How to Adapt SAM3 as a Strong Text-Prompted 3D Medical Segmentor?

Sangwoo Youn[1[0009−0007−1200−8463]], Xiaoke Huang[2[0000−0003−1658−6432]], Joonhyuk Seo[3[0009−0006−0245−2133]], Fei Xia[3[0000−0001−6591−8769]], Alvaro A. Cardenas[2[0000−0002−5142−9750]], and Yuyin Zhou[2[0000−0003−2232−9563]]⋆

[1] KAIST
[2] University of California, Santa Cruz
[3] University of California, Irvine
andy2884@kaist.ac.kr, xhuan192@ucsc.edu, joonhys1@uci.edu, fei.xia@uci.edu, alvaro.cardenas@ucsc.edu, yzhou284@ucsc.edu

Abstract. Recent advances in SAM3 integrate text-conditioned detection with memory-based tracking, enabling powerful promptable segmentation in natural videos. However, directly adapting SAM3 to biomedical volumes is challenging due to substantial domain differences, including heterogeneous imaging modalities, sparse targets, severe class imbalance, and the fixed through-plane continuity between adjacent slices. Existing adaptations largely overlook these characteristics and therefore achieve suboptimal performance. In this paper, we present Dr. SAM3, the first systematic investigation of SAM3 for text-prompted 3D medical segmentation, and introduce an effective modality- and volume-aware adaptation recipe without changing its core architecture. The key insight is that SAM3's visual-language and memory priors transfer best when modality identity, source-volume structure, and fixed slice geometry are made explicit. On the CVPR 2026 *Foundation Models for Text-guided 3D Biomedical Image Segmentation* Challenge, Dr. SAM3 achieves the strongest mean instance F1 among existing strong text-prompted methods. On the clinically important CT/MR oncologic tumor and lesion subset, it reaches the best class-macro instance F1 of 52.8 (+2.9). On the unseen RAOS-Real CT dataset, Dr. SAM3 attains 77.2 (+18.4) class-average DSC and outperforms strong text-prompted references, suggesting that SAM3's generalization ability can be effectively transferred to medical segmentation through data-interface adaptation.

Keywords: Text-prompted segmentation · 3D medical imaging · SAM3 · Foundation models · Biomedical segmentation.

1 Introduction

3D medical segmentation sits at the heart of modern volumetric biomedical imaging, underpinning how clinicians see, quantify, and act on disease inside the

⋆ Corresponding author.

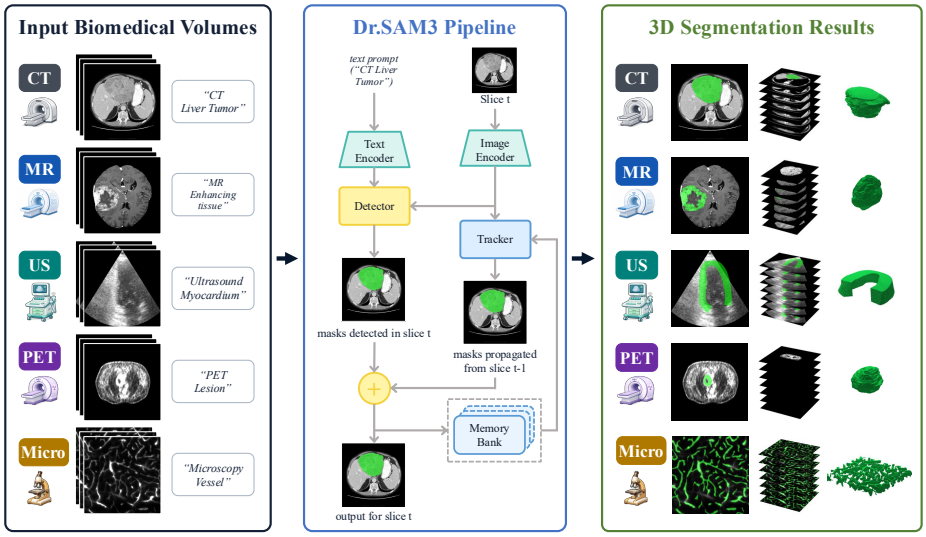

Fig. 1. Dr. SAM3 for text-prompted 3D medical segmentation. Pipeline overview for adapting SAM3 to biomedical volumes with modality-aware class-name prompts and ordered slice propagation.

body [11,24,3,7]. Across modalities ranging from Computed Tomography (CT) and Magnetic Resonance (MR) to Positron Emission Tomography (PET), Ultrasound (US), and Microscopy, accurate segmentation masks are no longer just useful annotations: they are the basis for reliable diagnosis, precise treatment planning, safe surgical navigation, robust biomarker extraction, and consistent quantitative follow-up. A practical 3D medical segmentor must handle heterogeneous imaging physics, large scale variation, sparse tumors and lesions, dense structures such as blood vessels, severe class imbalance, and limited annotations. In the text-prompted setting, the input is a volume and a class-name prompt such as "CT liver tumor" or "Microscopy vessel". The model must localize and segment the target from text, without boxes, clicks, or task-specific heads.

Earlier versions of the Segment Anything Model (SAM) [14,26] had a major impact on medical image segmentation by demonstrating strong generalization and promptable segmentation, motivating many adaptations to clinical and biomedical settings [19,22]. The latest SAM3 [1] advanced this line of work further by combining two useful abilities in one model. Its detector localizes visual concepts from text prompts, while its memory-based tracker propagates masks through an ordered frame sequence. This detector-tracker design suggests a simple route for medical volumes: treat a 3D scan as a sequence of 2D slices, use the class name as the text prompt, and let the tracker propagate objects along the slice axis. Motivated by this formulation, we investigate whether SAM3's general-domain visual-language and memory priors can be effectively transferred to 3D medical segmentation.

However, this transfer is not a plug-in application. The main challenge is not only the appearance gap between natural and medical images, but also in the mismatch between biomedical data characteristics and the interfaces used by SAM3. First, modality is part of the visual concept. The same label can look very different across modalities, so simply pooling multi-modality training set can cause interference if modality identity is ignored. Second, biomedical targets have very different slice-level densities. Tumors and lesions may appear in only a few slices, while blood vessels can create more instances than the detector can represent with its finite query budget. At the same time, some modalities contain many empty slices. The model therefore still needs negative examples to learn when no object is present, but too many empty slices can bias this calibration toward background predictions. Third, the long-tail distribution appears at the volume level. Rare labels often present in only one or a few source volumes, so repeatedly sampling isolated 2D slices fails to expose the model to the surrounding 3D anatomical context of those rare targets. Finally, a medical volume is not a natural video. Consecutive slices are spatially adjacent samples with a fixed through-plane order, so temporal sampling strategies such as random temporal strides designed for videos may skip small structures that occupy only a short slice range.

We tackle these challenges with **Dr. SAM3**, a data-interface adaptation of SAM3 that modifies the prompt strings, the volume and slice sampler, the treatment of dense and empty slices, and the video-style clips provided to the tracker. The underlying SAM3 core architecture remains unchanged. This lets us isolate a central question: how much of SAM3's text-conditioned segmentation ability can be transferred by aligning the input interface with the structure of medical volumes? To address this question, we designed our adaptation with three parts. First, modality-conditioned adaptation makes modality visible to the model and to the sampler. We prepend the modality name to each class prompt, balance per-epoch exposure across modalities, split very dense slices into fixed patches when the instance count exceeds the detector query capacity, and cap the sampled fraction of empty slices only for modalities where empty slices dominate. Second, volume-level repeat factor sampling repeats source volumes that contain rare labels, rather than repeating individual 2D slices. This preserves neighboring-slice context while increasing exposure to rare classes. Third, volume-to-video tracker adaptation presents a medical scan as an ordered slice sequence rather than as a natural video. We train and evaluate the tracker on adjacent slices with stride one, propagate from a central slice toward the periphery, and keep short target sequences by padding them with masked duplicate slices instead of discarding them.

Our design choices lead to two main insights. The first is that modality and volume identity are not metadata to be ignored. They define how text labels should be grounded and how rare classes should be sampled. The second is that the tracker should respect medical slice geometry. For small tumors and lesions, adjacent-slice propagation provides a more reliable signal than video-style random temporal sampling. Our ablation studies support these insights:

modality-balanced sampling recovers under-trained modalities, volume-level repeat sampling improves rare classes, and adjacent-slice sequencing substantially improves instance propagation.

To validate these design choices, we evaluate Dr. SAM3 on the CVPR 2026 *Foundation Models for Text-guided 3D Biomedical Image Segmentation* [20] Challenge, a large multi-modality volumetric testbed spanning CT, MR, US, PET, and Microscopy. Dr. SAM3 achieves the strongest mean instance F1 among compared text-prompted methods, while using only the challenge training data. On the clinically important CT/MR tumor and lesion subset, where targets are often small, sparse, and instance-sensitive, Dr. SAM3 reaches the best class-macro instance F1, that is F1 averaged over classes, of 52.8 (+2.9). To test external transfer, we also evaluate on the unseen RAOS-Real CT dataset [17] without RAOS-specific training or prompt tuning. Dr. SAM3 reaches 77.2 (+18.4) class-average Dice similarity coefficient (DSC), showing robust generalization under source-domain shift. Together, these results demonstrate that carefully adapting general-domain foundation models to the structural properties of medical volumes can substantially improve transferable 3D medical segmentation.

2 Related Work

Medical segmentation. Medical image segmentation has long been dominated by supervised models trained for a fixed label space. U-Net and V-Net established encoder-decoder designs for 2D and volumetric prediction [28,24]. nnU-Net then turned this family into a strong self-configuring specialist pipeline [11]. Transformer, hybrid, and state-space variants, represented by TransUNet and later 3D backbones, further improved context modeling for medical volumes [3,7,6,21]. Large-scale systems such as TotalSegmentator show the value of broad anatomical annotation for robust CT segmentation [31]. These advances are strong, but they are still centered on models that are trained, selected, or deployed for a specific modality, organ set, lesion family, or dataset. In this work, we instead study one text-prompted model that segments organs and lesions across CT, MR, US, PET, and Microscopy.

Spatial-prompted medical segmentation. The SAM family established promptable segmentation with points, boxes, and masks, and most medical SAM work adapts this spatial interface to clinical images [14,26]. MedSAM is the representative 2D adaptation [19], while MedSAM2 and SAM-Med3D extend the interface to medical videos and volumetric images [22,30]. SegVol and VISTA3D further show that 3D promptable models can use volumetric context for interactive segmentation [4,8]. Empirical SAM studies and adapter-based 3D variants also show that medical-domain appearance and volumetric context must be handled before SAM becomes reliable on medical data [23,9,5,2]. Spatial prompts are effective for annotation and interactive correction, where a human can localize the target. They are less convenient for clinical use at scale because the user must provide localization, often slice by slice.

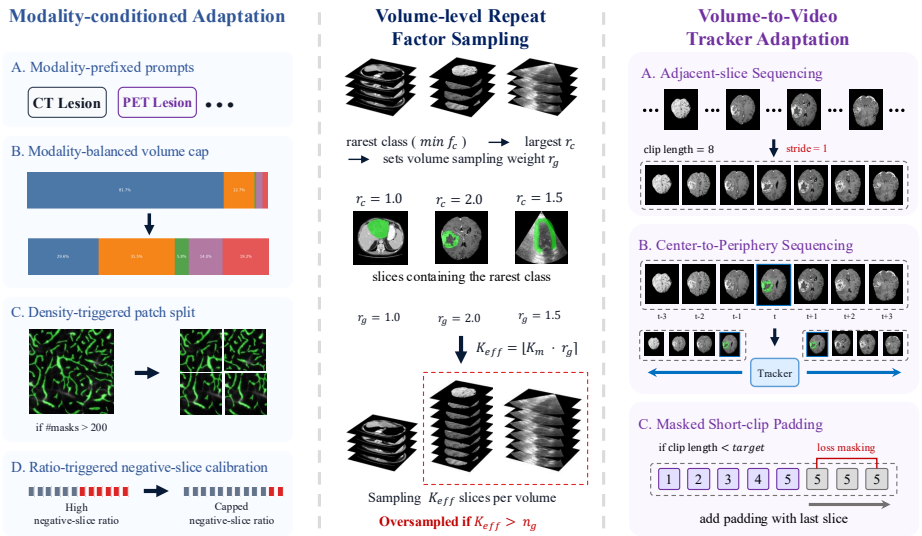

Fig. 2. Method overview. Dr. SAM3 adapts SAM3 to biomedical volumes through modality-conditioned adaptation, volume-level repeat factor sampling, and volume-to-video tracker adaptation.

Text-prompted medical segmentation. Text-prompted segmentation removes the need for spatial prompts by grounding class labels, noun phrases, or medical terms directly. Early approaches used language embeddings to share labels across organs and tumors [25,16,13], while CAT reduces ambiguity with coordinated anatomical and textual prompts [10]. SAT scales this idea to large-vocabulary 3D radiology using a terminology tree and SAT-DS, and BiomedParse and BiomedParse-V broaden text-guided segmentation across biomedical modalities and volumes [35,33,34]. Medal-S adds spatial self-refinement for MedSegFM [29]. In contrast, the SAM family is still mainly spatial-prompted; recent SAM3 medical adaptations improve with fine-tuning but remain behind strong text-prompted 3D systems on multi-modality volumes [1,15,12]. Dr. SAM3 addresses this by adapting SAM3's data interface, not its architecture. It uses far less fine-tuning compute than Medal-S and BiomedParse-V, avoids closed medical training data, and transfers SAM3 by making modality, volume structure, and slice geometry explicit.

3 Method

Dr. SAM3 adapts SAM3 for text-prompted 3D medical segmentation system through modality-conditioned adaptation, volume-level repeat factor sampling, and volume-to-video tracker adaptation, without modifying the underlying SAM3 architecture.

3.1 Modality-Conditioned Adaptation

Modality identity (CT, MR, etc.) is observable for every training sample, and ignoring it leaves cross-modality interference across the pipeline. Our modality-conditioned adaptation makes modality explicit at four interfaces: prompting, sampling, dense-slice handling, and negative-slice calibration.

Modality-prefixed prompts. We prepend the imaging modality to form the final prompt, such as `"CT Pancreas"` or `"MR Pancreas"`. This modality prefix disambiguates visually different concepts that share the same anatomical name across modalities and also separates labels in our frequency tables, so the sampler treats `"CT Liver"` and `"MR Liver"` as distinct modality-class pairs.

Modality-balanced volume cap. Multi-modality medical training pools often contain unequal slice counts across modalities, so uniform volume sampling can under-train smaller modalities. For a volume group g with modality $m(g)$, the base per-epoch slice cap is

$$K_g = K_{m(g)}.$$

We choose K_m by solving for caps that approximately equalize sampled slice exposure $B_m = \sum_{g:m(g)=m} \min(n_g, K_m)$ across modalities under a fixed epoch budget, with a lower bound of four slices per volume. The caps are computed from training-set volume counts only; modalities already below the target exposure remain uncapped, which is equivalent to setting $K_m = \infty$. This gives the detector a modality-balanced base sampler before applying the long-tail volume sampler. Exact caps, epoch mix, and the uniform-cap diagnostic are reported in Appendix A and Appendix D.

Density-triggered patch split. SAM3's detector has a query capacity $Q = 200$, which can be exceeded by dense instance slices. For class c and modality m, we trigger a fixed quadrant path when

$$\max_{s \in \mathcal{S}^+_{m,c}} I(s, c) > Q,$$

where $\mathcal{S}^+_{m,c}$ is the set of positive training slices for the modality–class pair and $I(s, c)$ is the number of class-c instances on slice s. The trigger depends only on training annotations and the prompt–modality pair, not on test-time annotation density. In this benchmark, only `Microscopy Vessel` triggers the patch-split path. All other prompts use the global-resize path. At inference, quadrant predictions are mapped back to the original slice coordinates before the standard per-class merging.

Ratio-triggered negative-slice calibration. The per-modality negative-slice ratio can vary by an order of magnitude. Let N^-_m and N^+_m denote the raw zero-annotation and positive slice counts in modality m. We compute

$$\rho_m = \frac{N^-_m}{N^+_m + N^-_m}, \qquad \text{trigger}(m) = \mathbf{1}[\rho_m > \rho_0],$$

with $\rho_0 = 0.2$. For triggered modalities, we cap negative sampling so that the sampled negative-slice ratio does not exceed 20%, while still retaining negatives for no-object calibration. In this benchmark, PET is the only modality that triggers this rule (Table 15), with a 48.8% raw negative-slice ratio. Other modalities use the natural training ratio.

3.2 Volume-Level Repeat Factor Sampling (RFS)

Medical anatomy follows a long tail that uniform random sampling cannot fix. The problem is sharper than in natural images because rare classes are concentrated in a small number of source volumes (sometimes 1–2), and the relevant axis for the long tail is the *volume*, not the slice. We therefore use volume-level repeat factor sampling, which assigns repeat expansion to source volumes rather than isolated slices, and compare Capped volume RFS with Oversampled volume RFS for small-volume cases. The sampler can be applied to any modality; in the final recipe, we enable volume RFS only for CT/MR volume groups because their long-tail imbalance is most severe in our training pool.

Algorithm. Our change is to move the repeat decision from image to volume while keeping frequency estimates at the slice level. After modality-prefixing, a class $c = (m, y)$ has frequency f_c within its own modality m: the numerator counts non-empty training slices in modality m that contain c, and the denominator counts all non-empty training slices in modality m. For enabled modalities $\mathcal{M}_{\text{RFS}}$, each class receives $r_c = \max(1, \sqrt{t/f_c})$. A volume group g initializes $r_g = 1$ and sets it to the maximum r_c over classes present anywhere in the volume. Its effective per-epoch slice budget is then

$$K_{\text{eff}} = \text{round}_s(K_m \cdot r_g),$$

where round_s is stochastic rounding. Modalities outside $\mathcal{M}_{\text{RFS}}$ use the base budget $K_{\text{base}}(g) = \min(n_g, K_m)$. Thus, rare classes increase exposure to the source volumes in which they occur, rather than causing isolated 2D slices to be repeated independently.

Algorithm 1 in Appendix D.1 gives the full sampling pseudocode.

Capped vs. Oversampled volume RFS. For small medical volumes, K_{eff} can exceed n_g. Capped volume RFS truncates K_{eff} at n_g, so rare classes receive no extra exposure once all available slices are already used; Oversampled volume RFS first uses all available slices from the volume, then fills the remaining budget by resampling slices from the same volume. We use Oversampled volume RFS with $t = 0.01$ in the final recipe; threshold rationale and the LVIS-default comparison are discussed in Appendix D.1.

3.3 Volume-to-Video Tracker Adaptation

SAM3's tracker is trained on video clips, but a 3D medical volume is not a natural video. The sequence axis is spatial rather than temporal, and target extent is

often limited to a few neighboring slices. We therefore adapt the volume-to-video interface rather than changing the tracker architecture.

Adjacent-slice sequencing. For tracker training, we sample consecutive axial slices instead of using SAM3's default random temporal stride. Equivalently, the tracker clip stride is fixed to 1. Adjacent slices are the closest available observations along the slice axis, so this reduces the chance that a larger slice stride skips the observations supporting small tumors and lesions.

Center-to-periphery sequencing. During training, each volume is split around a center slice into two videos: the pre-center side is reversed and the post-center side keeps its native order, so the temporal direction points from the center toward the periphery on both sides. We also disable SAM3's default random time-axis reversal, making temporal direction consistently mean center-to-periphery. We apply the same center-to-periphery sequencing at inference time.

Masked short-clip padding. SAM3's default tracker dataloader skips videos shorter than the required clip length. In medical volumes, however, tumor and lesion targets can occupy only a few annotated slices, so this default would preferentially discard small-volume examples. We keep them by padding to the target clip length with padded entries that duplicate the last valid slice and are masked out of both training losses and tracker memory updates.

4 Experiments

4.1 Dataset and Evaluation Metrics

We evaluate on the official CVPR 2026 *Foundation Models for Text-guided 3D Biomedical Image Segmentation* (MedSegFM) Challenge benchmark [20]. The benchmark is a multi-modality 3D biomedical segmentation testbed spanning CT, MR, Ultrasound, PET, and Microscopy. The validation set contains 2.1K volumes. We use only the organizer-provided preprocessed training and validation data for MedSegFM experiments, and the official 3D validation set is used only for final reporting.

Following the official segmentation protocol, semantic classes are evaluated by Dice similarity coefficient (DSC) and normalized surface Dice (NSD). Instance classes are evaluated by Instance F1 and DSC-TP, where DSC-TP is the mean Dice over true-positive matched instances. We report two operating points: Dr. SAM3-D, which applies the adapted SAM3 detector slice by slice, and Dr. SAM3-T, which initializes objects with the detector and propagates them through the adapted tracker.

4.2 Implementation Details

Preprocessing We use the challenge-provided preprocessed volumes and annotations. Each 2D slice is mapped to SAM3's 504×504 canvas. Class descriptions

are converted to short canonical anatomical noun phrases and prefixed with the imaging modality, such as `CT Liver tumor` or `MR Prostate lesion`. Dense slices are processed using a fixed quadrant split when the training annotation density exceeds the detector query budget. For the final tracker stage, each 3D volume is represented as an ordered slice sequence with adjacent-slice stride one.

Environment Settings The development environments and requirements are presented in Table 1.

Table 1. Development environments and requirements.

System	Ubuntu 24.04.4 LTS, Linux kernel 6.17.0-1007-oracle
CPU	AMD EPYC 7J13, 2 sockets, 64 cores per socket, 256 logical CPUs
RAM	2.0 TiB total system memory
GPU (number and type)	8× NVIDIA A100-SXM4-80GB
CUDA version	PyTorch CUDA runtime 12.1; NVIDIA driver 580.105.08; `nvidia-smi` reports CUDA 13.0
Programming language	Python 3.12.3
Deep learning framework	PyTorch 2.5.1+cu121, cuDNN 90100, SAM3 training stack

Training Protocols Training follows a detector-then-tracker schedule. Phase 1 fine-tunes the SAM3 detector on 2D medical slices using modality-prefixed prompts, modality-balanced exposure, dense-slice patch splitting, PET negative-slice calibration, and CT/MR volume-level repeat factor sampling. Phase 2 freezes the detector, prompt/text pathway, and image backbone, and fine-tunes tracker propagation modules on adjacent-slice clips. Checkpoints are selected without using official 3D validation metrics; the official validation set is reserved for final reporting.

Data Augmentation and Model Selection We use a deliberately limited augmentation policy to avoid changing biomedical geometry. For detector fine-tuning, we use SAM3's image-space random resizing on the 504×504 canvas and then map predictions back to the original slice coordinates. We do not use additional geometric or appearance augmentation, nor test-time augmentation. For tracker fine-tuning, ordered adjacent-slice clips, center-to-periphery ordering, and masked duplicate padding are part of the volume-to-video input construction rather than data augmentation. Sampling is controlled by modality-balanced volume caps and repeat factor sampling at the source-volume level. The final operating points are selected from the planned detector and tracker training phases rather than tuned on the official validation metrics.

Table 2. Training protocols for adapting SAM3 to text-prompted 3D medical segmentation.

Pretrained backbone	Public SAM3 checkpoint
Backbone tuning strategy	Detector full tuning; tracker-module tuning with detector frozen
Batch size	256 global batch
Patch / volume size	504×504 2D canvas; 8-slice tracker clips
Total epochs	10 detector epochs; 10 tracker epochs
Optimizer	AdamW, bf16
Initial learning rate	Detector main 2.24×10^{-4}; vision 7.0×10^{-5}; language 1.4×10^{-5}; tracker 5×10^{-5}
LR decay schedule	Inverse-square-root schedule
Class-imbalance handling	Modality-balanced caps, volume-level RFS, PET negative cap
Training time	Detector ∼20 h; tracker ∼28 h on 8×A100
Loss function	SAM3 detection, box/GIoU, mask focal/Dice, and tracker propagation losses
Number of trainable parameters	Full SAM3 detector in Phase 1; tracker memory/decoder modules in Phase 2

4.3 Post-processing and Inference Efficiency

Detector predictions are mapped back to the original slice coordinates. For patch-split prompts, quadrant predictions are merged in the original image space before per-class aggregation. We did not perform separate kernel-level, Docker-startup, or deployment-specific runtime optimization.

4.4 Main MedSegFM Benchmark

Tables 3 and 4 report the main MedSegFM validation results. Among public-data text-prompted methods, Dr. SAM3-D gives the best mean instance F1, while Dr. SAM3-T gives the strongest CT and MR instance F1. This shows that detector-only and detector-tracker serve different operating regimes.

Instance segmentation. Dr. SAM3-D reaches 40.48 mean instance F1, ahead of Medal-S (39.00), SAT (25.29), CAT (14.47), and Text3DSAM (11.38). It also uses far less MedSegFM fine-tuning compute than Medal-S, 160 GPU-hours versus 5,171 GPU-hours. Dr. SAM3-T reaches a similar mean instance F1 of 40.08 and is strongest on CT and MR instance F1, with 51.59 on CT and 57.38 on MR. These results indicate that SAM3's text-conditioned detector transfers well after modality and volume-aware adaptation, and that the tracker further helps CT/MR object recovery.

Semantic segmentation. Dr. SAM3-D obtains 78.07 mean DSC. This is the strongest result among public-data text-prompted methods, ahead of Medal-S (75.41), CAT (70.73), SAT (69.83), and Text3DSAM (67.57). Dr. SAM3-D is also

Table 3. Instance results on the MedSegFM 3D validation set. Values are F1/DSC-TP %; Mean averages the displayed instance-modality F1 columns. **Bold** indicates the best while underline indicates the second-best; best/second-best are marked among text-prompted models with open medical data. † marks data used for pretraining without MedSegFM. FT GPUh counts MedSegFM fine-tuning GPU hours.

			CT		MR		PET		Micro		Mean
Model	Training data	FT GPUh	F1	DSC-TP	F1	DSC-TP	F1	DSC-TP	F1	DSC-TP	F1
SAM3 [1]	–	–	2.81	2.94	0.00	0.00	0.00	0.00	0.00	0.00	0.70
MedSAM3 [15]	4 datasets	–	22.10	45.08	13.06	37.78	9.39	48.13	14.76	29.07	14.83
Medical SAM3 [12]	33 datasets	–	4.10	7.98	0.70	2.47	0.00	0.00	5.11	14.59	2.48
VoxTell [27]	158 datasets	–	3.24	11.94	3.80	9.89	16.93	52.31	0.00	0.00	5.99
Text3DSAM [32]	MedSegFM	240	11.46	19.95	11.63	33.29	8.98	49.95	13.44	55.05	11.38
CAT [10]	MedSegFM	576	30.03	37.00	13.75	28.13	10.98	27.79	3.13	36.28	14.47
SAT [35]	SAT-DS† + MedSegFM	288	25.25	39.40	12.28	27.28	**42.00**	**78.63**	21.62	45.74	25.29
Medal-S [29]	SAT-DS† + MedSegFM	5,171	40.36	51.32	48.43	66.66	33.68	71.02	**33.54**	72.26	39.00
BiomedParse-V [34]	Closed data† + MedSegFM	9,600	51.19	67.49	53.17	70.53	31.32	71.85	20.84	66.96	39.13
Dr. SAM3-T	MedSegFM	384	**51.59**	**55.03**	**57.38**	71.77	25.26	63.32	26.07	70.02	40.08
Dr. SAM3-D	MedSegFM	160	45.48	54.27	56.62	**76.03**	30.99	74.26	28.84	**75.53**	**40.48**

the strongest public-data method on US DSC and NSD, with 90.12 and 91.21. The closed-data BiomedParse-V reference remains stronger in mean semantic DSC, but Dr. SAM3-D exceeds it in mean instance F1.

The main gain is strong public-data instance performance with the strongest open-data mean semantic segmentation. The results do not imply uniform dominance on every modality. PET instance F1 remains better for SAT, CT semantic DSC and NSD remain stronger for Medal-S, and BiomedParse-V is a strong closed-data semantic reference. The overall trend supports our central claim: SAM3 can become a strong text-prompted 3D medical segmentor through data-interface adaptation rather than architecture changes.

4.5 MedSegFM CT/MR Tumor and Lesion Analysis

Table 5 focuses on oncologic CT/MR instance classes. This subset is clinically important and difficult because tumors and lesions are often small, sparse, and sensitive to instance fragmentation.

The subset includes all CT/MR tumor and lesion instance classes in the MedSegFM 3D validation set, covering 11 classes and 293 annotation-volume pairs. We report class-macro instance F1 and compare the same text-prompted baselines as in Table 3.

Dr. SAM3-T achieves the best class-macro instance F1, 52.8. It outperforms BiomedParse-V (49.9), Dr. SAM3-D (42.1), Medal-S (39.4), CAT (28.1), Text3DSAM (12.8), and SAT (7.7). The gain over Dr. SAM3-D is +10.7 pp,

Table 4. Semantic results on the MedSegFM 3D validation set. Values are DSC/NSD %; Mean averages the displayed semantic-modality DSC columns. **Bold** indicates the best while underline indicates the second-best among text-prompted models with open medical data.

Model	Training data	FT GPUh	CT		MR		US		Mean
			DSC	NSD	DSC	NSD	DSC	NSD	DSC
SAM3 [1]	–	–	22.62	21.26	10.76	11.69	0.00	0.00	11.13
MedSAM3 [15]	4 datasets	–	17.87	18.48	18.24	25.00	20.57	25.26	18.89
Medical SAM3 [12]	33 datasets	–	7.39	8.63	15.77	16.91	5.95	14.72	9.70
VoxTell [27]	158 datasets	–	46.55	46.70	38.62	45.20	0.23	1.22	28.47
Text3DSAM [32]	MedSegFM	240	67.07	60.95	52.26	59.35	83.37	81.17	67.57
CAT [10]	MedSegFM	576	72.11	72.27	54.15	61.93	85.94	83.60	70.73
SAT [35]	SAT-DS† + MedSegFM	288	67.80	67.26	56.10	66.69	85.58	79.24	69.83
Medal-S [29]	SAT-DS† + MedSegFM	5,171	**81.87**	**81.58**	61.91	70.85	82.46	79.53	75.41
BiomedParse-V [34]	Closed data† + MedSegFM	9,600	85.12	89.65	73.96	86.64	90.50	91.35	83.19
Dr. SAM3-T	MedSegFM	384	65.99	67.43	61.24	69.12	89.22	89.83	72.15
Dr. SAM3-D	MedSegFM	160	77.80	74.42	**66.29**	**75.23**	**90.12**	**91.21**	**78.07**

Table 5. CT/MR oncologic tumor–lesion subset. Values are instance F1 %.

Mod.	Class	n	Text3DSAM	SAT	CAT	Medal-S	BiomedParse-V	**Dr. SAM3-D**	**Dr. SAM3-T**
CT	Adrenocortical carcinoma	4	25.0	3.2	19.8	58.3	58.3	62.5	**100.0**
	Colon cancer primaries	12	12.5	12.6	29.2	32.1	58.3	57.5	**65.3**
	Lung lesions	16	0.0	9.4	13.4	24.0	44.8	36.5	**45.7**
	Head-neck cancer	24	62.5	14.2	89.6	86.1	**100.0**	87.8	80.6
	Lymph node	27	6.8	14.3	17.8	**38.0**	26.0	12.2	22.4
	Pancreas tumors	34	6.9	0.0	13.7	39.9	**58.3**	37.3	54.9
	Lesion	45	0.5	0.7	8.3	**9.4**	7.8	6.7	6.9
	Liver tumors	48	12.5	0.7	40.1	44.6	59.4	63.7	**70.8**
	Kidney lesions	48	7.3	7.7	39.5	56.5	**76.7**	63.2	72.3
MR	Prostate lesion	6	0.0	0.0	0.0	0.0	**11.1**	0.0	0.0
	GTVp and GTVn tumor	29	7.1	22.1	38.1	44.8	47.8	35.7	**61.9**
	Class avg.	**11 cls.**	12.8	7.7	28.1	39.4	49.9	42.1	**52.8**

showing that tracker adaptation is most useful when the target is sparse and object-level continuity matters.

This analysis separates instance recovery from semantic overlap. Instance F1 penalizes missed, fragmented, and poorly propagated objects. These are common failure modes for short-range tumor and lesion targets. The tumor and lesion subset therefore highlights the complementary role of Dr. SAM3-T: Dr. SAM3-D is the strongest overall detector-only setting in Table 3, while Dr. SAM3-T is preferable when preserving individual oncologic objects across adjacent slices is the priority.

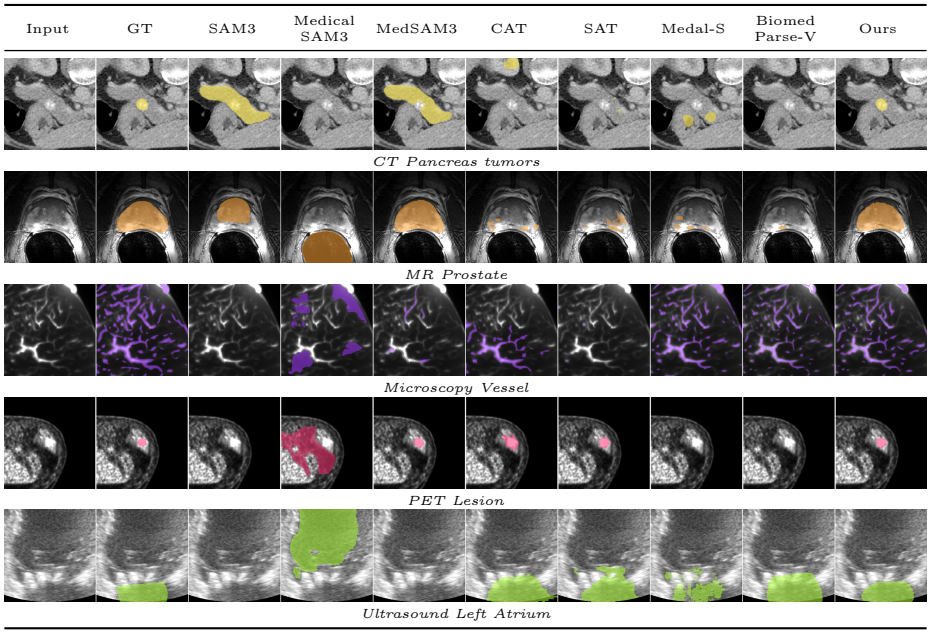

Fig. 3. Qualitative segmentation comparisons on MedSegFM 3D validation set. Each row shows one text-prompted example across the input, ground truth, baselines, and Dr. SAM3-T.

4.6 RAOS-Real External Transfer

Table 6 evaluates external CT generalization on RAOS-Real. This test measures whether the adapted SAM3 interface transfers beyond the MedSegFM validation distribution.

RAOS-Real is an abdominal-organ robustness benchmark from RAOS [17] that extends WORD [18]. We train Dr. SAM3 only on the MedSegFM training pool. We do not use RAOS-specific training, validation tuning, or prompt tuning. At inference, each organ name is prefixed with the modality, for example `CT Liver`, following Section 3.1. Medal-S uses components pretrained on SAT-DS, which includes WORD, so we treat Medal-S as a partially exposed reference rather than a strictly external baseline.

Dr. SAM3-D achieves the best 12-organ average DSC, 77.2. It is followed by Dr. SAM3-T (71.7), BiomedParse-V (58.8), Medal-S (43.9), and Text3DSAM (18.7). The improvement is broad: Dr. SAM3-D is best on 10 of the 12 organs, and Dr. SAM3-T is best on the remaining two, right kidney and gallbladder. On challenging structures, Dr. SAM3-D reaches 89.6 DSC on bladder versus 66.2 for the strongest non-Dr. SAM3 baseline, 67.9 on intestine versus 49.5, and 53.0 on duodenum versus 34.5. Dr. SAM3-T reaches 55.4 on gallbladder versus 31.5.

The RAOS-Real results suggest that the proposed adaptation preserves SAM3's pretrained visual-language priors while aligning the input interface with medical

Table 6. External CT generalization on RAOS-Real. Per-organ DSC percentages for displayed organs. Per-organ NSD is reported in Appendix E.1.

Model	Liver	Left kidney	Right kidney	Spleen	Bladder	Stomach	Pancreas	Esophagus	Intestine	Colon	Gallbladder	Duodenum	Class avg.
Medical SAM3	8.1	9.8	8.0	10.7	0.2	4.9	3.7	0.1	3.0	15.5	0.2	0.4	5.4
MedSAM3	15.1	1.8	13.2	16.9	0.7	18.1	10.5	0.5	12.2	2.1	8.1	2.5	8.5
Text3DSAM	51.1	25.1	25.9	38.8	0.6	23.3	14.0	18.5	8.1	8.6	1.4	9.0	18.7
Medal-S	80.4	50.6	58.3	45.0	39.9	51.7	48.5	13.9	49.5	52.8	7.6	28.0	43.9
BiomedParse-V	88.3	84.5	77.7	82.6	66.2	73.1	59.9	67.5	0.9	38.6	31.5	34.5	58.8
Dr. SAM3-T	91.4	91.5	**90.0**	89.7	85.1	82.7	70.9	71.6	46.9	52.3	**55.4**	32.4	71.7
Dr. SAM3-D	**93.2**	**91.7**	89.1	**90.0**	**89.6**	**86.2**	**76.7**	**73.8**	**67.9**	**62.2**	52.7	**53.0**	**77.2**

Table 7. Detector training ablation. DSC (%) on a held-out 2D diagnostic split from the 10% training coreset. Δ_{rare} is the rare-class mean change relative to the modality-prefix row.

Variant	CT	MR	PET	US	Micro	Mean	Δ_{rare}
Plain detector fine-tuning	**83.65**	47.69	5.36	29.16	26.26	38.42	–
+ Modality-balanced volume cap	82.02	46.91	38.43	**77.31**	28.38	54.61	–
+ Patch split	81.88	48.29	41.64	76.03	38.01	57.17	–
+ Modality-prefix prompts	82.05	49.79	42.97	75.31	39.23	57.87	–
+ Capped volume RFS, $t = 0.01$	82.74	**53.51**	36.77	76.55	**40.64**	58.04	+11.51
+ Oversampled volume RFS, $t = 0.01$	82.60	52.61	35.18	74.76	39.97	57.02	+12.40
+ Ratio-triggered negative-slice cap	82.66	52.53	**45.79**	72.66	40.44	**58.82**	**+12.67**

CT. Since RAOS-Real is excluded from training and prompt tuning, the gains are unlikely to be explained by overfitting to the MedSegFM validation set.

4.7 Ablations

Detector adaptation. Table 7 gives an additive detector ablation on a held-out 2D diagnostic split from the 10% training coreset. Plain detector fine-tuning uses a uniform per-volume cap ($K = 10$). Modality-balanced caps raise the mean by +16.19 pp, mainly by restoring PET and US exposure. Patch splitting improves dense microscopic blood vessel segmentation. Modality-prefixed prompts further improve the mean. Oversampled volume RFS gives the main rare-class gain (+12.40 pp relative to the modality-prefix row). The final ratio-triggered negative-slice cap preserves the rare-class gain (+12.67 pp) and improves PET and the overall mean. Appendix Figure 5 visualizes the resulting sampling distribution, and Appendix D.2 reports the negative-slice diagnostic.

Tracker adaptation. Table 8 isolates the slice sequencing choice for tracker training. Replacing SAM3's default random-stride clips with adjacent-slice clips (stride=1) raises mean instance F1 by +10.80 pp and mean DSC-TP by +6.82 pp. The largest gain is on MR, with +24.41 pp instance F1 and +17.42 pp DSC-TP. CT, PET, and Microscopy also improve. This supports the volume-to-video

Table 8. Tracker volume-to-video ablation. Values are instance F1 and DSC-TP (%) under the official 3D validation pipeline.

	CT		**MR**		**PET**		**Micro**		**Mean**	
Tracker interface	F1	DSC-TP	F1	DSC-TP	F1	DSC-TP	F1	DSC-TP	F1	DSC-TP
Default random stride	48.45	51.92	32.97	54.35	18.72	59.83	16.99	66.75	29.28	58.21
Adjacent-slice sequencing (stride=1, ours)	**51.59**	**55.03**	**57.38**	**71.77**	**25.26**	**63.32**	**26.07**	**70.02**	**40.08**	**65.04**

design in Section 3.3: medical volumes should be propagated along adjacent through-plane slices rather than treated as natural videos with random temporal strides.

5 Limitation and Future Work

Our recipe targets short canonical class-name prompts such as `CT Liver`, so robustness to arbitrary clinical free text remains untested. Training and inference use a 504×504 canvas, limiting very fine structures outside the patch-split trigger, and the tracker treats volumes as ordered slice sequences without explicit physical spacing. External generalization is tested only on CT; external MR, PET, Ultrasound, and Microscopy evaluation remains future work. Future work should incorporate physical spacing, richer clinical-language prompts, and external validation across all modalities.

6 Conclusion

We presented Dr. SAM3, a recipe for adapting SAM3 to text-prompted 3D medical segmentation without changing its architecture. The main takeaway is that biomedical volumes should be handled with their modality, source-volume structure, and ordered-slice geometry in mind, rather than flattened into independent 2D images or treated as natural videos. On the MedSegFM 3D validation set, this yields two useful operating points: Dr. SAM3-D for strong slice-wise text-prompted inference, and Dr. SAM3-T for instance continuity and sparse-object recovery, most clearly on oncologic tumor–lesion instance segmentation. The RAOS-Real results further suggest that preserving SAM3's pretrained visual and text priors while adapting the data interface supports external CT transfer under source-domain shift.

Acknowledgements. We thank all data owners for making the biomedical images publicly available and the challenge organizers for hosting the benchmark platform.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Carion, N., Gustafson, L., Hu, Y.T., et al.: SAM 3: Segment anything with concepts. In: International Conference on Learning Representations (2026) 2, 5, 11, 12
2. Chen, C., Miao, J., Wu, D., Zhong, A., Yan, Z., Kim, S., Hu, J., Liu, Z., Sun, L., Li, X., et al.: MA-SAM: Modality-agnostic SAM adaptation for 3D medical image segmentation. Medical Image Analysis **98**, 103310 (2024). https://doi.org/10.1016/j.media.2024.103310 4
3. Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., et al.: TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis **97**, 103280 (2024). https://doi.org/10.1016/j.media.2024.103280 2, 4
4. Du, Y., Bai, F., Huang, T., Zhao, B.: SegVol: Universal and interactive volumetric medical image segmentation. In: Advances in Neural Information Processing Systems. vol. 37 (2024). https://doi.org/10.52202/079017-3516 4
5. Gong, S., Zhong, Y., Ma, W., Li, J., Wang, Z., Zhang, J., Heng, P.A., Dou, Q.: 3DSAM-adapter: Holistic adaptation of SAM from 2D to 3D for promptable tumor segmentation. Medical Image Analysis **98**, 103324 (2024). https://doi.org/10.1016/j.media.2024.103324 4
6. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Lecture Notes in Computer Science, vol. 12962, pp. 272–284. Springer (2022). https://doi.org/10.1007/978-3-031-08999-2_22 4
7. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: UNETR: Transformers for 3D medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584 (2022) 2, 4
8. He, Y., Guo, P., Tang, Y., et al.: VISTA3D: A unified segmentation foundation model for 3D medical imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20863–20873 (2025). https://doi.org/10.1109/CVPR52734.2025.01943 4
9. Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., et al.: Segment anything model for medical images? Medical Image Analysis **92**, 103061 (2024). https://doi.org/10.1016/j.media.2023.103061 4
10. Huang, Z., Jiang, Y., Zhang, R., Zhang, S., Zhang, X.: CAT: Coordinating anatomical-textual prompts for multi-organ and tumor segmentation. In: Advances in Neural Information Processing Systems. vol. 37 (2024). https://doi.org/10.52202/079017-0119 5, 11, 12
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 2, 4, 23, 30

12. Jiang, C., Ding, T., Song, C., et al.: Medical SAM3: A foundation model for universal prompt-driven medical image segmentation (2026), arXiv preprint 5, 11, 12
13. Jiang, Y., Huang, Z., Zhang, R., Zhang, X., Zhang, S.: ZePT: Zero-shot pan-tumor segmentation via query-disentangling and self-prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11386–11397 (2024) 5
14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023) 2, 4
15. Liu, A., Xue, R., Cao, X.R., et al.: MedSAM3: Delving into segment anything with medical concepts (2025), arXiv preprint 5, 11, 12
16. Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., Landman, B.A., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: CLIP-driven universal model for organ segmentation and tumor detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21152–21164 (2023) 5
17. Luo, X., Li, Z., Zhang, S., Liao, W., Wang, G.: Rethinking abdominal organ segmentation (RAOS) in the clinical scenario: A robustness evaluation benchmark with challenging cases. In: Medical Image Computing and Computer Assisted Intervention. Lecture Notes in Computer Science, vol. 15009, pp. 531–541. Springer Nature Switzerland (2024). https://doi.org/10.1007/978-3-031-72114-4_51 4, 13
18. Luo, X., Liao, W., Xiao, J., et al.: WORD: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from CT image. Medical Image Analysis **82**, 102642 (2022). https://doi.org/10.1016/j.media.2022.102642 13
19. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 654 (2024). https://doi.org/10.1038/s41467-024-44824-z 2, 4, 30
20. Ma, J., Kim, S., Zhou, Y., Wang, B.: CVPR-BiomedSegFM: Cvpr 2025 competition on foundation models for 3D biomedical image segmentation. Hugging Face dataset (2025), https://huggingface.co/datasets/junma/CVPR-BiomedSegFM, accessed 2026-05-06 4, 8, 24
21. Ma, J., Li, F., Wang, B.: U-Mamba: Enhancing long-range dependency for biomedical image segmentation (2024), arXiv preprint 4
22. 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 (2025), arXiv preprint 2, 4, 30
23. Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment anything model for medical image analysis: An experimental study. Medical Image Analysis **89**, 102918 (2023). https://doi.org/10.1016/j.media.2023.102918 4
24. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: International Conference on 3D Vision. pp. 565–571. IEEE (2016). https://doi.org/10.1109/3DV.2016.79 2, 4
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning. pp. 8748–8763. PMLR (2021) 5
26. Ravi, N., Gabeur, V., Hu, Y.T., et al.: SAM 2: Segment anything in images and videos. In: International Conference on Learning Representations (2025) 2, 4

27. Rokuss, M., Langenberg, M., Kirchhoff, Y., et al.: VoxTell: Free-text promptable universal 3D medical image segmentation (2025), arXiv preprint 11, 12
28. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015). https://doi.org/10.1007/978-3-319-24574-4_28 4
29. Shi, P., Chen, J., Liu, J., Zhang, X., Chen, T., Li, L.: Medal S: Spatio-textual prompt model for medical segmentation (2025), arXiv preprint 5, 11, 12
30. Wang, H., Guo, S., Ye, J., et al.: SAM-Med3D: Towards general-purpose segmentation models for volumetric medical images. In: European Conference on Computer Vision Workshops. pp. 51–67. Springer (2025). https://doi.org/10.1007/978-3-031-91721-9_4 4
31. Wasserthal, J., Breit, H.C., Meyer, M.T., et al.: TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. Radiology: Artificial Intelligence **5**(5), e230024 (2023). https://doi.org/10.1148/ryai.230024 4
32. Xin, Y., Ates, G.C., Shao, W.: Text3DSAM: Text-guided 3D medical image segmentation using SAM-inspired architecture. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop on Foundation Models for Text-Guided 3D Biomedical Image Segmentation (2025) 11, 12
33. Zhao, T., Gu, Y., Yang, J., et al.: A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. Nature Methods **22**(1), 166–176 (2025). https://doi.org/10.1038/s41592-024-02499-w 5
34. Zhao, T., Lee, H.H., Santamaria-Pang, A., et al.: BiomedParse-V: Scaling foundation model for universal text-guided volumetric biomedical image segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop on Foundation Models for Text-Guided 3D Biomedical Image Segmentation (2025) 5, 11, 12
35. Zhao, Z., Zhang, Y., Wu, C., et al.: Large-vocabulary segmentation for medical images with text prompts. npj Digital Medicine **8**, 566 (2025). https://doi.org/10.1038/s41746-025-01964-w 5, 11, 12

A Training Details

Training follows the detector-then-tracker schedule in Section 4.2. The detector phase updates the full SAM3 image detector from the public SAM3 checkpoint using the full MedSegFM detector pool and volume-level repeat factor sampling enabled for CT/MR groups. The tracker phase starts from the detector checkpoint, freezes the detector and prompt/image pathways, and updates only tracker propagation modules. Neither phase uses official 3D validation metrics for checkpoint selection; final reporting uses the official 3D validation set.

The benchmark provides class names as long descriptive sentences (e.g., "Visualization of the pancreas in abdominal CT imaging"), but SAM3's prompt encoder is designed for short noun phrases. We therefore use the benchmark metadata to map each class description to a canonical anatomical noun phrase. The same mapping is applied consistently during both training and evaluation.

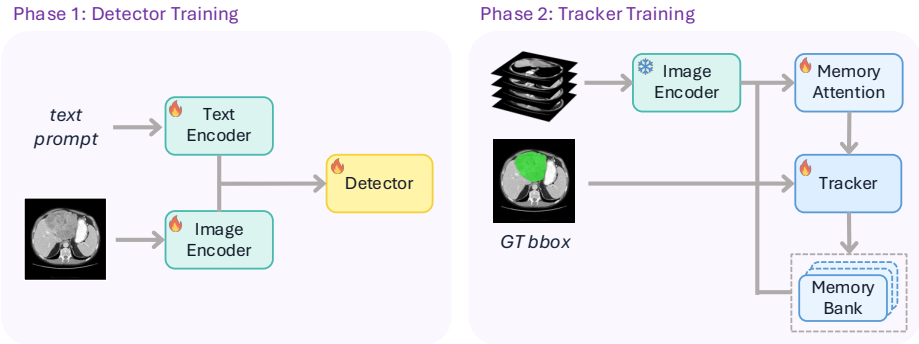

Fig. 4. Detector–tracker training pipeline. Phase 1 fine-tunes the SAM3 detector on 2D medical slices, and Phase 2 freezes the detector pathway while fine-tuning tracker propagation modules on volume-derived slice clips.

SAM3 uses 4-slice tracker clips by default. Table 10 compares this default with an 8-slice random-stride clip and adjacent-slice sequencing (stride=1). The 4-slice default is stronger on several MR metrics, but the 8-slice setting improves CT instance metrics, MR DSC-TP, US semantic metrics, and PET/Microscopy instance propagation by large margins under the random-stride interface. The final tracker further replaces random stride with adjacent-slice sequencing (stride=1), yielding the strongest instance propagation metrics.

B Training Data Composition

Training-pool scale and imbalance diagnostics. The official MedSegFM training pool comprises 44.6K volume groups spanning 4.3M annotated 2D slices/frames, with 17.4M annotations across 257 observed modality-prefixed classes. Per-class

slice counts span four orders of magnitude. In our frequency analysis, 204 of the 257 classes (79%) trigger volume RFS scaling ($r > 1$ at $t = 0.01$), and the rare tail extends to single-volume classes such as CT vasculature pelvis, which contains 12 source slices from one volume.

The raw training distribution is strongly modality-skewed. CT accounts for 81.7% of annotated slices, whereas PET, US, and Microscopy together contribute under 6%. Negative-slice ratios and annotation density are also modality dependent: PET reaches a 48.8% negative-slice ratio, US is nearly always positive, and a small number of high-density slices exceed SAM3's 200-query detector capacity. These dataset-side properties motivate the modality-balanced volume cap, volume-level repeat factor sampling, density-triggered patch split, and ratio-triggered negative-slice calibration described in Section 3.1.

B.1 Source Dataset Counts

Our recipe trains on the official MedSegFM training pool, organized into 75 source datasets across 5 modalities. Detector and tracker share the same source datasets but generate different training artifacts: the detector consumes 2D slices/frames, while the tracker consumes generated 3D videos. Table 11 reports per-modality totals, and Table 12 lists every source dataset with its detector image count and tracker video count.

Observations. The training pool is heavily long-tailed at the dataset level: a single CT dataset (AbdomenAtlas) accounts for 1.96M detector images (45.6% of total), while 14 datasets contribute fewer than 5,000 detector images each. Modality coverage is similarly imbalanced: PET has a single source dataset (autoPET, 62 raw volumes), and Microscopy aggregates 11 small datasets totaling 14 raw volumes. This long tail at both the dataset and modality level is the primary motivation for our modality-balanced volume caps (Section 3.1) and volume-level repeat factor sampling (Section 3.2).

B.2 Rare-Class Subset ($N = 31$)

The rare-class subset used to compute the rare-class ΔDice summary in Section 4.7 covers classes affected by low frequency, limited source-volume diversity, and small-volume tumor/lesion structure. The 31 classes are summarized in Table 13.

C Evaluation Details

We use the official 3D validation set only for final reporting. Hyperparameters such as the RFS threshold, modality caps, negative-slice trigger, detector threshold, and tracker post-processing were fixed from training-set statistics or architectural constraints, not selected by official 3D validation performance. Detector-side ablations are run on a held-out 2D diagnostic split from the 10% training coreset, i.e., a split carved from the training data, and are not used for checkpoint selection.

Detector ablation protocol. Table 7 uses 4-epoch detector runs on the held-out 2D diagnostic split from the 10% training coreset. Plain detector fine-tuning uses a uniform per-volume cap ($K = 10$), and subsequent rows add components cumulatively. The volume RFS rows use the modality-agnostic algorithm in Appendix D.1, enabled only for CT/MR volume groups in our final recipe. Δ_{rare} is computed on the $N = 31$ rare-class subset relative to the modality-prefix row.

For BiomedParse-V, SAT, and CAT, we score the per-volume prediction outputs released by the challenge organizers. Other text-prompted baselines in Tables 3 and 4 were run by us on the same official 3D validation set. For VoxTell, Text3DSAM, and Medal-S, we use the authors' public inference code and released checkpoints when available, adapting dataset paths, class prompts, and output formatting to the official evaluator. For SAM3-family rows, including base SAM3, MedSAM3, and Medical SAM3, we use a detector–tracker 3D evaluation setup: MedSAM3 and Medical SAM3 use their released detector weights with the base SAM3 tracker attached according to the detector–tracker setup described in their papers. We also add the MedSegFM-specific adapters needed for a fair text-prompted 3D run: modality-prefixed prompts and density-triggered patch split. We do not tune baseline-specific thresholds on the official 3D validation set.

Unless otherwise stated, detector-side thresholds and NMS follow each SAM3-based row's own defaults: official SAM3 defaults for the SAM3 reference row and our models, and released repository defaults for MedSAM3 and Medical SAM3. When a base SAM3 tracker is attached for 3D evaluation, tracker-side thresholds and mask/track merging use official SAM3 defaults. We do not tune these values on the official MedSegFM 3D validation set.

D Sampling Algorithm and Ablations

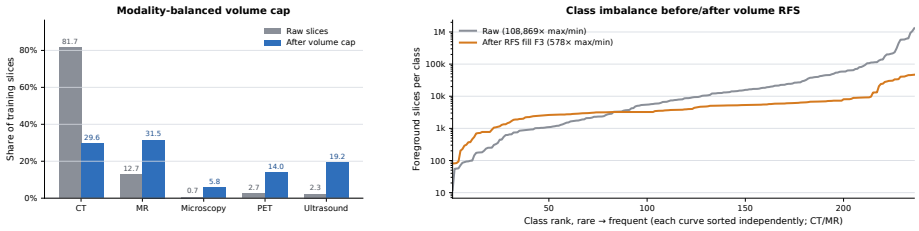

Fig. 5. Effect of modality caps and volume RFS on detector sampling. The sampler compresses the raw CT-dominated slice distribution while increasing effective exposure for smaller modalities and rare-class volumes.

Modality-cap details. The modality caps are computed from training-set volume counts only. In the final detector sampler, $K_{\text{CT}} = 4$ and $K_{\text{MR}} = 7$, while PET,

US, and Microscopy are uncapped. This changes the full-data detector epoch mix to approximately 151K CT, 161K MR, 98K US, 72K PET, and 29K Microscopy slices, reducing CT from 81.7% of the raw slice pool to about 30% of a sampled epoch.

D.1 Volume RFS Algorithm and Threshold Choice

Algorithm 1 Per-epoch sampling with Capped or Oversampled volume RFS.

Require: Training set of volume groups; modality-wise non-empty-slice class frequencies f_c; threshold t; per-modality base K_m; RFS modalities $\mathcal{M}_{\text{RFS}}$; oversampling $\in \{\text{off}, \text{on}\}$

Ensure: List of (volume, slice index) training samples for the epoch

for each volume group g of modality m **do**
  $n_g \leftarrow$ number of unique annotated slices in g
  if $m \notin \mathcal{M}_{\text{RFS}}$ **then**
    Sample $\min(n_g, K_m)$ slices from g without replacement ▷ base sampler
  else
    $r_g \leftarrow 1$
    for each class c with at least one annotated slice in g **do**
      $r_c \leftarrow \max(1, \sqrt{t/f_c})$ ▷ slice-level per-class repeat factor
      $r_g \leftarrow \max(r_g, r_c)$
    end for
    $K_{\text{eff}} \leftarrow \text{round}_s(K_m \cdot r_g)$ ▷ effective slices sampled from volume g this epoch
    if $K_{\text{eff}} \leq n_g$ **then**
      Sample K_{eff} slices from g without replacement
    else if oversampling = off **then**
      Take all n_g slices once ▷ Capped volume RFS
    else
      Take all n_g slices once, then sample $(K_{\text{eff}} - n_g)$ additional slices from g with replacement
    end if
  end if
end for

Threshold choice. Algorithm 1 exposes the threshold t, which controls how broadly volume RFS amplifies the long tail. In our training-set frequency analysis, the LVIS default $t = 0.001$ affects too few volume groups, so we use $t = 0.01$ to cover the rare-class tail more broadly. Table 14 shows that Capped volume RFS on CT/MR groups with $t = 0.001$ gives only a +2.61 pp gain on the rare-class subset, compared with +11.51 pp for $t = 0.01$. Oversampled volume RFS at $t = 0.01$ further raises the rare-class gain to +12.40 pp, although it slightly lowers the aggregate mean before ratio-triggered negative-slice calibration. We therefore use Oversampled volume RFS with $t = 0.01$ on CT/MR groups in the final recipe.

D.2 Negative-Slice Calibration Diagnostic

Table 15 reports the training-set negative-slice ratio by modality under the final detector sampler analysis. The ratio is computed as zero-annotation negative slices divided by raw slices within each modality. A modality triggers ratio-based negative-slice calibration when this ratio exceeds 20%; PET is the only triggered modality, with a 48.8% raw negative-slice ratio, and its final sampler caps negative slices at 20%.

Table 16 gives the Neg-FP diagnostic on a held-out negative-slice subset. Only the detector regime matters for this: the tracker cannot rescue a detector that produces non-empty masks on target-absent slices. The collapse under positive-only training, especially MR where 94.6% of negative slices receive a non-empty prediction when re-prompted, motivates retaining natural negative slices for calibration. A 33% positive-forcing baseline performs worse than the natural-ratio baseline (aggregate 0.043 vs. 0.019), indicating that global positive/negative forcing is less effective than ratio-triggered calibration.

E Additional Quantitative Results

This section collects quantitative analyses that support the main results but are not part of the primary text-prompted ranking in Table 3.

E.1 RAOS-Real Per-Organ Results

Table 17 reports RAOS-Real NSD for the displayed organs and the unweighted class average over the displayed 12 organs.

E.2 Spatial-Prompted Baselines

Spatial-prompted models use oracle bounding-box prompts and therefore solve a different task from text-prompted segmentation. We report them separately in Table 18 for context, not as part of the text-prompted ranking set in Table 3.

E.3 nnU-Net Per-Modality Specialist Comparison

nnU-Net [11] is trained as a per-modality specialist (one model per modality) on the official MedSegFM 10% coreset, with no text prompting. We use nnU-Net v2, the official second-generation reimplementation; it improves code structure and extensibility while preserving the original self-configuring methodology. Because both its training regime (per-modality, smaller data) and its inference setup (no shared text-prompted multi-modality model) differ from every other entry in Tables 3 and 4, we keep it out of the main comparison and report it here against both Dr. SAM3 operating points. All rows are scored under the same official MedSegFM evaluation pipeline.

The modality split is informative. On CT and MR, both Dr. SAM3 operating points substantially beat the per-modality nnU-Net specialists across all four metrics. Dr. SAM3-T is strongest for CT/MR instance F1 and CT DSC-TP, while Dr. SAM3-D is strongest for semantic DSC/NSD and MR DSC-TP. On the smaller modalities, nnU-Net remains strongest on PET instance metrics, whereas Dr. SAM3-D is strongest on US semantic DSC/NSD and both Microscopy metrics. We do not claim our recipe replaces specialists on every modality; the contribution is a single text-prompted multi-modality system that outperforms specialists on the two largest-data modalities while remaining competitive on the rest.

F Broader Impacts

Dr. SAM3 is a research system for text-prompted 3D biomedical image segmentation. Its potential positive impact is to reduce the amount of spatial prompting required for volumetric segmentation, enabling users to request targets with short modality-aware class names such as "CT Liver" or "MR Prostate". If made reliable, this direction could support scalable annotation, model prototyping, and retrospective analysis across heterogeneous biomedical modalities.

At the same time, the system is not intended for direct clinical deployment. Incorrect segmentations could affect downstream measurements, treatment planning, or clinical interpretation if used without expert review. This risk is especially important for small tumors, sparse lesions, fine structures, and modalities where performance remains lower, such as PET and some Microscopy classes. The model also relies on benchmark class names and fixed prompt formats, so robustness to arbitrary clinical free text, institution-specific terminology, acquisition protocols, and population shifts remains untested.

Responsible use therefore requires human oversight, task-specific validation, and calibration on the intended deployment distribution before any clinical or operational use. The model should be treated as an assistive segmentation tool rather than an autonomous diagnostic system. We also note that the present work evaluates existing benchmark data and released model resources; it does not introduce new patient data or make deployment decisions.

G Licenses for training dataset

This work uses existing public datasets, benchmark resources. The primary training and evaluation benchmark is MedSegFM [20]. According to the benchmark release, the dataset is distributed under the CC-BY-NC-SA license: the images are publicly available to the research community, while the ground-truth annotations are hidden on the challenge platform for long-term benchmarking. We follow the benchmark usage protocol and evaluate through the official validation pipeline.

H Additional Qualitative Results

Figure 6 and Figure 7 illustrates additional qualitative results.

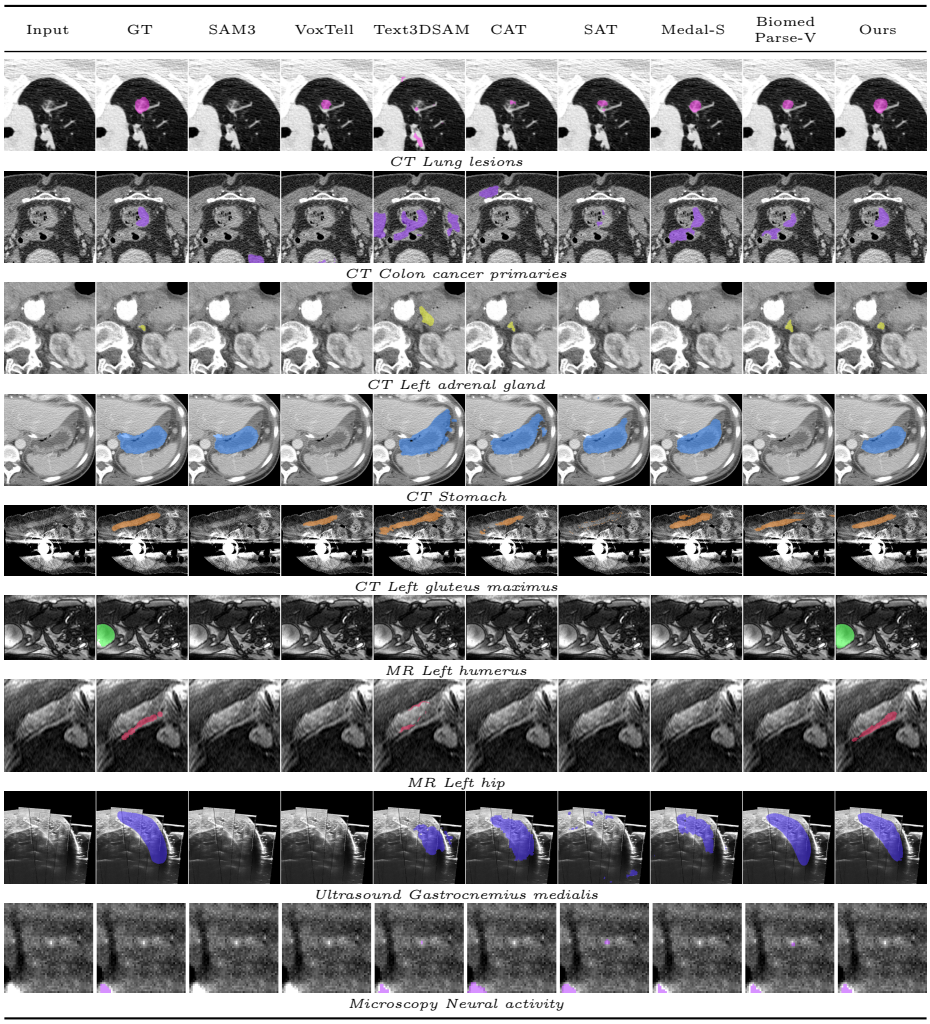

Fig. 6. Additional qualitative segmentation comparisons extending Figure 3 with VoxTell and Text3DSAM. Each row shows one text-prompted example across the input, ground truth, external baselines, and Dr. SAM3-T.

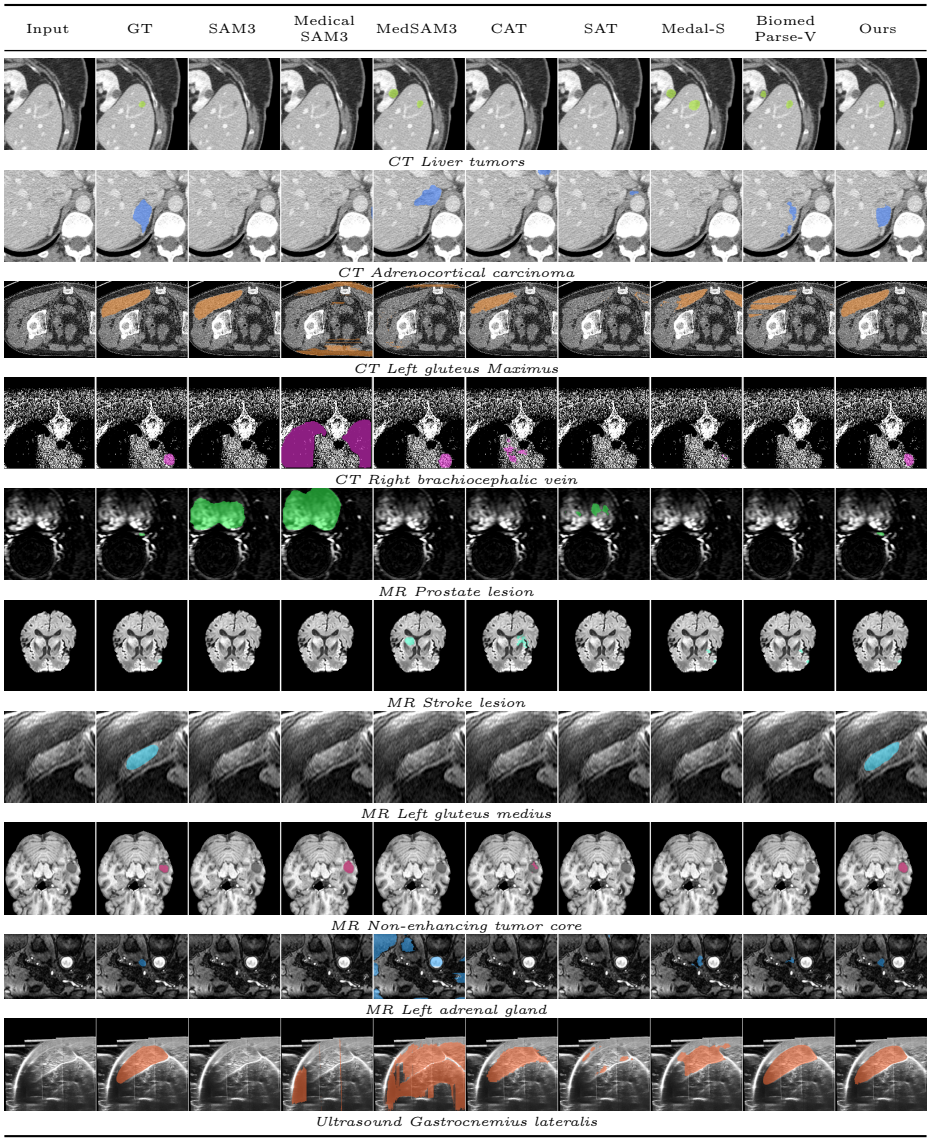

Fig. 7. Additional qualitative segmentation comparisons extending Figure 3. Each row shows one text-prompted example across the input, ground truth, external baselines, and Dr. SAM3-T.

Table 9. Final training configuration. Core hyperparameters for the detector and tracker phases.

Phase	Parameter	Value
Both	Hardware	8× A100 80G
	Global batch	256
	Epochs	10
	Optimizer	AdamW, bf16
	Weight decay	0.1
	Gradient clip	0.1
	LR schedule	Inverse-square-root
Detector	Trainable	Detector, image encoder, text encoder
	Frozen	Tracker
	Data	4.3M slices, 17.4M annotations, 257 classes
	Prompt/canvas	Modality-prefixed names, 504×504
	RFS threshold	$t = 0.01$ for CT/MR volume groups
	Volume caps	$K_{\mathrm{CT}} = 4$, $K_{\mathrm{MR}} = 7$
	Epoch mix	CT 151K, MR 161K, US 98K, PET 72K, Micro 29K
	Steps	20.0K
	Training time	∼20 h, 160 A100 GPU-hours
	LR	Main 2.24×10^{-4}, vision 7.0×10^{-5}, language 1.4×10^{-5}
	Loss weights	Box/GIoU = 5/2, class/presence = 20/20, mask focal/Dice = 200/10, semantic/presence/Dice = 20/1/30
Tracker	Trainable	Tracker modules: memory attention, memory encoder, mask decoder
	Frozen	Detector, prompt/text pathway, image backbone
	Clip length	8 slices
	Stride	1
	Temporal layout	Center-to-periphery sequencing; random time reversal off
	Padding/input	Short-clip padding, 504×504, RandomResize min 400
	Steps	15.8K
	Training time	∼28 h, 224 A100 GPU-hours
	LR	Tracker 5×10^{-5}, frozen modules 0
	Loss weights	Box/GIoU = 10/10, occlusion = 10, focal/Dice = 200/10

Table 10. Tracker clip length and volume-to-video interface ablation. Values are Dice/F1 % on the official 3D validation pipeline. The first two rows compare SAM3-style random-stride tracker clips with 4 versus 8 slices. The final row uses adjacent-slice sequencing (stride=1) with center-to-periphery ordering and masked short-clip padding.

		CT				MR				US		PET		Micro	
Clip	Interface	DSC	NSD	F1	DSC-TP	DSC	NSD	F1	DSC-TP	DSC	NSD	F1	DSC-TP	F1	DSC-TP
4 slices	Default random stride	61.63	63.76	47.40	49.92	61.40	69.32	41.55	49.96	88.78	89.40	14.76	33.03	14.04	47.32
8 slices	Default random stride	**62.30**	**63.80**	48.45	51.92	60.00	68.56	32.97	54.35	**89.12**	**89.76**	18.72	59.83	16.99	66.75
8 slices	Adjacent-slice sequencing (stride=1)	61.20	63.72	**51.59**	**55.03**	**61.71**	**69.54**	**57.38**	**71.77**	89.00	89.74	**25.26**	**63.32**	**26.07**	**70.02**

Table 11. Training data summary by modality. Detector training images are 2D slices/frames under each modality root. Tracker counts are generated training videos, not raw subject counts.

Modality	#datasets	Detector imgs	Tracker vids
CT	29	3,516,524	59,227
MR	30	546,053	25,499
PET	1	115,223	1,246
Ultrasound	4	98,080	2,244
Microscopy	11	29,478	698
Total	**75**	**4,305,358**	**88,914**

Table 12. Per-dataset training counts across all 75 source datasets. "Det." is detector 2D image count; "Trk." is tracker generated video/volume count.

CT (29 datasets)

Dataset	Det.	Trk.
AbdomenAtlas	1,961,552	16,252
TotalSeg-muscles	281,051	2,194
TotalSeg-cardiac	237,349	2,114
TotalSeg-organs	232,894	2,118
TotalSeg-vertebrae	166,861	1,114
AirwayTree	114,966	560
Abdomen1K	110,739	2,000
DeepLesion	105,844	9,429
LungLesion	43,035	13,938
Lungs	42,460	838
WholeBodyTumor	39,986	816
LymphNode	29,968	1,962
AMOS	27,186	540
LiverTumor	20,912	1,312
Aorta	19,101	102
PancreasTumor	14,598	1,186
KidneyTumor	12,569	882
Aeropath	11,033	54
Aortic-Dissection	10,325	80
TCIA-LCTSC	7,137	108
COVID19-Infection	6,226	396
TopCoW24	5,387	250
SegRap-HNTumor	4,820	432
HaN-Seg	4,199	76
AdrenalTumor	1,710	96
MedSegLiver	1,647	100
MedSegIVC	1,165	40
ColonTumor	1,142	228
MedSeg	662	10

MR (30 datasets)

Dataset	Det.	Trk.
BraTS-T2w	80,532	2,430
BraTS-T1c	80,153	2,430
BraTS-T2f	79,959	2,430
BraTS-T1n	79,735	2,430
LiverTumor-LLD	62,687	7,205
TotalSeg	41,597	454
Spider-Vertebrae	21,537	806
HVSMR	16,509	240
Spider-IVD	14,255	806
AMOS	9,469	108
Spider-Spine	8,904	806
HNTS-MRG-HeadTumor	8,234	522
TopCoW24	7,688	250
ProstateT2	5,149	772
ISLES-ADC	3,364	424
ISLES-DWI	3,359	424
ProstateADC	3,287	514
IBD	2,987	228
WMH-T1	2,896	306
WMH-FLAIR	2,837	306
Heart-ACDC	2,384	540
crossMoDA-Tumor-Cochlea	2,220	410
Spine2014	2,169	344
LeftAtrium	1,201	36
SegThy	1,105	28
RESECT	744	46
CHAOS-T1	437	36
CHAOS-T2	420	36
QIN-PROSTATE-Lesion	179	118
CervicalCancer	56	14

Microscopy / US / PET

Dataset	Det.	Trk.
Microscopy (11)		
MUCIC-HL60	10,127	240
SELMA3D-vessel	5,145	168
nucmm	3,136	98
SELMA3D-ADplaques	2,335	60
NIS3D	1,803	12
SELMA3D-nucleus	1,754	22
SELMA3D-neural-activity	1,351	36
urocell-Endolysosome	1,280	10
urocell-Mitochondria	1,280	10
AxonEM	894	36
cremi	373	6
Ultrasound (4)		
US-Low-limb-Leg	40,006	64
US-RESECT	28,977	316
US-Cardiac	16,364	1,800
US-SegThy	12,733	64
PET (1)		
autoPET	115,223	1,246

Table 13. Rare-class subset by failure mode.

Category	N	Classes
CT HaN-Seg head OAR (volume RFS recoverable)	10	Oral cavity, lips, cricopharyngeal inlet, larynx-glottis, left/right submandibular gland, right/left anterior eyeball, buccal mucosa, arytenoids
CT HaN-Seg source-volume floor (extreme rarity, limited volume RFS effect)	8	Right/left cochlea, optic chiasm, pituitary gland, left/right lacrimal gland, left/right optic nerve
CT tumor / lesion	5	Colon cancer primaries, pancreas tumors, kidney lesions, lung lesions, lymph node
MR BraTS / brain tumor	3	Non-enhancing tumor core, resection cavity, enhancing tissue
MR small-volume saturation target	2	Left atrium, intra-meatal schwannoma
MR rare tumor	3	Cervical cancer tumor, extra-meatal schwannoma (crossMoDA), prostate lesion

Table 14. Volume RFS threshold ablation. Values are DSC percentages on a held-out 2D diagnostic split from the 10% training coreset after 4 epochs. Δ_{rare} is the change in mean DSC on the $N = 31$ rare-class subset relative to the modality-prefix row in Table 7.

Variant	CT	MR	PET	US	Micro	Mean	Δ_{rare}
Capped volume RFS, $t = 0.001$	81.92	49.76	35.20	**79.09**	31.71	55.54	+2.61
Capped volume RFS, $t = 0.01$	**82.74**	**53.51**	**36.77**	76.55	**40.64**	**58.04**	+11.51
Oversampled volume RFS, $t = 0.01$	82.60	52.61	35.18	74.76	39.97	57.02	+12.40

Table 15. Training-set negative-slice ratios by modality. Values are computed from raw detector-training slices before PET negative-slice capping. Calibration is triggered when the raw negative-slice ratio exceeds 20%.

Modality	Raw slices	Pos./neg. slices	Negative ratio	Triggered?
CT	3,516,524	3,164,003 / 352,521	10.0%	No
MR	546,053	504,882 / 41,171	7.5%	No
Microscopy	29,478	28,086 / 1,392	4.7%	No
PET	115,223	59,017 / 56,206	**48.8%**	**Yes**
Ultrasound	98,080	98,001 / 79	0.08%	No

Table 16. Neg-FP rate by detector training regime (lower is better). Per-modality sample sizes: CT $n = 300$, MR $n = 294$, PET $n = 300$, Microscopy $n = 58$, aggregate $n = 952$. US is excluded because its negative-slice ratio is below 0.1% (no meaningful negative-slice subset exists).

Detector training regime	CT	MR	PET	Micro	Aggregate
Positive-only sampling	0.287	0.946	0.217	0.000	0.420
33% positive forcing	0.010	0.109	0.000	0.000	0.043
Baseline (K-cap, natural ratio)	0.013	0.048	0.000	0.000	0.019
+ Ratio-triggered negative-slice cap	**0.007**	**0.044**	**0.000**	**0.000**	**0.013**

Table 17. Per-organ RAOS-Real NSD results. Values are NSD percentages for displayed organs.

Organ	n	Medical SAM3	MedSAM3	Text3DSAM	Medal-S	BiomedParse-V	Dr. SAM3-T (detector–tracker)	Dr. SAM3-D (detector-only)
Liver	126	8.7	20.9	19.9	56.3	66.5	84.9	**88.1**
Left kidney	123	8.4	5.7	12.4	33.7	77.9	**90.9**	85.9
Right kidney	115	7.2	12.5	11.5	40.0	72.2	**88.0**	82.5
Spleen	126	9.3	13.7	19.5	30.1	75.3	**91.1**	82.6
Bladder	126	0.7	1.8	0.5	18.6	53.4	83.3	**90.0**
Stomach	126	3.7	16.0	9.3	28.1	50.9	75.9	**78.7**
Pancreas	126	3.4	10.2	9.2	37.0	55.1	79.1	**82.7**
Esophagus	124	0.4	1.7	21.0	14.0	74.9	83.8	**86.2**
Intestine	126	2.7	14.6	6.8	34.3	0.9	50.2	**68.6**
Colon	126	10.3	5.2	3.4	26.0	13.3	46.0	**53.7**
Gallbladder	86	0.2	6.5	3.2	7.6	30.7	**64.5**	53.1
Duodenum	126	0.7	3.5	7.3	24.0	33.0	36.3	**57.8**
Class avg.	**12 organs**	4.6	9.4	10.3	29.1	50.3	72.8	**75.8**

Table 18. Spatial-prompted (oracle bounding-box) baselines on the same MedSegFM 3D validation set. Values are Dice/F1 %.

	CT				**MR**				**US**		**PET**		**Micro**	
Model	DSC	NSD	F1	DSC-TP	DSC	NSD	F1	DSC-TP	DSC	NSD	F1	DSC-TP	F1	DSC-TP
MedSAM [19]	75.96	75.97	81.65	80.50	64.39	78.01	37.32	51.01	47.96	51.08	31.88	73.64	21.64	37.37
MedSAM2 [22]	61.18	64.28	77.02	78.59	51.65	63.22	39.35	57.43	88.79	87.45	60.40	80.17	46.77	75.64

Table 19. Our text-prompted multi-modality model vs. nnU-Net per-modality specialists. Values are Dice/F1 %. nnU-Net is trained as five separate models, one per modality, on the official 10% coreset; Dr. SAM3 is a single text-prompted system trained jointly across modalities.

	CT				**MR**				**US**		**PET**		**Micro**	
Model	F1	DSC-TP	DSC	NSD	F1	DSC-TP	DSC	NSD	DSC	NSD	F1	DSC-TP	F1	DSC-TP
nnU-Net [11]	14.86	19.70	30.07	29.73	12.73	29.61	57.85	64.69	90.11	90.15	**38.85**	**75.58**	22.86	73.08
Dr. SAM3-T (Ours)	**51.59**	**55.03**	65.99	67.43	**57.38**	71.77	61.24	69.12	89.22	89.83	25.26	63.32	26.07	70.02
Dr. SAM3-D (Ours)	45.48	54.27	**77.80**	**74.42**	56.62	**76.03**	**66.29**	**75.23**	**90.12**	**91.21**	30.99	74.26	**28.84**	**75.53**

Table 20. Checklist Table. Please fill out this checklist table in the answer column. (Delete this table in the camera-ready submission.)

Requirements	Answer
A meaningful title	Yes
The number of authors ($\leq$6)	6
Author affiliations and ORCID	Yes
Corresponding author email is presented	Yes
Task and track are clearly stated in the abstract	Yes
Validation metrics are presented in the abstract	Yes
Introduction includes background, related work, and motivation	Yes
A pipeline/network figure is provided	Figures 1 and 2
Pre-processing	Section 4.2
Strategies for data augmentation	Section 4.2
Strategy for class imbalance	Sections 3.1, 3.2
Strategies to improve model inference	Section 4.3
Post-processing	Section 4.3
Environment setting table is provided	Table 1
Training protocol table is provided	Table 2
Ablation study	Section 4.7
Efficiency evaluation results are provided	Tables 3 and 4
Qualitative example is provided	Appendix H
Limitation and future work are presented	Yes
Reference format is consistent	Yes
Main text $\geq$ 8 pages, not including references and appendix	Yes