\mathcal{T}^-|} \sum_{t \in \mathcal{T}^-} \big\|g_t\, \hat{\mathbf{m}}_t\big\|_1. \tag{8}$$

The overall objective is $\mathcal{L} = \mathcal{L}_{\text{seg}} + \lambda_{\text{gate}} \mathcal{L}_{\text{gate}} + \lambda_{\text{neg}} \mathcal{L}_{\text{neg}}$.

Training robustness. We apply stochastic modality dropout (randomly masking one or more input modalities) and simulate clinical prompts from full masks with structured noise (endpoint jitter, box slack, mask dilation/erosion), forcing the model to remain robust to incomplete inputs and imprecise annotations encountered in practice. Full augmentation details are detailed in Tab. S1.

3 Datasets and Experiments

3.1 Dataset and Annotation

The training dataset consists of 3,256 prostate bpMRI examinations collected from seven clinical institutions (Private-1 to Private-7). Each study includes axial T2w, high b-value DWI, and ADC modalities. T2w and DWI are acquired in the same examination and rigidly registered with visual quality control. After registration and preprocessing, the dataset contains 145,974 2D slices used to train the multimodal model.

To evaluate generalization, we benchmark the model on an independent test set of 3,504 studies (228,531 2D slices) collected from 11 institutions, including eight private clinical cohorts (Private-8 to Private-15) and three public datasets (ProstateX, Prostate158, and TCIA [1,16,20]). In addition, we curate 516 studies (23,682 slices) from three routine-care cohorts derived from real-world clinical practice, where lesion annotations are available for both standard clinical measurements (lesion diameters) and curated 2D lesion contours.

Annotations were produced in two stages: routine clinical 2D annotation (typically a single-slice contour or longest-axis diameter on the maximal-lesion slice) was extended through the volume by a trained annotation team into full 3D masks, reviewed by a fellowship-trained prostate-MRI radiologist. The 2D prompts in our experiments are thus grounded in real clinical annotations, while the 3D masks serve as expert-reviewed references for prompt-to-volume evaluation. For a unified protocol across sites with differing raw prompt formats, we standardize prompts into common diameter/box/circle/oracle-mask formats derived from the finalized masks. Lesions with PI-RADS ≥ 3 were labeled as positive. Table 1 summarizes the dataset splits. Additional details on the datasets and preprocessing pipeline, including a per-site inventory (data size, annotation type, and spacing statistics), are provided in the Supplementary. All datasets were retrospectively collected, de-identified before analysis, and used under site-specific governance, including local IRB approvals or waivers where applicable.

Table 1: Overview of multi-center datasets. Train and primary test sets contain full 3D lesion masks; routine-care cohorts provide only standard clinical annotations (2D contours or lesion diameters).

Split	Institutions	#Scans	#Images	Label type
Train	7	3,256	145,974	Full 3D masks
Test	11	3,504	228,531	Full 3D masks
Test (2D)	3	516	23,682	2D contour / diameter

3.2 Experimental Setup

Prompt regimes. We evaluate prompts under two regimes:

1. Image-level: On cohorts with full 3D masks, we evaluate on all axial slices that contain a lesion and generate three routine geometric prompts from the slice's 2D ground truth: **(i) diameter** as the endpoints of the longest in-mask chord (maximum-length line segment fully contained in the lesion), **(ii) box** as the tight axis-aligned bounding box, and **(iii) circle** as the lesion's circumscribed (minimum enclosing) circle. Bounded perturbations are applied to mimic clinical imprecision (Table S1). On routine-care cohorts, we use the site-provided clinical prompt directly.
2. Volume-level: For propagation, we focus on the largest 3D connected component in the ground truth. All slice-k prompts (diameter/box/circle) are generated from the 2D mask of this component on slice k. We additionally report an **(iv) oracle-mask** upper bound, where the prompt is the unperturbed 2D ground-truth mask on slice k.

Evaluation protocol and metric. Following prior prostate SAM adaptations, we use Dice similarity coefficient (DSC) as the primary metric [34]. For *prompted-slice* evaluation, we compute **2D DSC** on slice k only and report mean±std across lesions per cohort. For *prompt-to-volume* evaluation, we run bidirectional propagation from slice k and compute **3D DSC** between the full predicted volume and the 3D ground-truth mask, reported as mean±std per cohort. Predictions are not clamped to the prompt region, reflecting real clinical prompts that can be imprecise. For methods that output multiple candidate masks, we select the one with the highest model-predicted IoU (no ground-truth-based selection). We additionally report HD95 (mm) and false-positive slice rate (FPR-S; fraction of predicted-positive slices outside the GT lesion extent) in Table S6 to quantify boundary error and over-propagation.

Baselines. For *image-level* (2D) comparison we include nnU-Net [11] trained on our dataset, pretrained SAM (ViT-B) [13] and MedSAM [18], and ScribblePrompt [28], I-MedSAM [26], and PCaSAM [34] (we disable automatic prompt generation for fair prompted setting) using the same training dataset. For *volume-level* (3D) comparison we include nnU-Net [11] trained on our dataset, pretrained SAM 2 (ViT-B+) [21], MedSAM2 [19], and Medical SAM 2 [35]. All baseline methods used stacked 3 modalities as input. All prompt-based methods receive identical simulated prompts and are evaluated on the same held-out cohorts. The public SAM 2-family baselines are evaluated using their released checkpoints without additional finetuning on our private prostate-bpMRI training set; protocol-specific finetuning could improve these baselines and remains an important future benchmark.

Implementation Details. We adapted pretrained SAM 2-image (ViT-b) for prompted-slice segmentation and SAM 2-video for volumetric propagation [21], trained on four NVIDIA H100 80 GB GPUs with AdamW (weight decay 0.01, cosine schedule, 5% warmup). Stage I trains AMF and LoRA adapters for 100

Table 2: Prompted-slice (image-level) lesion segmentation across 11 held-out cohorts. Values are DSC (mean±std) computed on the prompted slice and aggregated across lesions within each cohort. Best in **bold**, second underlined.

Method	Private-8	Private-9	Private-10	Private-11	Private-12	Private-13	Private-14	Private-15	ProstateX	Prostate158	TCIA	Avg
nnUNet [11]	0.47±0.23	0.45±0.24	0.56±0.19	0.55±0.21	0.60±0.16	0.57±0.22	0.49±0.20	0.43±0.24	0.50±0.18	0.44±0.23	0.39±0.25	0.50±0.21
Diameter prompt (2D slice)												
SAM [13]	0.51±0.22	0.49±0.24	0.61±0.18	0.59±0.20	0.64±0.15	0.60±0.21	0.54±0.19	0.47±0.22	0.55±0.17	0.48±0.21	0.42±0.24	0.54±0.20
ScribblePrompt [28]	0.53±0.21	0.50±0.23	0.63±0.17	0.62±0.19	0.67±0.14	0.64±0.20	0.56±0.18	0.49±0.21	0.57±0.16	0.51±0.20	0.45±0.23	0.56±0.19
I-MedSAM [26]	0.56±0.20	0.52±0.22	0.65±0.16	0.64±0.18	0.69±0.14	0.66±0.19	0.59±0.17	0.51±0.20	0.59±0.16	0.53±0.18	0.47±0.22	0.58±0.18
MedSAM [18]	0.66±0.18	0.62±0.21	0.76±0.14	0.75±0.16	0.80±0.12	0.77±0.18	0.70±0.15	0.63±0.19	0.68±0.14	0.60±0.18	0.56±0.19	0.68±0.17
PCaSAM [34]	0.77±0.15	0.73±0.18	0.85±0.11	0.86±0.12	0.90±0.09	0.88±0.13	0.80±0.12	0.73±0.15	0.79±0.11	0.69±0.14	0.65±0.15	0.79±0.13
Ours	**0.81**±0.14	**0.75**±0.17	**0.89**±0.10	**0.90**±0.11	**0.92**±0.08	**0.92**±0.12	**0.83**±0.11	**0.76**±0.14	**0.81**±0.10	**0.74**±0.13	**0.67**±0.14	**0.82**±0.12
Box prompt (2D slice)												
SAM [13]	0.55±0.21	0.52±0.23	0.64±0.17	0.63±0.19	0.67±0.15	0.65±0.20	0.58±0.18	0.50±0.21	0.57±0.17	0.51±0.20	0.44±0.23	0.57±0.19
ScribblePrompt [28]	0.58±0.20	0.54±0.22	0.66±0.16	0.66±0.18	0.70±0.14	0.68±0.19	0.60±0.17	0.53±0.20	0.59±0.16	0.54±0.19	0.47±0.22	0.60±0.18
I-MedSAM [26]	0.60±0.19	0.56±0.21	0.68±0.15	0.68±0.17	0.72±0.13	0.70±0.18	0.63±0.16	0.55±0.19	0.61±0.15	0.56±0.18	0.49±0.21	0.62±0.17
MedSAM [18]	0.70±0.17	0.67±0.20	0.79±0.13	0.81±0.15	0.84±0.11	0.82±0.17	0.75±0.14	0.66±0.18	0.71±0.13	0.62±0.17	0.58±0.18	0.72±0.16
PCaSAM [34]	0.80±0.14	0.78±0.17	0.89±0.10	0.91±0.11	0.93±0.08	0.92±0.12	0.86±0.11	0.78±0.14	0.83±0.10	0.73±0.13	0.70±0.14	0.83±0.12
Ours	**0.84**±0.13	**0.81**±0.16	**0.92**±0.09	**0.94**±0.10	**0.95**±0.07	**0.94**±0.11	**0.89**±0.10	**0.80**±0.13	**0.84**±0.09	**0.76**±0.12	**0.72**±0.13	**0.86**±0.11

epochs (lr 1×10^{-4}); Stage II initializes from Stage I and trains propagation-specific parameters (spacing-aware attention, confidence calibration, stop gate) for 100 epochs (lr 5×10^{-5}). The development split is patient-level (2,605/651 train/validation), and the loss weights are $\lambda_{\text{BCE}} = 1.0$, $\lambda_{\text{gate}} = 0.3$, and $\lambda_{\text{neg}} = 0.1$. We apply modality dropout to each input with probability p_{drop}=0.1 per sample. At inference, we propagate bidirectionally from the prompted slice, storing up to 8 memory entries with gating threshold τ=0.5. Training-time augmentations and further hyperparameters are provided in Table S1 and S3.

3.3 Quantitative Results

Prompted-slice lesion segmentation. We first isolate the image-level stage I performance by evaluating segmentation on the prompted slice alone, without volumetric propagation. Table 2 reports prompted-slice DSC across 11 held-out cohorts under bounding-box and diameter prompts. Circle prompt results are provided in Table S5.

As a detection-and-segmentation reference, nnU-Net (2D) achieves 0.50 mean DSC without any user-provided localization, confirming that cross-site lesion delineation remains challenging when the model must both detect and delineate. Among prompted methods, vanilla SAM reaches 0.57/0.54 (box/diameter), exposing a substantial natural-to-medical domain gap. Medical pretraining narrows this gap (MedSAM: 0.72/0.68), and prostate-specific multimodal modeling improves further (PCaSAM: 0.83/0.79). Our Stage I achieves the highest accuracy at 0.86/0.82, with consistent gains across cohorts including the most challenging sites (Private-8/9, TCIA). We attribute this to AMF learning spatially adaptive modality weighting via cross-attention, combined with prompt augmentation that exposes the model to realistic annotation noise during training.

Table 3: Prompted-slice segmentation on routine-care diameter labels (516 studies, 3 sites). DSC mean±std.

Method / DSC	SAM-B [13]	MedSAM [18]	PCaSAM [34]	**SAPP (Stage I)**
Diameter	0.54±0.14	0.69±0.18	0.74±0.13	**0.83**±0.06

Table 4: Prompt-to-volume (volume-level) lesion segmentation across 11 held-out cohorts. Starting from slice k, we run bidirectional propagation and report 3D DSC (mean±std) per cohort. Best in **bold**, second underlined within each prompt block; circle results are in Table S5.

Method	Private-8	Private-9	Private-10	Private-11	Private-12	Private-13	Private-14	Private-15	ProstateX	Prostate158	TCIA	Avg
nnUNet [11]	0.40±0.25	0.37±0.24	0.43±0.23	0.45±0.20	0.51±0.15	0.47±0.24	0.46±0.19	0.36±0.21	0.48±0.22	0.39±0.18	0.33±0.22	0.42±0.21
Single-slice diameter prompt (3D propagation)												
SAM 2 [21]	0.48±0.22	0.47±0.23	0.62±0.20	0.60±0.18	0.65±0.16	0.66±0.19	0.53±0.20	0.46±0.22	0.56±0.18	0.43±0.23	0.37±0.21	0.53±0.20
Medical SAM2 [35]	0.55±0.20	0.52±0.21	0.66±0.18	0.64±0.16	0.70±0.14	0.69±0.17	0.61±0.18	0.51±0.20	0.62±0.16	0.47±0.21	0.41±0.19	0.58±0.18
MedSAM2 [19]	0.58±0.19	0.54±0.20	0.67±0.17	0.66±0.15	0.72±0.13	0.71±0.16	0.63±0.17	0.53±0.19	0.61±0.15	0.49±0.20	0.43±0.18	0.60±0.17
SAPP	**0.75**±0.15	**0.72**±0.16	**0.84**±0.13	**0.86**±0.11	**0.89**±0.09	**0.90**±0.12	**0.80**±0.14	**0.71**±0.15	**0.81**±0.12	**0.67**±0.17	**0.61**±0.15	**0.78**±0.14
Single-slice box prompt (3D propagation)												
SAM 2 [21]	0.52±0.21	0.50±0.22	0.63±0.19	0.62±0.17	0.67±0.15	0.66±0.18	0.57±0.19	0.48±0.21	0.59±0.17	0.46±0.22	0.40±0.20	0.55±0.19
Medical SAM2 [35]	0.58±0.19	0.55±0.20	0.68±0.17	0.69±0.15	0.73±0.13	0.74±0.16	0.63±0.17	0.53±0.19	0.64±0.16	0.50±0.20	0.44±0.18	0.61±0.17
MedSAM2 [19]	0.60±0.18	0.57±0.19	0.72±0.16	0.70±0.14	0.75±0.12	0.74±0.15	0.65±0.16	0.55±0.18	0.66±0.15	0.52±0.19	0.46±0.17	0.63±0.16
SAPP	**0.76**±0.15	**0.76**±0.15	**0.87**±0.13	**0.91**±0.11	**0.91**±0.09	**0.91**±0.12	**0.82**±0.13	**0.72**±0.15	**0.80**±0.12	**0.70**±0.16	**0.62**±0.14	**0.80**±0.13
Single-slice oracle mask prompt (3D propagation)												
SAM 2 [21]	0.64±0.19	0.61±0.21	0.74±0.17	0.76±0.15	0.80±0.13	0.82±0.16	0.69±0.18	0.58±0.20	0.70±0.16	0.55±0.21	0.50±0.19	0.67±0.18
Medical SAM2 [35]	0.69±0.17	0.66±0.18	0.78±0.15	0.81±0.13	0.84±0.11	0.85±0.14	0.74±0.15	0.64±0.17	0.75±0.14	0.61±0.18	0.56±0.16	0.72±0.15
MedSAM2 [19]	0.71±0.16	0.68±0.17	0.82±0.14	0.81±0.12	0.86±0.10	0.87±0.13	0.76±0.14	0.66±0.16	0.77±0.13	0.63±0.17	0.58±0.15	0.74±0.14
SAPP	**0.85**±0.12	**0.82**±0.14	**0.92**±0.10	**0.94**±0.09	**0.96**±0.07	**0.97**±0.10	**0.89**±0.11	**0.81**±0.13	**0.90**±0.10	**0.78**±0.14	**0.72**±0.12	**0.87**±0.11

Replacing a box with a diameter prompt reduces spatial constraint and degrades all methods by ∼0.04 DSC on average. In Table S5, we additionally report circle prompting (0.75±0.12 DSC for our method); circles can include substantial surrounding tissue, making axis-aligned boxes the most informative prompt type in this setting. Pairwise improvements of SAPP over PCaSAM are statistically significant across all cohorts (Wilcoxon signed-rank test, $p<0.01$ after Bonferroni correction for 11 comparisons).

As a complementary evaluation, Table 3 reports results on 516 routine-care studies from three additional sites (Private-16–18) using site-provided diameter measurements as prompts and curated 2D contours as ground truth. Since these annotations are coarse, DSC should be interpreted as a consistency check rather than an absolute accuracy measure. The ranking from the fully annotated cohorts is preserved: our method achieves 0.83 mean DSC (+5 points over PCaSAM, +14 points over MedSAM), with the wider margin suggesting that prompt augmentation confers disproportionate robustness to real-world measurement noise.

Prompt-to-volume lesion segmentation. We next evaluate the full pipeline (Stage I + II) when supervision is limited to a single prompted slice. Table 4 reports volumetric DSC across 11 held-out cohorts under diameter, box, and oracle mask prompts; circle prompt results (0.74±0.12 mean DSC) are reported in Table S5.

Our method achieves 0.78/0.80/0.87 mean DSC under diameter/box/oracle prompts, improving over the strongest baseline (MedSAM2) by +18/+17/+13 points respectively. The gains widen under weaker prompts, indicating that spacing-aware propagation and the stop gate are most beneficial when the 2D predicted mask is imprecise and propagation must compensate for residual uncertainty. The prompt-type ranking mirrors the image-level results, with boxes outperforming diameters by 2–3 points. As an unprompted reference, nnU-Net,(3D) reaches only 0.42 mean DSC, about half our diameter-prompted result, under-

Table 5: Robustness to imperfect single-slice prompts, averaged over all held-out cohorts. We perturb prompt geometry (mask dilation/erosion, box jitter) and prompt-slice selection. Metric: 3D DSC (mean±std).

Method	Mask dilate (5 px)	Mask erode (5 px)	Box boundary jitter	Second-largest slice	Random lesion slice
MedSAM2 [19]	0.63±0.18	0.62±0.16	0.65±0.20	0.69±0.14	0.56±0.17
SAPP w/o prompt aug	0.71±0.18	0.70±0.17	0.68±0.22	0.77±0.20	0.63±0.19
SAPP (full)	0.75±0.16	0.74±0.16	0.72±0.18	0.82±0.15	0.71±0.17

scoring the value of minimal clinical input. Per-site HD95 and FPR-S are in Table S6.

3.4 Robustness analysis

Robustness to imperfect prompts. In practice, clinicians do not always prompt on the slice of maximal lesion extent; the reported slice may correspond to whatever cross-section was most conspicuous during reading. We evaluate Stage II under five realistic perturbations applied to the single-slice prompt, averaged over all held-out cohorts (Table 5).

Mask dilation, erosion, and box jitter each degrade MedSAM2 to 0.62–0.65 DSC, while our full model retains 0.72–0.75 DSC. Removing prompt augmentation during training yields comparable fragility, confirming that structured prompt noise is essential and that the propagation modules alone do not confer robustness to systematic 2D mask bias. The more clinically relevant perturbation is prompt-slice selection. On the second-largest lesion slice, our method retains 0.82 DSC (−4 points), indicating that a slightly sub-optimal slice choice has limited impact. However, prompting on a random lesion-containing slice, which can correspond to a marginal cross-section near the lesion boundary, reduces performance to 0.71 (−15 points); MedSAM2 drops to 0.56 (−18 points).

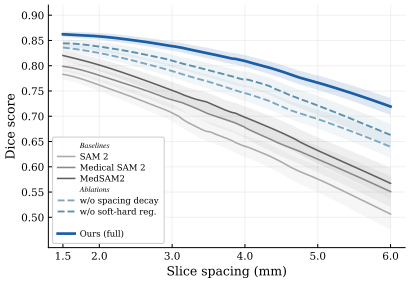

Fig. 4: Effect of anisotropic spacing. We stratify test volumes by native through-plane spacing and report mean 3D DSC within each bin under the oracle-mask prompt propagation setting. All methods degrade with coarser spacing; SAPP shows the shallowest decline. Curves marked as ablations remove spacing decay or the stop gate.

This gap reflects a fundamental asymmetry: propagating outward from a small 2d mask requires the model to expand into unseen lesion extent, whereas propagating inward from a large mask only requires progressive contraction of a well-initialized mask. Spacing-aware decay and the stop gate partially mitigate drift over longer propagation distances, but small-to-large expansion remains the most challenging regime.

Slice-spacing stratification. Prostate MRI protocols vary widely in through-plane spacing (Sec. 2.4; spacing statistics in Figure S2). To isolate how spacing

affects propagation, we bin all test volumes by native slice spacing and report mean DSC within each bin under the oracle-mask prompt propagation setting (Fig. 4).

At fine spacing (≤ 2.0 mm), adjacent slices are highly correlated and all methods perform reasonably. As spacing increases, performance degrades for all methods, but the rates differ markedly. SAM 2 drops from $\sim$0.78 at 1.5 mm to $\sim$0.51 at 6.0 mm (about -28 DSC points), consistent with the increasing mismatch between video-style propagation and sparsely sampled volumes. Medical adaptation improves the curve but remains spacing-sensitive (MedSAM2: $\sim$0.82 $\rightarrow$ $\sim$0.57, about -25 points). Our full model shows the shallowest decline ($\sim$0.86 $\rightarrow$ $\sim$0.72, about -14 points) and retains the highest DSC across all spacing bins.

Ablations confirm the complementarity of Module 2's components: removing spacing decay steepens the decline to $\sim$20 points, while removing the stop gate yields $\sim$18 points, indicating that the decay attenuates memory across large physical gaps and the gate prevents propagation beyond the lesion extent.

3.5 Ablation studies

Table 6 ablates key design choices of SAPP under the *oracle-mask* 3D setting, averaged over all held-out cohorts. This isolates (i) multimodal fusion (Module 1) and (ii) modifications to SAM 2 streaming-memory propagation (Module 2).

Module 1: Fusion strategy. Replacing our fusion with **late fusion** (PCaSAM-style) degrades both overlap and boundary accuracy (DSC 0.78 vs. 0.85). Late fusion is also heavier, running the image encoder separately per modality. **Early fusion** (channel stacking w/o cross-attention) is worse still (DSC 0.75), suggesting that learning modality-specific tokens before context-conditioned fusion is important. Removing the gland-prior input in an auxiliary ablation reduces oracle-mask 3D DSC by $\sim$ 0.08, indicating that the prior helps localization but is not the sole source of the gain.

Module 2: Spacing-aware propagation. Module 2 consists of **(a)** spacing-decayed memory attention and **(b)** the soft-to-hard stop gate. Removing spacing decay hurts more than removing the stop gate alone (-11 vs. -9 points), indicating that explicit awareness of *physical inter-slice gaps* is the primary driver, while

Model variation	DSC ↑
Late fusion	<u>0.78</u>$_{\pm 0.14}$
Early fusion w/o cross-attn	0.75$_{\pm 0.18}$
w/o spacing decay	0.74$_{\pm 0.16}$
w/o soft-hard gate	0.76$_{\pm 0.17}$
w/o Module 2	0.73$_{\pm 0.15}$
SAPP (full)	**0.85**$_{\pm 0.12}$

Table 6: Ablations under the oracle-mask prompt setting, averaged over held-out cohorts. Metric: 3D DSC (mean±std). Module 2 consists of spacing decay + stop gate.

Table 7: Inference efficiency on a single H100 GPU (batch size 1), one lesion per volume; wall-clock time and peak GPU memory (lower is better). Prompt-based methods use one oracle prompt. PCaSAM reports repeated slice-wise 2D inference.

Method	Time (s) ↓	Peak GPU (GB) ↓
SAM 2 [21]	5.8	9.6
MedicalSAM2 [35]	6.3	10.4
MedSAM2 [19]	7.1	12.8
PCaSAM [34]	10.3	15.3
SAPP (full)	**4.5**	**8.9**

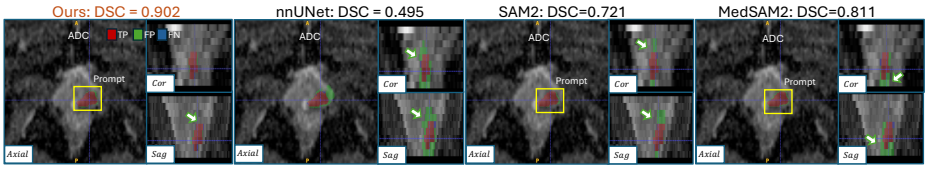

Fig. 5: Qualitative comparison of prompt-to-volume segmentation (ADC shown) from a single-slice prompt (yellow box). We visualize axial/coronal (*Cor*)/sagittal (*Sag*) views with error overlays (TP/FP/FN; colors as indicated). Dice scores are per-volume for this example. Baselines show false-positive leakage into lesion-absent slices, while SAPP better preserves lesion extent with fewer FP (e.g., 0.902 vs. nnU-Net 0.495, SAM 2 0.721, MedSAM2 0.811).

the stop gate complements it by preventing beyond-extent leakage. Removing both (w/o Module 2) yields the largest drop (DSC 0.73). Notably, w/o Module 2 (0.73) matches the strongest public MedSAM2 checkpoint (0.74, Table 4), so the full gain emerges only after adding propagation. This isolates Module 2's contribution beyond prostate-specific adaptation. This leakage reduction also shows in over-propagation metrics: under the oracle prompt, SAPP attains a false-positive slice rate of 0.06 versus 0.30 for SAM 2 (Table S6). The confidence factor is not ablated separately, as it is tightly coupled to the stop decision; the w/o soft-hard gate row removes confidence-controlled stopping.

Inference efficiency. Table 7 compares wall-clock time and peak GPU memory. SAPP is faster than SAM 2 and Medical SAM 2 variants (4.5 s vs. 5.8–7.1 s) because the stop gate terminates propagation early in most volumes, avoiding unnecessary decoding on lesion-absent slices. PCaSAM is slower and uses more memory in part because it computes multiple image-encoder embeddings across modalities. Peak memory is also lower for SAPP (8.9 GB), consistent with early termination reducing the number of decoded slices and written memory entries.

3.6 Qualitative comparison: propagation failures and false positives

Figure 5 visualizes a representative lesion example under a single-slice prompt. We overlay true positives (TP), false positives (FP), and false negatives (FN) to highlight where errors arise across views. While prior prompt-based baselines can recover the lesion core, we frequently observe over-propagation into lesion-absent slices, producing extensive FP regions. In contrast, our spacing-aware propagation suppresses leakage and preserves a coherent 3D extent, yielding substantially fewer FP and higher overlap. These qualitative findings align with our quantitative gains and support the role of physical spacing and explicit stopping in prompt-to-volume segmentation. Additional qualitative results are provided in Sec. S7 (Figs. S3 and S4).

4 Discussion and Limitations

Our method works best when the prompt falls on or near the maximal lesion slice, matching typical reporting practice. The main failure modes are extremely

small lesions, severe diffusion distortion or misregistration, and diffuse margins where even expert delineation is ambiguous; the stop gate can also terminate prematurely in heterogeneous lesions with abrupt appearance changes. Current limitations include a global spacing decay parameter and fixed stopping threshold (per-volume adaptation may help), the assumption of aligned multimodal inputs, and retrospective-only evaluation. By design, SAPP performs prompt-to-volume completion, lifting a routine 2D prompt into a coherent 3D mask rather than discovering unprompted lesions; extending it toward automatic detection is a natural next step.

5 Conclusion

This paper introduces SAPP, a prompt-to-volume framework that adapts SAM 2 to multimodal prostate bpMRI by combining adaptive multimodal fusion for *image-level* prompted segmentation with spacing-aware, confidence-gated prompt propagation and a learned stop gate for *volume-level* completion. Across 14 held-out cohorts, extensive experiments and robustness analyses show that SAPP reliably lifts minimal clinical prompts to coherent 3D lesion masks and outperforms specialist and foundation-model baselines. Finally, given evidence that lesion-mask priors can improve PI-RADS scoring on external cohorts [34], prompt-propagated 3D labels may support label-efficient training of downstream detection and risk stratification pipelines at scale.

Disclaimer The concepts and information presented in this paper are based on research results that are not commercially available. Future commercial availability cannot be guaranteed.

ProstateAI Clinical Collaborators Dr. Henkjan Huisman (Radboud University Medical Center, Nijmegen, Netherlands), Dr. Angela Tong (New York University, New York City, NY, USA), Dr. David Winkel (Universitätsspital Basel, Basel, Switzerland), Dr. Tobias Penzkofer (Charité, Universitätsmedizin Berlin, Berlin, Germany), Dr. Ivan Shabunin (Patero Clinic, Moscow, Russia), Dr. Moon Hyung Choi (Eunpyeong St. Mary's Hospital, Catholic University of Korea, Seoul, Republic of Korea), Dr. Qingsong Yang (Changhai Hospital of Shanghai, Shanghai, China), Dr. Dieter Szolar (Diagnostikum Graz Süd-West, Graz, Austria), Dr. Steven Shea (Loyola University Medical Center, Maywood, IL, USA), Dr. Fergus Coakley (Oregon Health and Science University, Portland, OR, USA), Dr. Mukesh Harisinghani (Massachusetts General Hospital, Boston, MA, USA), Dr. Lidia Alcalá Mata (HT Médica, Jaén, Spain), Dr. Justus Mahnke (University Hospital Schleswig Holstein, Kiel, Germany), Dr. Heiner Steffens (Radprax, Wuppertal, Germany), Dr. Magdalena Zagrodzka (QUADIA Diagnostic Centre, Piaseczno, Poland), and Dr. Edyta Szurowska (Medical University of Gdansk, Gdansk, Poland).

References

1. Adams, L.C., Makowski, M.R., Engel, G., Rattunde, M., Busch, F., Asbach, P., Niehues, S.M., Vinayahalingam, S., van Ginneken, B., Litjens, G., Bressem, K.K.: Prostate158 — an expert-annotated 3t mri dataset and algorithm for prostate cancer detection. Computers in Biology and Medicine **148**, 105817 (2022). https://doi.org/10.1016/j.compbiomed.2022.105817
2. Ahdoot, M., Wilbur, A.R., Reese, S.E., Lebastchi, A.H., Mehralivand, S., Gomella, P.T., Bloom, J., Gurram, S., Siddiqui, M., Pinsky, P., Parnes, H., Linehan, W.M., Merino, M., Choyke, P.L., Shih, J.H., Turkbey, B., Wood, B.J., Pinto, P.A.: MRI-targeted, systematic, and combined biopsy for prostate cancer diagnosis. New England Journal of Medicine **382**(10), 917–928 (2020). https://doi.org/10.1056/NEJMoa1910038
3. An, D., Gu, P., Sonka, M., Wang, C., Chen, D.Z.: Sli2vol+: Segmenting 3d medical images based on an object estimation guided correspondence flow network. arXiv preprint arXiv:2411.13873 (2024). https://doi.org/10.48550/arXiv.2411.13873, https://arxiv.org/abs/2411.13873
4. Boellaard, T.N., van Erck, R., van der Graaf, S.H., de Boer, L., van der Poel, H.G., Mertens, L.S., van Leeuwen, P.J., Dashtbozorg, B.: Comparing AI and manual segmentation of prostate MRI: Towards AI-driven 3d-model-guided prostatectomy. Diagnostics **15**(9), 1141 (2025). https://doi.org/10.3390/diagnostics15091141
5. Chen, M., Zhang, F., Zhao, Z., Yao, J., Zhang, Y., Wang, Y.: Probabilistic conformal distillation for enhancing missing modality robustness. Advances in Neural Information Processing Systems **37**, 36218–36242 (2024)
6. Chen, Z., Nan, X., Li, J., Zhao, J., Li, H., Lin, Z., Li, H., Chen, H., Liu, Y., Tang, L., Zhang, L., Dong, B.: PAM: A propagation-based model for segmenting any 3d objects across multi-modal medical images. arXiv preprint arXiv:2408.13836 (2024). https://doi.org/10.48550/arXiv.2408.13836, https://arxiv.org/abs/2408.13836
7. Cheng, H.K., Schwing, A.G.: Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In: European conference on computer vision. pp. 640–658. Springer (2022)
8. Du, Y., Bai, F., Huang, T., Zhao, B.: SegVol: Universal and interactive volumetric medical image segmentation. arXiv preprint arXiv:2311.13385 (2023). https://doi.org/10.48550/arXiv.2311.13385, https://arxiv.org/abs/2311.13385
9. Havaei, M., Guizard, N., Chapados, N., Bengio, Y.: Hemis: Hetero-modal image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 469–477. Springer (2016)
10. He, Y., Guo, P., Tang, Y., Myronenko, A., Nath, V., Xu, Z., Yang, D., Zhao, C., Simon, B., Belue, M., Harmon, S., Turkbey, B., Xu, D., Li, W.: VISTA3D: A unified segmentation foundation model for 3d medical imaging. arXiv preprint arXiv:2406.05285 (2024). https://doi.org/10.48550/arXiv.2406.05285, https://arxiv.org/abs/2406.05285
11. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z
12. Jeganathan, T., Salgues, E., Schick, U., Tissot, V., Fournier, G., Valéri, A., Nguyen, T.A., Bourbonne, V.: Inter-rater variability of prostate lesion segmentation on

multiparametric prostate MRI. Biomedicines **11**(12), 3309 (2023). https://doi.org/10.3390/biomedicines11123309

13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023). https://doi.org/10.48550/arXiv.2304.02643, https://arxiv.org/abs/2304.02643
14. Li, H., Liu, H., Hu, D., Wang, J., Oguz, I.: Prism: A promptable and robust interactive segmentation model with visual prompts. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 389–399. Springer (2024)
15. Li, H., Liu, H., Hu, D., Wang, J., Oguz, I.: Promise: Prompt-driven 3d medical image segmentation using pretrained image foundation models. In: 2024 IEEE international symposium on biomedical imaging (ISBI). pp. 1–5. IEEE (2024)
16. Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N., Huisman, H.: Computer-aided detection of prostate cancer in mri. IEEE Transactions on Medical Imaging **33**(5), 1083–1092 (2014). https://doi.org/10.1109/TMI.2014.2303821, https://pubmed.ncbi.nlm.nih.gov/24770913/
17. Liu, H., Fan, Y., Li, H., Wang, J., Hu, D., Cui, C., Lee, H.H., Zhang, H., Oguz, I.: Moddrop++: A dynamic filter network with intra-subject co-training for multiple sclerosis lesion segmentation with missing modalities. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 444–453. Springer (2022)
18. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B., et al.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024). https://doi.org/10.1038/s41467-024-44824-z
19. Ma, J., Yang, Z., Kim, S., Chen, B., Baharoon, M., Fallahpour, A., Asakereh, R., Lyu, H., Wang, B.: Medsam2: Segment anything in 3d medical images and videos. arXiv preprint arXiv:2504.03600 (2025). https://doi.org/10.48550/arXiv.2504.03600, https://arxiv.org/abs/2504.03600
20. Natarajan, S., Priester, A., Margolis, D., Huang, J., Marks, L.: Prostate mri and ultrasound with pathology and coordinates of tracked biopsy (prostate-mri-us-biopsy) (version 2) [data set] (2020). https://doi.org/10.7937/TCIA.2020.A61IOC1A
21. Ravi, N., et al.: SAM 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024), https://arxiv.org/abs/2408.00714
22. Rui, S., Chen, L., Tang, Z., Wang, L., Liu, M., Zhang, S., Wang, X.: Multi-modal vision pre-training for medical image analysis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5164–5174 (2025)
23. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians **71**(3), 209–249 (2021). https://doi.org/10.3322/caac.21660
24. Turkbey, B., Rosenkrantz, A.B., Haider, M.A., Padhani, A.R., Villeirs, G., Macura, K.J., Tempany, C.M., Choyke, P.L., Cornud, F., Margolis, D.J., Thoeny, H.C., Verma, S., Barentsz, J., Weinreb, J.C.: Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. European Urology **76**(3), 340–351 (2019). https://doi.org/10.1016/j.eururo.2019.02.033

25. Wang, J., Liu, H., Li, H., Bagnato, F., Oguz, I.: Modular cross-attention fusion for multi-modal ms spine lesion segmentation with missing modalities. Ms-MultiSpine challenge proceedings:"Multiple Sclerosis Spinal Cord Lesion Detection from MultiSequence MRIs" p. 1 (2025)
26. Wei, X., Cao, J., Jin, Y., Lu, M., Wang, G., Zhang, S.: I-medsam: Implicit medical image segmentation with segment anything. arXiv preprint arXiv:2311.17081 (2024). https://doi.org/10.48550/arXiv.2311.17081, https://arxiv.org/abs/2311.17081
27. Winkel, D.J., Tong, A., Lou, B., Kamen, A., Comaniciu, D., Disselhorst, J.A., Rodríguez-Ruiz, A., Huisman, H., Szolar, D., Shabunin, I., et al.: A novel deep learning based computer-aided diagnosis system improves the accuracy and efficiency of radiologists in reading biparametric magnetic resonance images of the prostate: results of a multireader, multicase study. Investigative radiology **56**(10), 605–613 (2021). https://doi.org/10.1097/RLI.0000000000000780, https://pubmed.ncbi.nlm.nih.gov/33787537/
28. Wong, H.E., Rakic, M., Guttag, J., Dalca, A.V.: Scribbleprompt: Fast and flexible interactive segmentation for any biomedical image. arXiv preprint arXiv:2312.07381 (2024). https://doi.org/10.48550/arXiv.2312.07381, https://arxiv.org/abs/2312.07381
29. Wu, J., Wang, Z., Hong, M., Ji, W., Fu, H., Xu, Y., Xu, M., Jin, Y.: Medical SAM adapter: Adapting segment anything model for medical image segmentation. Medical Image Analysis **102**, 103547 (2025). https://doi.org/10.1016/j.media.2025.103547
30. Wu, J., Xu, M.: One-prompt to segment all medical images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11302–11312 (2024)
31. Yang, Z., Wei, Y., Yang, Y.: Associating objects with transformers for video object segmentation. Advances in Neural Information Processing Systems **34**, 2491–2502 (2021)
32. Yang, Z., Yang, Y.: Decoupling features in hierarchical propagation for video object segmentation. Advances in Neural Information Processing Systems **35**, 36324–36336 (2022)
33. Zhang, Y., He, N., Yang, J., Li, Y., Wei, D., Huang, Y., Zhang, Y., He, Z., Zheng, Y.: mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 107–117. Springer (2022)
34. Zhang, Y., Ma, X., Li, M., Huang, K., Zhu, J., Wang, M., Wang, X., Wu, M., Heng, P.A.: Generalist medical foundation model improves prostate cancer segmentation from multimodal MRI images. npj Digital Medicine **8**(1), 372 (2025). https://doi.org/10.1038/s41746-025-01756-2
35. Zhu, J., Hamdi, A., Qi, Y., Jin, Y., Wu, J.: Medical SAM 2: Segment medical images as video via segment anything model 2. arXiv preprint arXiv:2408.00874 (2024). https://doi.org/10.48550/arXiv.2408.00874