Comparing SAM 2 and SAM 3 for Zero-Shot Segmentation of 3D Medical Data

Satrajit Chakrabarty[1] SATRAJIT.CHAKRABARTY@GEHEALTHCARE.COM
Ravi Soni[1] RAVI.SONI@GEHEALTHCARE.COM
[1] *GE HealthCare, San Ramon, CA, USA*

Editors: Under Review for MIDL 2026

Abstract

Foundation models, such as the Segment Anything Model (SAM), have heightened interest in promptable zero-shot segmentation. Although these models perform strongly on natural images, their behavior on medical data remains insufficiently characterized. While SAM 2 has been widely adopted for annotation in 3D medical workflows, the recently released SAM 3 introduces a new architecture that may change how spatial prompts are interpreted and propagated. Therefore, to assess whether SAM 3 can serve as an out-of-the-box replacement for SAM 2 for zero-shot segmentation of 3D medical data, we present the first controlled comparison of both models under purely spatial prompting, with concept mechanisms of SAM 3 disabled. We benchmark using a variety of prompting strategies on 16 public datasets (CT, MRI, Ultrasound, endoscopy) covering 54 anatomical structures, pathologies, and surgical instruments. We further quantify three failure modes: prompt-frame over-segmentation, over-propagation after object disappearance, and temporal retention of well-initialized predictions. Our results show that SAM 3 provides stronger initialization than SAM 2 for click prompts and maintains higher Dice and more stable retention for complex, vascular, and soft-tissue anatomies. Under bounding box and mask, SAM 2 remains competitive and often more conservative for compact organs by terminating tracks earlier and hallucinating less. The overall results position SAM 3 as the superior default choice for most medical segmentation tasks, while clarifying when SAM 2 remains a preferable propagator.

Keywords: Foundation models, Segment Anything Model, Zero-shot segmentation, SAM 2, SAM 3.

1. Introduction

Foundation models for promptable segmentation have reshaped interactive medical image analysis. The Segment Anything Model (SAM) (Kirillov et al., 2023) introduced a general-purpose framework for zero-shot segmentation of 2D images using point, box, and mask prompts. SAM 2 (Ravi et al., 2024) extended this approach to videos and 3D-like sequences with a memory-based transformer for frame-to-frame propagation, enabling consistent segmentation across volumes and cine series. The most recent iteration, SAM 3 (Carion et al., 2025), replaces the separate image and memory encoders with a unified Perception Encoder and a DETR-style detector–tracker (Carion et al., 2020), and adds concept-level prompting modules for open-vocabulary segmentation. These architectural changes are designed for recognition and semantics, but they also modify the model's behaviour under purely spatial prompts compared with SAM 2.

Although the SAM family performs strongly on natural images, its behaviour on medical imaging, especially 3D data and long temporal sequences, remains inadequately characterized. Medical data impose modality-specific contrast, low signal-to-noise ratios, and substantial variation across depth or time. Over the past year, SAM 2 has become a widely used baseline for zero-shot segmentation of 3D medical data because its memory module provides stable slice-to-slice propagation. Prior studies have compared SAM and SAM 2 on medical datasets under various prompting modes (Sengupta et al., 2025; Dong et al., 2024; Ma et al., 2024a), effectively serving as backwards-compatibility checks for newer models in the SAM family. With the introduction of SAM 3's detector–tracker pipeline and redesigned mask heads, a similar question arises: can SAM 3 safely replace SAM 2 in 3D medical annotation workflows operating purely with visual prompts?

To answer this question, we conduct a large-scale, controlled comparison of SAM 2 and SAM 3 across sixteen publicly available datasets spanning CT, MRI, ultrasound, and endoscopy, covering 54 anatomical structures, pathologies, and surgical instruments. We disable all concept-based mechanisms in SAM 3 so that both models operate strictly in the visual prompting regime. We benchmark single-click, multi-click, bounding-box, and mask prompts applied only to the first frame. Beyond prompt-frame and full-volume performance, we explicitly quantify three failure modes that are critical for interactive annotation workflows: prompt-frame over-segmentation (poor initialization), temporal retention (forgetting), and over-propagation after object disappearance.

This study makes three main contributions:

- A unified, cross-modality evaluation framework for comparing SAM 2 and SAM 3 under identical visual prompts, with all concept-level mechanisms in SAM 3 disabled.
- A comprehensive empirical analysis of prompt-frame and full-volume/sequence performance across sixteen datasets, revealing a structural divergence in behaviour: SAM 3 dominates initialization and tracking of complex topologies, while SAM 2 retains specific advantages for some compact, rigid anatomy under strong spatial guidance.
- A cross-model failure-mode analysis that quantifies prompt-frame over-segmentation, temporal decay of prediction, and over-propagation, providing the first systematic evidence on when SAM 3 can serve as an out-of-the-box replacement for SAM 2 and when SAM 2 remains the more conservative propagator.

By isolating visual-prompt behaviour and conducting extensive cross-modality experiments, this work clarifies the complementary strengths of SAM 2 and SAM 3 and provides practical guidance for selecting between these models in clinical and research settings.

2. Methods

2.1. Comparison rationale

The objective of this study is to compare SAM 2 and SAM 3 under controlled and identical prompting conditions for medical image segmentation in 3D volumes and medical video sequences. SAM 3 contains components designed for Promptable Concept Segmentation, such as the presence head and text–exemplar fusion modules. All concept-based mechanisms are

disabled in our evaluation so that both models operate strictly within the visual prompting regime. This configuration creates a direct comparison between the two architectures and isolates the effects of visual prompt encoding and propagation without the influence of text, concept embeddings, or exemplar-based recognition.

2.2. Model Overview

SAM 2 (Ravi et al., 2024) is an encoder–decoder architecture built on the Hiera (Hierarchical Vision Transformer) backbone (Ryali et al., 2023). Its defining feature is a streaming memory mechanism designed for semi-supervised video object segmentation: a memory bank stores features and masks from past frames, and a memory-attention module aggregates these to enforce spatio-temporal consistency during propagation through a 3D volume or cine sequence. SAM 3 (Carion et al., 2025) instead uses a unified Perception Encoder and a DETR-style detector–tracker (Carion et al., 2020). Learnable object queries are used to localize and track targets over time, and additional heads support presence prediction and concept-level prompting. In this work, we disable all language and concept modules so that SAM 3 operates purely as a visual tracker and segmenter. The resulting configuration exposes the impact of its new backbone and tracking pipeline while keeping the comparison with SAM 2 focused on visual prompts and propagation behaviour.

2.3. Prompting Strategy

We evaluate three standard prompting strategies: (i) *click prompting*, using either a single positive click (1,0) or a mixed positive–negative configuration (1,2), where the positive click is placed near the centroid of the target and negative clicks are sampled from a dilated region around the structure; (ii) *bounding-box prompting*, where a tight axis-aligned box around the ground-truth structure in the first frame provides coarse geometric context; and (iii) *mask prompting*, where a binary ground-truth mask from the first frame in which the structure appears is supplied as the initial prompt. All prompts are provided only on the first frame; thereafter, the models receive no additional interactions and propagate their predictions sequentially from the first to the last frame without forward–backward refinement, temporal smoothing, or post-processing.

2.4. Datasets and Implementation Details

We evaluate SAM 2 and SAM 3 on sixteen publicly available medical imaging datasets spanning four imaging modalities: 3D CT, 3D MRI, ultrasound (2D cine and 3D volumes), and endoscopy video (Table 1, Figure 1). Our data selection covers a broad spectrum of anatomical structures, pathological conditions, and clinical instruments across modalities, ensuring that the evaluation reflects the diversity encountered in real-world clinical imaging workflows. The CT cohorts include multi-organ abdominal benchmarks (AMOS (Ji et al., 2022), BTCV (Landman et al., 2015), FLARE22 (Ma et al., 2024b), TotalSegmentator (Wasserthal et al., 2023)) together with oncologic and thoracic tasks from the MSD collection (lung tumors, pancreas and pancreatic tumors, spleen, and colon cancer) (Antonelli et al., 2022; Simpson et al., 2019). MRI coverage comes from AMOS22 (Ji et al., 2022), ACDC (Bernard et al., 2018), MSD Task02 Heart and Task04 Hippocampus (Antonelli et al., 2022; Simpson et al., 2019), and the MRI subset of TotalSegmentator (D'Antonoli

Table 1: Descriptions of the medical imaging datasets used for evaluation.

Modality	Dataset	Anatomy	#Volumes / #Frames	#Classes
CT	AMOS	Abdominal	200 / 26069	9
	BTCV	Abdominal	30 / 3779	13
	FLARE22	Abdominal	50 / 4794	13
	MSD Lung	Lung tumor	63 / 17657	1
	MSD Pancreas	Pancreas, tumor	281 / 26719	2
	MSD Spleen	Spleen	41 / 3650	1
	MSD Colon	Colon cancer	126 / 13486	1
	TotalSegmentator	Abdominal	1113 / 304346	13
MRI	ACDC	Cardiac	149 / 1482	3
	AMOS	Abdominal	40 / 9455	9
	MSD Heart	Cardiac	20 / 2271	2
	MSD Hippocampus	Hippocampus	260 / 9270	2
	TotalSegmentator	Abdominal	1880 / 287217	13
US	CAMUS	Cardiac	500 / 9964	3
	SegThy	Thyroid/vascular	32 / 15820	5
Endoscopy	CholecSeg8K	Cholecystectomy	17 / 8080	12

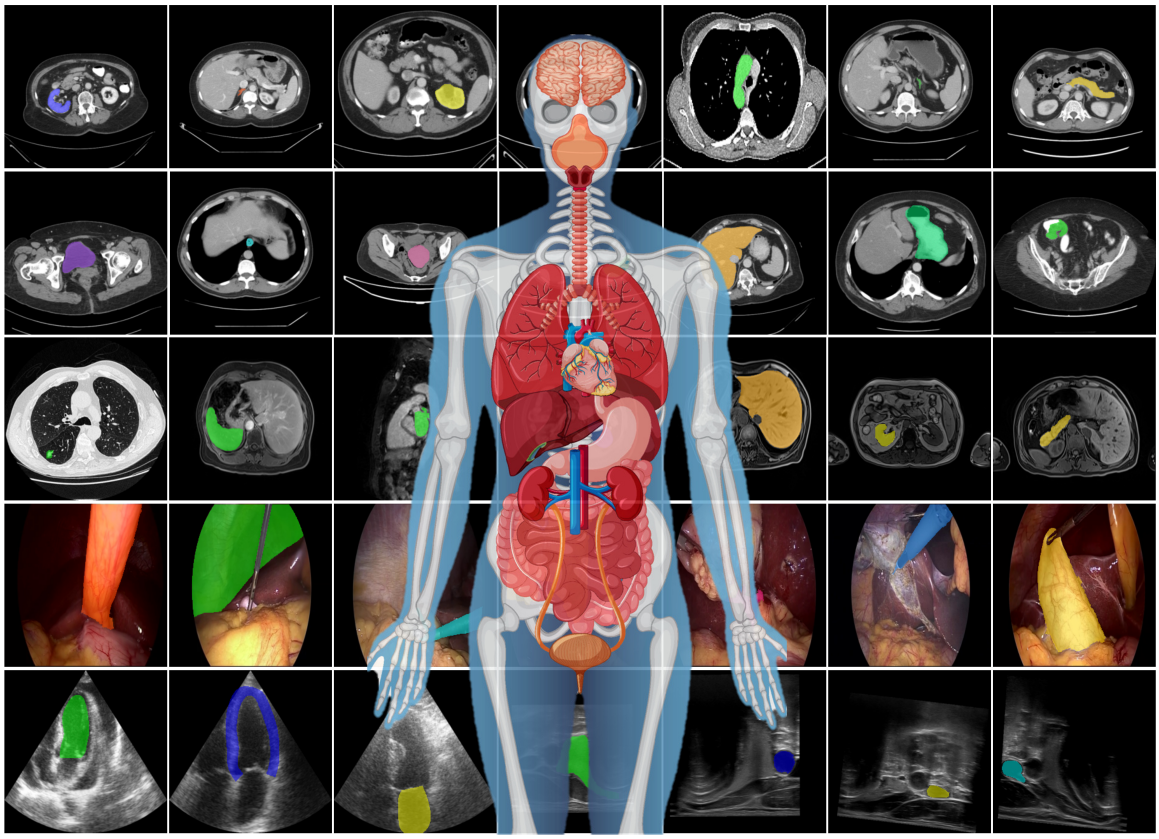

Figure 1: **Overview of the multi-modality benchmark dataset.** Representative images with ground-truth masks from the sixteen public datasets used in this study, illustrating variability in modality, anatomy, pathology, contrast, and acquisition. [Human illustration adapted from Vecteezy.com].

et al., 2024). Ultrasound is represented by cardiac cine sequences from CAMUS (Leclerc et al., 2019) and 3D thyroid ultrasound from SegThy (Krönke et al., 2022), while CholecSeg8K (Hong et al., 2020; Twinanda et al., 2016) provides endoscopy video frames with organ and instrument labels. Segmentation accuracy is measured using the Dice similarity coefficient (DSC). Statistical significance is assessed using paired Wilcoxon signed-rank tests on video/volume-level DSC, with significance defined at $\alpha = 0.05$. All preprocessing and checkpoint details are provided in Appendix A.

2.5. Failure Mode Analysis

To complement volume-level DSC, we quantify three failure modes at a *case* level, where *case* is a single target structure within a volume per dataset. All metrics are computed per case and summarized as distributions across all cases, stratified by prompting mode.

1. Prompt-frame oversegmentation (flooding). To measure whether the model accurately resolves the target's spatial extent or "floods" into the background, we compute an *area ratio* on the prompt frame t_0. Let $M_{gt}^{(t_0)}$ and $M_{pred}^{(t_0)}$ denote the ground-truth and predicted binary masks at t_0, and $|\cdot|$ the foreground pixel count. For all cases with $|M_{gt}^{(t_0)}| > 0$ we define

$$R = \frac{|M_{pred}^{(t_0)}|}{|M_{gt}^{(t_0)}|}, \tag{1}$$

and analyze both the distribution of R and the fraction of *severe flooding* events ($R > 2$).

2. Temporal retention (forgetting). To measure how segmentation quality evolves across a volume/sequence while the object is present, we model the decay of Dice over the object's lifespan. For each case, we consider all frames where the ground-truth mask is non-empty and DSC is defined, re-index the frame IDs to a normalized time variable $\tau \in [0, 1]$, and fit a simple linear model

$$DSC(\tau) \approx \alpha + \beta\,\tau. \tag{2}$$

The *normalized decay slope* β serves as a retention score: values closer to zero indicate stable performance, whereas more negative β correspond to faster forgetting. We compute β for all cases as well as focus on a subset of cases with good initialization (prompt-frame $DSC \geq 0.7$).

3. Over-propagation after object disappearance. To quantify how long a model continues to hallucinate a mask after the physical object has disappeared, we count the number of *over-propagated frames*. Let t_{last} be the final frame where the ground-truth object is present (non-empty mask). The over-propagation length for a case is then

$$L = \#\{t > t_{\text{last}} \mid M_{pred}^{(t)} \text{ is non-empty}\},$$

i.e., the number of frames with any predicted foreground after the object's last annotated frame. We summarize L via boxen plots and empirical cumulative distribution functions, and report percentiles such as the 90th percentile (P_{90}), which indicates the over-propagation length below which 90% of cases fall.

3. Results

3.1. Prompt-Frame Accuracy

To isolate the effect of prompt interpretation, defined here as the model's ability to accurately resolve the spatial extent of the target structure on the initial frame based on user input, we measured segmentation performance on the prompt-frame only. While detailed numerical results for all 54 anatomical structures are provided in Appendix B (Table 3), Figure 2 summarizes two key aspects: (a) the distribution of the prediction-to-ground-truth area ratio R for each prompt type on a log scale, and (b) the fraction of cases with severe over-segmentation ($R > 2$) stratified by object ground-truth size.

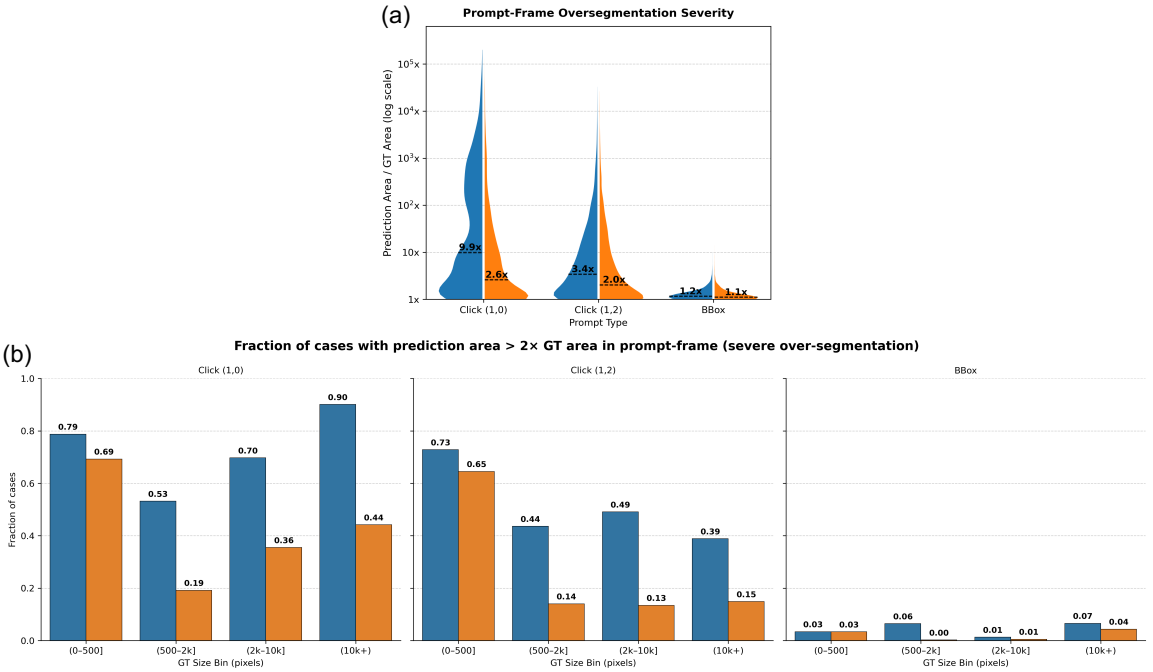

Figure 2: **Analysis of Prompt-Frame Over-segmentation.** **(a)** Distribution of the log_{10} Area Ratio (A_{pred}/A_{gt}). **(b)** Proportion of severe oversegmentation failures ($A_{pred} > 2 \times A_{gt}$), stratified by ground truth object size. [Color: SAM 2, SAM 3]

Across all structures and prompt types, SAM 3 provides markedly stronger and more stable initialization than SAM 2. Under click prompting, SAM 2 exhibits a severe instability manifested as a heavy-tailed distribution in Figure 2a: its predicted masks are frequently $10\times$ to $10^5\times$ larger than the ground truth. Specifically, the median area ratio for single-click prompts is $9.9\times$ for SAM 2 versus $2.6\times$ for SAM 3; for multi-click prompts, the medians drop to $3.4\times$ and $2.0\times$, respectively, and for bounding boxes they are close to unity at $1.16\times$ (SAM 2) and $1.11\times$ (SAM 3). Thus, even under sparse clicks, SAM 3 keeps the predicted area much closer to the ground-truth support, whereas SAM 2 frequently floods large portions of the frame.

Figure 2b quantifies the frequency of the severe over-segmentations ($R > 2$). For single-click prompting on small targets (< 500 pixels), 79% of SAM 2 initializations are severely over-segmented, compared with 69% for SAM 3; for medium-sized structures (500–2k pixels), the gap widens to 53% vs. 19%, and even for the largest objects ($\geq$ 10k pixels), severe over-segmentation occurs in 90% of SAM 2 cases but only 44% of SAM 3 cases. Multi-click prompting reduces these failure rates for both models, but SAM 2 still shows substantially higher severe-error frequencies (e.g., 73% vs. 65% in the smallest bin and 49% vs. 39% in the 2k–10k bin). Bounding-box prompts largely suppress over-segmentation in both models, with severe over-segmentation falling below 6% across all size bins and below 1% for most large structures. In this strong-prompt regime, the initialization advantage of SAM 3 becomes modest, confirming that SAM 2's instability is primarily a sparse-prompt phenomenon.

3.2. Full-Volume/Sequence Segmentation Accuracy

While prompt-frame accuracy captures initialization quality, clinical applications require accurate segmentation across full 3D volumes or complete temporal sequences. Full-volume DSC, therefore, reflects the combined effect of both initialization and propagation under SAM 2's memory-based architecture and SAM 3's redesigned tracking pathway. Table 2 summarizes structure-wise performance across all prompting regimes.

Across modalities, a consistent pattern emerges. Under sparse guidance (single- and multi-click prompting), SAM 3 generally achieves higher full-volume DSC than SAM 2 for most targets, indicating that its stronger prompt-frame initialization translates into better sequence-level performance, especially for anatomically complex or elongated structures such as vessels, gastrointestinal segments, and cardiac chambers. As prompt strength increases to bounding boxes and masks, this global advantage narrows: both models approach similar accuracy for many large, well-contrasted organs, and performance instead splits by anatomical type.

In this stronger-prompt regime, SAM 2 is frequently more competitive or superior for compact, encapsulated organs (e.g., kidneys, spleen, bladder), reflecting more conservative propagation once a reliable mask is provided. In contrast, SAM 3 retains an advantage for low-contrast, highly deformable, or tubular anatomy, where tracking stability is more challenging. Representative failure cases, such as MR bladder and SegThy thyroid/vascular targets, illustrate that excellent prompt-frame DSC can still collapse to near-zero full-volume DSC for one model while the other maintains stable masks. These discrepancies foreshadow the retention and over-propagation behaviour quantified by the failure-mode analysis.

3.3. Failure-Mode Analysis: Temporal Retention and Over-Propagation

Volume-level DSC aggregates initialization and propagation into a single number, but interactive workflows care about how masks evolve over time. Here we examine two temporal failure modes defined in Section 2.5: (i) *retention*, i.e., how quickly a well-initialized mask drifts or degrades while the object is still present, and (ii) *over-propagation*, i.e., how long a model continues to hallucinate a mask after the object has disappeared.

Retention of well-initialized objects. To isolate propagation behaviour from pure initialization failures, we restrict this analysis to cases with good starting masks (prompt-

Table 2: Full-volume/sequence DSC (%) for zero-shot segmentation across modalities and anatomical structures using single-click (1,0), multi-click (1,2), bounding-box, and mask prompts. For each pair, the higher DSC is shown in **bold**. Color shading in the table denotes statistical significance for the better model: $p < 0.001$, $0.001 < p < 0.05$, and no shading for $p > 0.05$.

Modality	Structure	Click (1,0)		Click (1,2)		BBox		Mask	
		SAM 2	SAM 3	SAM 2	SAM 3	SAM 2	SAM 3	SAM 2	SAM 3
CT	Adrenal Gland (L)	19.13	**25.14**	20.79	**34.54**	**49.02**	46.67	**47.59**	41.78
	Adrenal Gland (R)	8.93	**11.28**	10.18	**19.86**	**45.52**	44.86	**44.41**	39.53
	Aorta	68.71	**78.72**	72.07	**81.17**	68.41	**74.58**	67.22	**73.42**
	Bladder	3.63	**10.32**	6.56	**10.93**	10.60	**12.03**	9.72	**11.61**
	Colon Tumor	11.16	**15.98**	13.03	**17.27**	16.58	**18.24**	18.53	**19.36**
	Duodenum	25.68	**30.68**	26.92	**32.36**	31.34	**33.23**	32.78	**34.35**
	Esophagus	3.88	**37.00**	8.12	**48.28**	60.60	**68.44**	59.75	**68.22**
	Gallbladder	22.88	**30.39**	31.68	**34.53**	**49.62**	38.00	**48.47**	36.68
	Inferior Vena Cava	70.08	**79.28**	65.21	**78.77**	69.89	**78.54**	70.41	**78.47**
	Kidney (L)	58.72	**59.35**	**66.18**	61.38	**75.75**	64.70	**76.67**	64.43
	Kidney (R)	54.02	**65.66**	66.01	**67.43**	**78.15**	72.52	**78.78**	72.32
	Liver	44.18	**65.85**	52.10	**71.53**	67.72	**74.94**	67.21	**74.99**
	Lung Tumor	6.78	**18.72**	13.95	**30.06**	**44.22**	42.48	**46.33**	43.20
	Pancreas	19.24	**32.71**	23.93	**34.87**	28.00	**33.93**	27.38	**33.73**
	Pancreas Tumor	11.64	**17.00**	13.32	**18.72**	27.09	**28.47**	26.49	**29.06**
	Portal & Splenic Veins	27.70	**31.44**	31.00	**31.63**	**36.55**	34.05	**34.69**	34.33
	Prostate	2.56	**7.24**	5.32	**7.58**	**11.88**	8.70	**9.26**	8.75
	Spleen	46.20	**57.13**	56.51	**59.77**	**74.25**	63.03	**74.96**	62.56
	Stomach	36.80	**50.34**	45.73	**52.07**	49.43	**53.84**	49.42	**55.86**
MR	Aorta	42.58	**58.67**	46.05	**63.01**	40.83	**58.59**	42.12	**58.30**
	Bladder	0.48	**7.49**	**55.49**	6.80	**76.91**	7.29	**47.90**	6.34
	Gallbladder	17.92	**19.03**	**26.15**	25.06	**44.60**	30.19	**43.74**	30.57
	Hippocampus (Ant)	12.19	**16.96**	11.16	**17.86**	22.16	**23.62**	24.86	**25.37**
	Hippocampus (Post)	12.94	**33.89**	14.00	**33.10**	16.91	**18.43**	**23.74**	23.46
	Kidney (L)	**45.19**	44.44	**51.67**	48.77	**58.07**	49.23	**56.81**	48.07
	Kidney (R)	52.73	**54.43**	**60.53**	55.93	**63.28**	58.29	**64.40**	57.42
	Left Atrium	17.72	**30.41**	26.22	**43.65**	22.47	**44.23**	18.94	**39.46**
	Left Ventricle	80.38	**93.17**	74.32	**92.54**	88.21	**93.62**	89.88	**94.21**
	Liver	36.37	**57.46**	42.94	**63.12**	51.25	**56.39**	48.65	**56.42**
	Myocardium	36.92	**78.24**	39.56	**74.10**	52.95	**72.88**	82.46	**84.75**
	Pancreas	6.49	**22.87**	11.26	**24.57**	18.24	**24.35**	17.72	**23.83**
	Prostate	12.22	**17.70**	15.90	**19.59**	**28.12**	23.61	**28.10**	23.65
	Right Ventricle	52.39	**77.29**	46.01	**82.76**	80.41	**85.60**	82.61	**86.37**
	Spleen	26.91	**49.94**	36.22	**56.99**	56.52	**58.78**	**59.15**	59.11
US	Carotid Artery (L)	**13.55**	10.99	**23.53**	23.46	5.98	**56.65**	5.65	**41.36**
	Carotid Artery (R)	1.14	**11.92**	17.83	**21.67**	17.00	**51.12**	22.86	**60.69**
	Jugular Vein (L)	7.76	**30.69**	19.62	**30.55**	5.45	**35.30**	5.57	**39.52**
	Jugular Vein (R)	2.33	**17.84**	7.93	**31.92**	10.71	**28.16**	21.49	**29.05**
	Left Atrium	19.49	**28.59**	30.34	**66.11**	79.08	**83.82**	90.34	**90.60**
	LV Endocardium	27.47	**67.48**	62.50	**72.98**	85.38	**85.79**	**91.93**	91.19
	LV Epicardium	24.15	**28.04**	25.84	**27.24**	41.54	**42.29**	**82.20**	78.17
	Thyroid	10.13	**32.65**	19.87	**53.90**	10.53	**28.10**	7.27	**27.58**
Endoscopy	Abdominal Wall	55.77	**67.42**	58.14	**79.41**	69.20	**82.56**	81.23	**87.81**
	Blood	5.11	**12.94**	7.91	**38.80**	25.77	**33.39**	10.14	**38.22**
	Connective Tissue	**70.45**	66.32	**65.73**	61.16	61.14	**72.24**	69.91	**74.55**
	Cystic Duct	**0.15**	0.14	**0.20**	0.14	**1.42**	0.20	**0.16**	0.16
	Fat	60.00	**71.43**	**61.71**	60.90	38.42	**40.78**	87.20	**88.14**
	Gallbladder	72.49	**79.18**	74.95	**75.38**	**82.81**	81.17	78.73	**83.80**
	GI Tract	37.83	**65.02**	31.49	**70.83**	69.57	**75.84**	**76.77**	73.30
	Grasper	74.59	**82.47**	75.64	**78.21**	77.65	**78.93**	76.35	**80.76**
	Hepatic Vein	19.49	**21.62**	20.05	**21.70**	20.71	**21.64**	20.76	**23.08**
	L-Hook Electrocautery	**66.84**	64.50	65.91	**68.58**	65.91	**69.64**	66.78	**69.98**
	Liver	57.40	**59.78**	60.89	**68.48**	**67.76**	67.12	88.49	**90.33**
	Liver Ligament	**98.69**	98.29	**98.65**	96.16	**98.76**	98.55	**98.72**	98.51

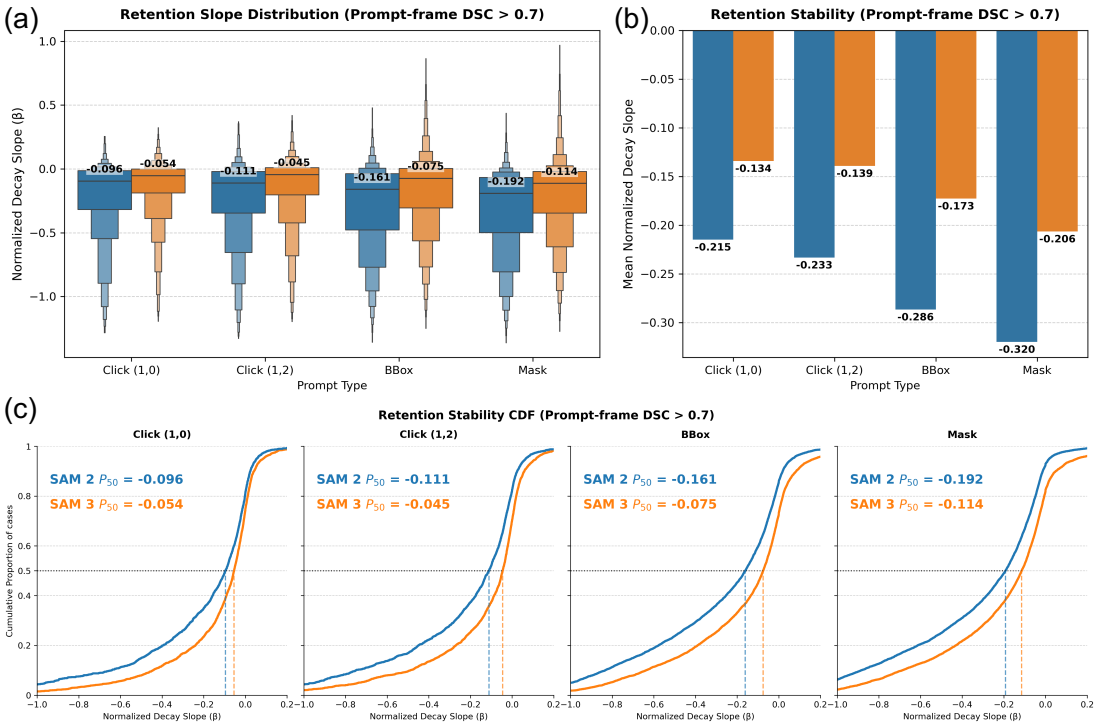

Figure 3: **Retention decay analysis for well-initialized cases.** Analysis is restricted to cases with $DSC \geq 0.7$ on the prompt frame. **(a)** Boxen plots showing the distribution of normalized decay slopes across prompting modes, where more negative values correspond to faster degradation in DSC as the object evolves over time. **(b)** Mean normalized decay slopes, summarizing the average retention behavior for each model and prompt type. **(c)** Cumulative distribution functions of the decay slopes, with annotated medians (P_{50}) highlighting that SAM 3 consistently exhibits less negative slopes than SAM 2. [Color: SAM 2, SAM 3]

frame $DSC \geq 0.7$), so that the decay slopes primarily reflect how well each model maintains a reasonable segmentation rather than how quickly an already-bad mask collapses. Figure 3 summarizes retention for cases with good initialization (prompt-frame $DSC \geq 0.7$). The boxen plots show the distribution of normalized decay slopes β for each prompt type, where more negative values correspond to faster loss of accuracy from the first to the last frame. The bar plot reports mean slopes by prompt type, and the ECDF curves show, for any threshold on β, what fraction of cases have decay no worse than that value; the annotated P_{50} markers indicate the median slope (half of the cases decay faster, half more slowly).

Across all well-initialized cases, both models exhibit negative normalized decay slopes on average, indicating that segmentation quality tends to deteriorate as the object evolves (Figure 3). However, SAM 3 consistently forgets more slowly. Under single-click prompts, the mean decay slope is -0.215 for SAM 2 versus -0.134 for SAM 3, and the median slopes (P_{50}) are -0.096 and -0.054, respectively, implying that a typical SAM 2 sequence

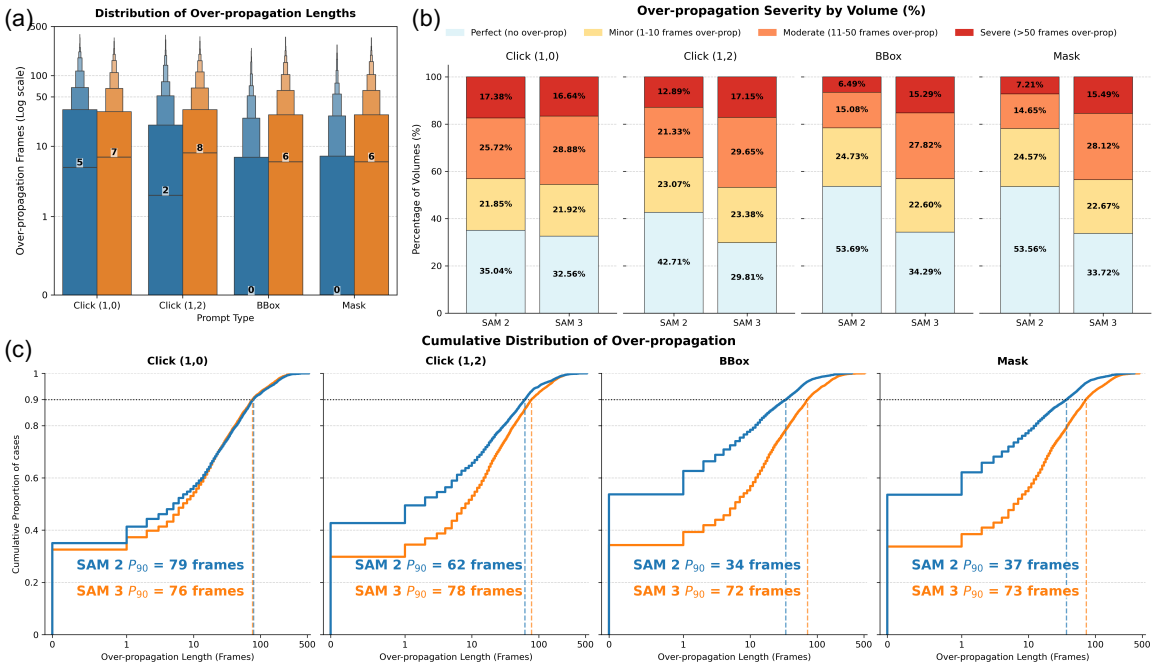

Figure 4: **Analysis of over-propagation after object disappearance.** Over-propagation length is defined as the number of frames with non-empty prediction after the last frame where the ground-truth object is present. **(a)** Boxen plots (log-scaled y-axis) showing the distribution of over-propagation lengths for each model and prompting mode. **(b)** Stacked bar plots summarizing the severity distribution across four bins: perfect termination, minor, moderate, and severe. **(c)** Cumulative distribution functions of over-propagation length, with annotated 90th-percentile values (P_{90}). [Color: SAM 2, SAM 3]

loses roughly twice as much DSC over its lifespan as a typical SAM 3 sequence. Multi-click prompts show a similar pattern (mean slopes -0.233 vs. -0.139; medians -0.111 vs. -0.045). The difference widens with stronger prompts: for bounding boxes, mean slopes are -0.286 (SAM 2) and -0.173 (SAM 3), with medians of -0.161 and -0.075; for masks, the means are -0.320 vs. -0.206 and medians -0.192 vs. -0.114. In the ECDFs, SAM 3's curves are consistently shifted toward less negative values, indicating that, conditional on a good start, SAM 3 maintains segmentation quality better across all prompt types.

Over-propagation after object disappearance. Figure 4 reveals the cost of SAM 3's retention stability: a tendency to be "sticky". The distributions of over-propagation length highlight how many frames each model continues to predict foreground after the last ground-truth frame, the stacked bars group volumes into none/minor/moderate/severe hallucination (0, 1–10, 11–50, >50 frames), and the ECDF curves describe the cumulative distribution of hallucinated length; the annotated P_{90} gives the number of frames below which 90% of cases fall.

Under single-click (1,0) prompts, both models behave similarly: about 35% of SAM 2 volumes and 33% of SAM 3 volumes terminate perfectly with zero over-propagation, and the P_{90} values are comparable (79 vs. 76 frames). With multi-click prompts, SAM 2 becomes slightly more conservative, with about 43% of volumes showing no over-propagation compared with about 30% for SAM 3, and P_{90} dropping to 62 frames for SAM 2 versus 78 frames for SAM 3.

The contrast is sharper once strong spatial guidance is provided. For bounding-box prompts, 54% of SAM 2 volumes exhibit no over-propagation, compared with only 34% for SAM 3, and severe tails of more than 50 hallucinated frames occur in 6.5% of SAM 2 cases but 15.3% of SAM 3 cases; the corresponding P_{90} values are 34 vs. 72 frames. Mask prompting shows a similar trend: roughly 54% of SAM 2 volumes versus 34% of SAM 3 volumes have zero over-propagation, while severe tails appear in 7.2% vs. 15.5% of cases and P_{90} increases from 37 frames (SAM 2) to 73 frames (SAM 3).

Taken together with the prompt-frame over-segmentation analysis, these failure-mode results highlight a complementary trade-off between the models. SAM 3 offers more reliable initialization and better retention for well-initialized objects, particularly under stronger prompts, but is more "sticky" and prone to long-lived hallucinated masks after the object disappears. SAM 2 is less capable under sparse prompts and struggles with complex anatomy, yet it tends to terminate tracks earlier and exhibits fewer extreme over-propagation failures under bounding-box and mask prompting.

3.4. Performance Behavior as a Function of Prompt Strength

Across modalities, both models follow a consistent pattern as prompt strength increases from single-click to multi-click, bounding-box, and mask prompts. Under sparse guidance (clicks), SAM 3 dominates because it interprets minimal prompts more reliably, leading to higher prompt-frame DSC, fewer prompt-frame over-segmentation failures, and better temporal retention. As prompts become more informative and provide explicit spatial support, the global advantage narrows and performance instead splits by anatomical type: SAM 2 becomes competitive or superior for compact, well-delineated organs, while SAM 3 retains a clear advantage for elongated, low-contrast, or vascular structures that are more susceptible to drift and over-propagation during tracking.

3.5. Overall Interpretation and Summary of Findings

Taken together, the prompt-frame, full-volume, and failure-mode evaluations show that SAM 2 and SAM 3 offer complementary strengths rather than a single performance hierarchy, driven by a trade-off between prompt interpretation (what to segment) and temporal consistency (how well it is remembered). We summarize our findings as follows:

- **Initialization advantages for SAM 3.** Under click prompts, SAM 3 has a clear advantage: its unified perception encoder infers structure from minimal input, yielding higher prompt-frame DSC and substantially fewer flooding failures than SAM 2 across most targets.
- **Propagation trade-off for compact versus complex anatomy.** Once initialized via bounding-box or mask prompts, propagation behaviour splits by anatomy. SAM 2

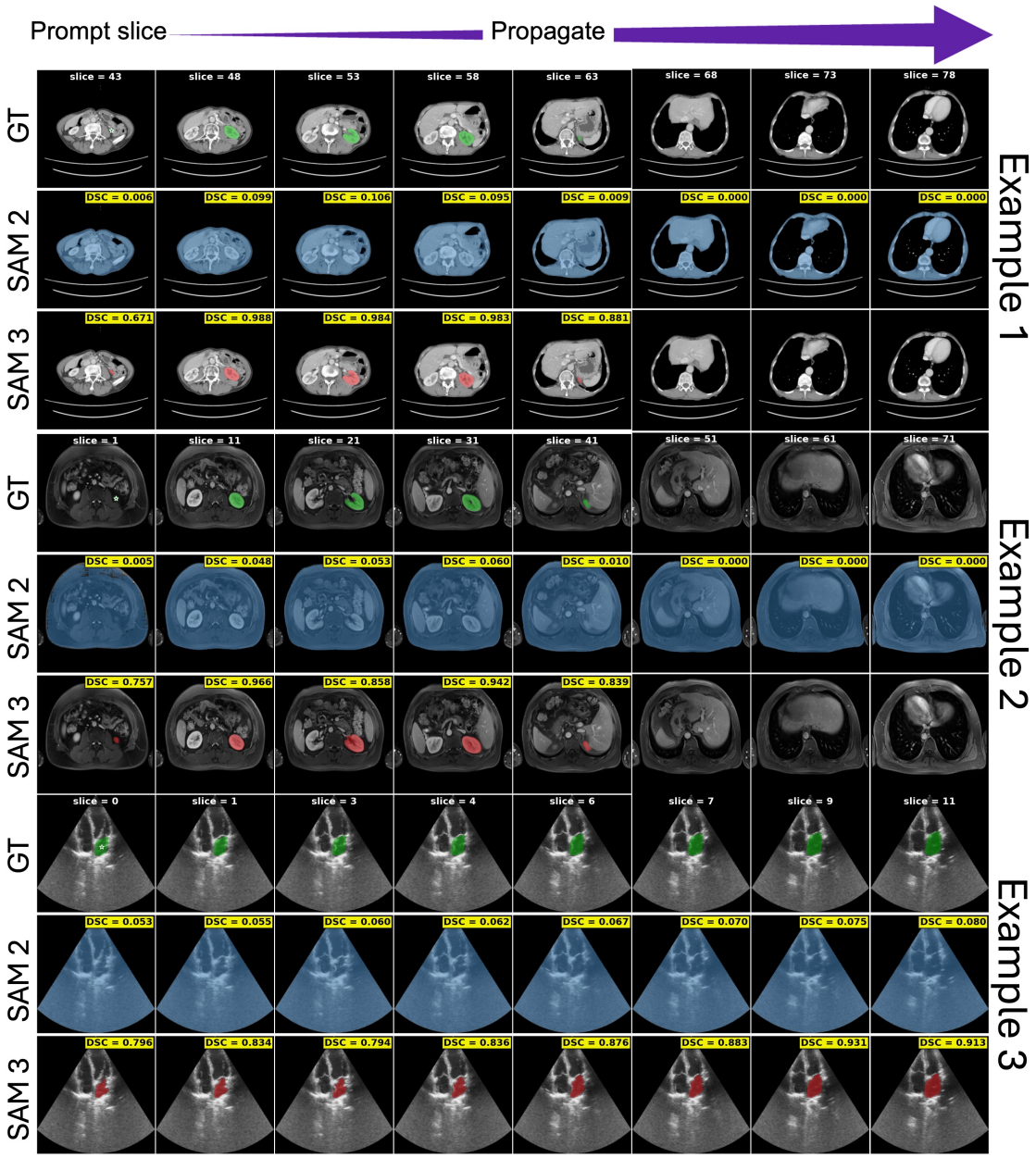

Figure 5: **Qualitative examples of cases where SAM 3 outperforms SAM 2**. All three examples are for single-click (1,0) prompting and show SAM 3's superior prompt initialization by better localizing the structure even under sparse prompts. SAM 3 produces accurate, spatially coherent segmentations even for small or low-contrast structures, whereas SAM 2 exhibits failure to localize the target on the prompted frame, resulting in over-segmentation and notably lower DSC. [Colors: GT, SAM 2, SAM 3]

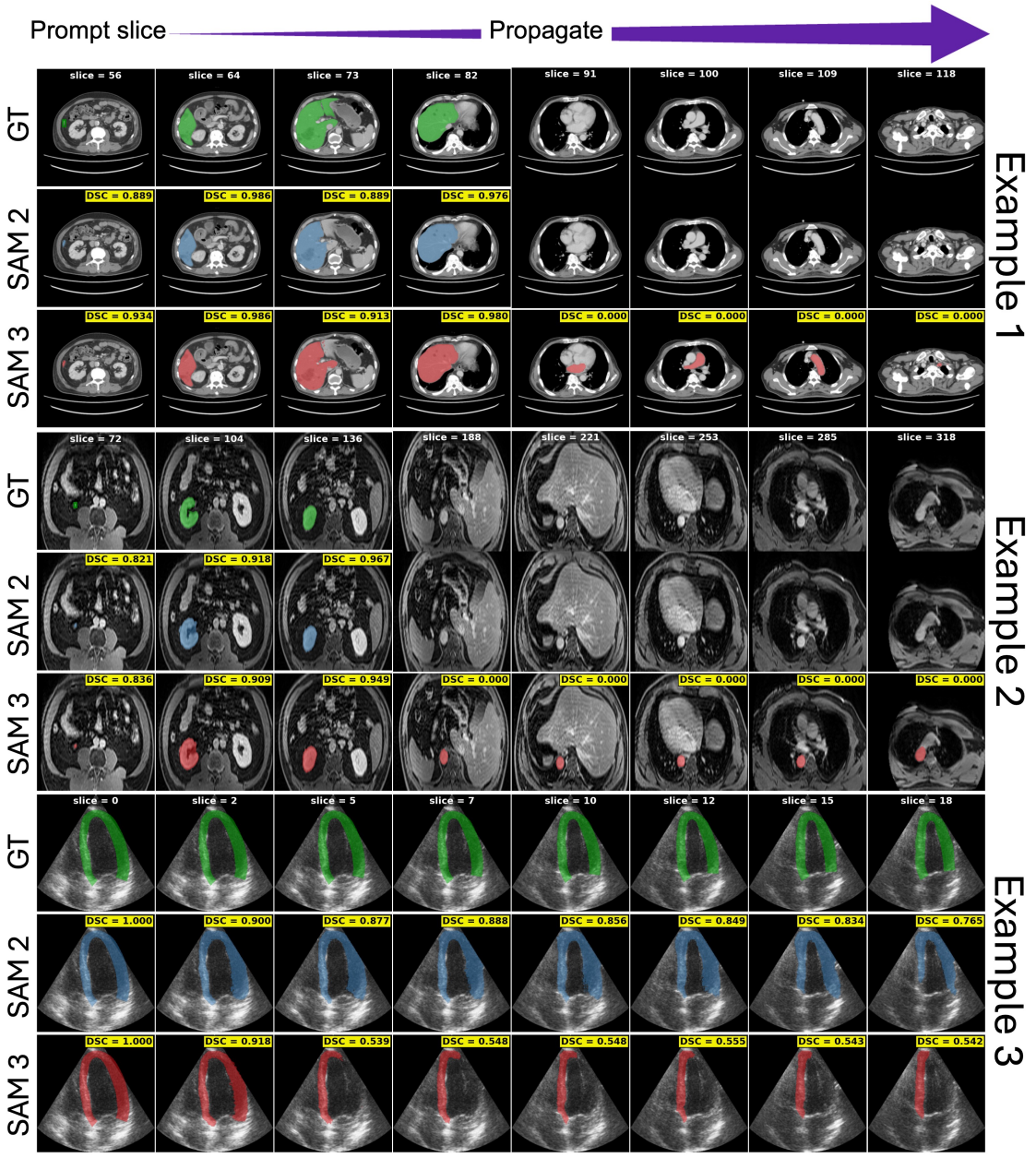

Figure 6: **Qualitative examples of cases where SAM 2 outperforms SAM 3**. Examples 1–2 are for bbox and example 3 is for mask prompt. In these examples, SAM 3 provides strong initial localization but exhibits propagation failures, including hallucinated residual masks in later slices (Examples 1–2) and erosion or collapse of structure boundaries under low contrast or motion (Example 3). In contrast, SAM 2 maintains more stable slice-to-slice consistency and suppresses spurious predictions, yielding higher DSC. [Colors: GT, SAM 2, SAM 3]

is more stable for compact, encapsulated organs (kidney, spleen, bladder), often showing less temporal drift and fewer over-propagation tails under strong spatial guidance. SAM 3 is more reliable for elongated, continuous, or poorly contrasted structures (e.g., vessels, gastrointestinal tract, thyroid), where SAM 2 often fails to maintain the segmentation despite good initialization.

- **The "unreliable propagator" risk.** High initialization accuracy does not guarantee successful propagation. In several datasets (e.g., MR bladder, SegThy ultrasound), one model attains excellent prompt-frame DSC but then collapses or hallucinates for many frames. This highlights the need to evaluate temporal retention and over-propagation beyond prompt-frame or volume-averaged DSC.

Overall, for interactive tasks with sparse clicks, SAM 3 is the natural default due to its stronger prompt interpretation and retention. When stronger prompts (bounding boxes or masks) are available, the preferred model becomes anatomy dependent: SAM 2 is often safer for compact organs, whereas SAM 3 is better suited to vascular and irregular soft tissues where propagation is more fragile. In practice, the optimal choice depends on the balance between initialization difficulty and propagation demands in a given modality and task.

4. Conclusion

This work presents the first large-scale, controlled comparison of SAM 2 and SAM 3 for zero-shot segmentation of 3D medical data under identical visual prompting. By evaluating single-click, multi-click, bounding-box, and mask initialization across sixteen datasets and 54 anatomical structures, we disentangle how architectural changes in SAM 3 affect prompt interpretation, temporal retention, and failure behaviour relative to SAM 2.

Our results show that SAM 3 offers markedly stronger prompt interpretation: it delivers higher prompt-frame DSC, substantially fewer flooding failures, and slower temporal decay for well-initialized objects, especially under click and bounding-box prompts. These advantages translate into superior full-volume performance for most complex, elongated, or low-contrast targets. SAM 2, however, remains a competitive and often preferable choice for compact, rigid organs under strong spatial guidance, where its propagation is more conservative and less prone to long-lived hallucinated masks. The failure-mode analysis highlights that high initialization accuracy alone is not sufficient: models can still suffer catastrophic collapse or prolonged over-propagation, underscoring the need to explicitly evaluate temporal retention and termination behaviour.

Overall, our findings position SAM 3 as the stronger default backbone for broad 3D medical segmentation workflows, while clarifying scenarios in which SAM 2 remains the safer propagator for specific organ types and prompt regimes. A key limitation of this study is that we restrict the comparison to purely visual prompts, deliberately disabling the concept- and text-based mechanisms introduced in SAM 3. As vision–language approaches such as Voxtell (Rokuss et al., 2025) gain traction for open-vocabulary 3D medical segmentation, extending our framework to include semantic prompting and language-guided concepts, with SAM 3's full capabilities enabled, is an important direction for future work.

References

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11): 2514–2525, 2018.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

Tugba Akinci D'Antonoli, Lucas K Berger, Ashraya K Indrakanti, Nathan Vishwanathan, Jakob Weiß, Matthias Jung, Zeynep Berkarda, Alexander Rau, Marco Reisert, Thomas Küstner, et al. Totalsegmentator mri: Sequence-independent segmentation of 59 anatomical structures in mr images. *arXiv preprint arXiv:2405.19492*, 2024.

Haoyu Dong, Hanxue Gu, Yaqian Chen, Jichen Yang, Yuwen Chen, and Maciej A Mazurowski. Segment anything model 2: an application to 2d and 3d medical images. *arXiv preprint arXiv:2408.00756*, 2024.

W-Y Hong, C-L Kao, Y-H Kuo, J-R Wang, W-L Chang, and C-S Shih. Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. *arXiv preprint arXiv:2012.12453*, 2020.

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

Markus Krönke, Christine Eilers, Desislava Dimova, Melanie Köhler, Gabriel Buschner, Lilit Schweiger, Lemonia Konstantinidou, Marcus Makowski, James Nagarajah, Nassir Navab, et al. Tracked 3d ultrasound and deep neural network-based thyroid segmentation reduce interobserver variability in thyroid volumetry. *Plos one*, 17(7):e0268550, 2022.

Bennett Landman, Zhoubing Xu, Juan Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge*, volume 5, page 12. Munich, Germany, 2015.

Sarah Leclerc et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019.

Jun Ma, Sumin Kim, Feifei Li, Mohammed Baharoon, Reza Asakereh, Hongwei Lyu, and Bo Wang. Segment anything in medical images and videos: Benchmark and deployment. *arXiv preprint arXiv:2408.03322*, 2024a.

Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Shihao Mae, Adamo Young, Cheng Zhu, Xin Yang, Kangkang Meng, Ziyan Huang, et al. Unleashing the strengths of unlabelled data in deep learning-assisted pan-cancer abdominal organ quantification: the flare22 challenge. *The Lancet Digital Health*, 6(11):e815–e826, 2024b.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Maximilian Rokuss, Moritz Langenberg, Yannick Kirchhoff, Fabian Isensee, Benjamin Hamm, Constantin Ulrich, Sebastian Regnery, Lukas Bauer, Efthimios Katsigiannopulos, Tobias Norajitra, et al. Voxtell: Free-text promptable universal 3d medical image segmentation. *arXiv preprint arXiv:2511.11450*, 2025.

Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International conference on machine learning*, pages 29441–29454. PMLR, 2023.

Sourya Sengupta, Satrajit Chakrabarty, and Ravi Soni. Is sam 2 better than sam in medical image segmentation? In *Medical Imaging 2025: Image Processing*, volume 13406, pages 666–672. SPIE, 2025.

Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5), 2023.

Appendix A. Preprocessing Details

All datasets were converted into a unified slice-based format to enable consistent evaluation across models and modalities. For CT datasets, images were first windowed using clinically standard ranges (e.g., soft-tissue windowing with level–width of 40/400 for abdominal CT and lung windowing of −600/1500 for thoracic CT) before clipping and rescaling to $[0, 255]$. MRI volumes were normalized by extracting the intensity values between the 0.5th and 99.5th percentiles within each volume and linearly rescaling this clipped range to $[0, 255]$, ensuring robustness to modality-specific dynamic range differences. Ultrasound images were min–max normalized per sequence and similarly rescaled to $[0, 255]$. Endoscopy datasets provided color-coded semantic masks, which were converted into per-class binary masks via RGB-to-class lookup. No smoothing, interpolation, or artifact removal was applied. This standardized preprocessing ensures that SAM 2 and SAM 3 operate on identically prepared inputs across all imaging modalities.

All experiments use publicly released checkpoints without any fine-tuning. For SAM 2, we use the SAM 2.1 Hiera-B+ checkpoint. The SAM 3 release does not specify multiple model variants in the paper, and we therefore adopt the standard configuration provided by the authors. All evaluations were performed on NVIDIA H100 GPUs.

Appendix B. Detailed Prompt-Frame Results

This appendix provides the structure-wise breakdown of prompt-frame accuracy to supplement the analysis in Section 3.1. Table 3 summarizes the DSC across all structures, modalities, and prompt types. In CT, the gains for SAM 3 are substantial for anatomically small or low-contrast targets such as bladder, pancreas, esophagus, prostate, and spleen. Improvements are similarly pronounced in MRI, especially for cardiac structures where SAM 3 significantly outperforms SAM 2 for LV, RV, and myocardium under both single- and multi-click prompting. Multi-click prompting (1,2) reduces ambiguity for both models, yet SAM 3 retains a clear advantage in nearly all CT and MRI structures with statistically significant gains, frequently with $p < 0.001$.

For bounding-box prompts, where the spatial support is considerably less ambiguous, the performance gap narrows but does not disappear. SAM 3 continues to produce higher DSC for most CT and MRI structures, although with smaller margins. We also see some instances of SAM 2 performing slightly better than SAM 3 (e.g., MR Bladder and MR Hippocampus Posterior). Box prompts achieve the highest absolute accuracy for both models, and here the differences between the two models typically fall within a modest range (∼5 DSC points) with the exception of MR Myocardium where SAM 3 beats SAM 2 by about 20 DSC points.

Ultrasound exhibits a mixed pattern. For segmentation of cardiac chambers in cine sequences (LA, LV endocardium, LV epicardium), SAM 3 achieves substantially higher DSC for click prompts, reflecting improved localization. In contrast, for the SegThy dataset (thyroid, carotid arteries, and jugular veins), both models exhibit near-total failure under click prompting, with DSCs frequently remaining in the single digits. Segmentation accuracy becomes meaningful only when bounding-box prompts are supplied; in this viable regime, SAM 2 consistently outperforms SAM 3 across the thyroid and all vascular targets.

For endoscopy (CholecSeg8K), SAM 3 shows an advantage under single-click (1,0) prompting, outperforming SAM 2 for the majority of the tissue and instrument classes. However, under multi-click (1,2) and bounding-box prompts, the results are more balanced: SAM 2 and SAM 3 each achieve higher DSC for different categories, and no model dominates across all structures. Notably, even when numerical differences are large between the two models, none of these comparisons reach statistical significance because CholecSeg8K contains only a small number of annotated videos, which limits the power of paired significance testing.

Appendix C. Retention on all data

In the main text (Section 3.3) we focused on retention behavior for the subset of cases with good initialization ($DSC \geq 0.7$ on the prompt frame). For completeness, Figure 7 extends the same analysis to all volume–object pairs, including those with poor initialization. As expected, the distributions of normalized decay slopes broaden for both models, particularly under click prompting where failures at the first frame lead to rapid apparent decay. Under single-click (1,0) prompts, the mean slopes of SAM 2 and SAM 3 are nearly identical (approximately -0.071 vs. -0.073) and the median slopes are close to zero (-0.006 vs. -0.028), reflecting that most degradation is driven by a minority of highly unstable cases. For multi-click prompting, SAM 3 retains an advantage (mean slopes -0.130 vs. -0.085; medians -0.059 vs. -0.030). The differences become more pronounced under bounding-box and mask prompts: SAM 2 exhibits substantially more negative mean slopes (about -0.262 and -0.309) than SAM 3 (about -0.152 and -0.201), and the ECDFs show that SAM 3's decay distributions are consistently shifted toward less negative values. Overall, when poor initializations are included, SAM 3 continues to forget more slowly than SAM 2 once a sufficiently strong spatial prompt is provided.

Table 3: Prompt–frame DSC (%) for zero-shot segmentation across modalities and anatomical structures using single-click (1,0), multi-click (1,2), and bounding-box prompts. For each pair, the higher DSC is shown in **bold**. Color shading in the table denotes statistical significance for the better model: $p < 0.001$, $0.001 < p < 0.05$, and no shading for $p > 0.05$.

Modality	Structure	Click (1,0)		Click (1,2)		BBox	
		SAM 2	SAM 3	SAM 2	SAM 3	SAM 2	SAM 3
CT	Adrenal Gland (L)	20.28	**25.56**	25.81	**37.11**	77.04	**78.38**
	Adrenal Gland (R)	9.60	**10.65**	15.56	**20.14**	77.45	**79.91**
	Aorta	60.30	**68.14**	65.54	**69.59**	84.05	**86.20**
	Bladder	10.33	**50.53**	22.13	**59.97**	84.49	**87.51**
	Colon Tumor	25.45	**44.05**	35.21	**52.08**	74.76	**76.82**
	Duodenum	47.65	**52.93**	52.80	**61.04**	82.07	**85.77**
	Esophagus	5.20	**33.93**	20.68	**46.13**	85.97	**86.85**
	Gallbladder	24.38	**44.24**	37.32	**52.06**	82.68	**84.38**
	Inferior Vena Cava	86.25	**90.79**	84.81	**91.24**	91.25	**93.63**
	Kidney (L)	51.79	**65.32**	60.48	**67.76**	84.48	**87.78**
	Kidney (R)	49.94	**65.13**	61.59	**67.24**	84.67	**88.17**
	Liver	30.62	**52.88**	44.08	**57.99**	78.42	**81.99**
	Lung Tumor	5.08	**19.21**	20.12	**29.90**	75.89	**76.62**
	Pancreas	17.96	**36.65**	28.60	**41.79**	78.72	**81.22**
	Pancreas Tumor	22.01	**37.43**	28.10	**42.81**	85.55	**88.24**
	Portal & Splenic Veins	60.85	**76.99**	66.13	**76.82**	86.91	**88.50**
	Prostate	9.06	**36.79**	25.03	**47.22**	86.01	**87.25**
	Spleen	37.91	**60.77**	54.95	**65.60**	80.81	**85.09**
	Stomach	45.89	**62.82**	56.87	**69.60**	84.40	**87.21**
MR	Aorta	29.88	**38.59**	37.97	**40.79**	83.66	**84.97**
	Bladder	13.42	**86.99**	84.28	**87.97**	**92.42**	90.53
	Gallbladder	16.25	**26.63**	28.62	**33.63**	82.00	**83.32**
	Hippocampus (Ant)	6.89	**18.70**	**23.69**	20.90	82.12	**82.31**
	Hippocampus (Post)	3.16	**8.92**	6.14	**13.37**	**82.82**	79.41
	Kidney (L)	41.18	**44.71**	48.50	**49.75**	81.88	**82.86**
	Kidney (R)	40.84	**49.34**	49.83	**52.22**	82.26	**84.88**
	Left Atrium	4.58	**9.97**	15.16	**18.30**	75.25	**79.67**
	Left Ventricle	88.08	**95.64**	89.10	**94.34**	96.06	**96.75**
	Liver	19.82	**34.04**	29.60	**39.75**	76.87	**78.63**
	Myocardium	36.61	**79.95**	44.74	**72.40**	53.80	**74.25**
	Pancreas	4.20	**17.38**	16.13	**23.58**	75.27	**78.24**
	Prostate	13.82	**29.82**	26.80	**35.58**	84.08	**84.39**
	Right Ventricle	73.64	**86.55**	74.90	**89.23**	95.07	**95.61**
	Spleen	20.78	**39.74**	38.06	**49.25**	77.78	**79.50**
US	Carotid Artery (L)	0.38	**0.68**	**3.62**	1.61	**64.05**	57.98
	Carotid Artery (R)	0.18	**1.46**	**14.53**	3.30	**61.06**	60.97
	Jugular Vein (L)	1.14	**2.65**	**6.74**	2.93	**58.41**	55.61
	Jugular Vein (R)	0.40	**1.63**	**6.99**	2.50	**58.32**	54.97
	Left Atrium	17.61	**25.31**	47.37	**60.30**	76.99	**82.61**
	LV Endocardium	31.73	**70.46**	69.49	**73.01**	**86.77**	86.16
	LV Epicardium	23.98	**27.65**	**30.73**	26.04	34.15	**34.93**
	Thyroid	0.99	**3.89**	**5.88**	3.70	**67.19**	65.14
Endoscopy	Abdominal Wall	58.07	**72.06**	77.34	**80.86**	**87.33**	87.24
	Blood	3.06	**3.08**	42.03	**63.01**	**73.81**	72.91
	Connective Tissue	**66.34**	54.60	**68.74**	67.61	85.01	**88.95**
	Cystic Duct	30.50	**32.92**	**29.70**	13.45	41.92	**48.84**
	Fat	62.33	**74.23**	**72.84**	62.25	**53.74**	52.80
	Gallbladder	73.32	**83.38**	82.24	**83.48**	88.38	**88.44**
	GI Tract	53.26	**79.72**	80.53	**86.75**	92.02	**93.13**
	Grasper	81.52	**90.12**	**90.30**	88.98	**87.00**	86.53
	Hepatic Vein	**87.78**	85.87	**88.03**	86.91	**88.74**	87.64
	L-Hook Electrocautery	**73.67**	72.20	**74.06**	71.93	89.42	**90.96**
	Liver	55.80	**66.08**	63.87	**67.55**	72.76	**73.40**
	Liver Ligament	**98.38**	98.25	**98.29**	95.66	**98.55**	98.52

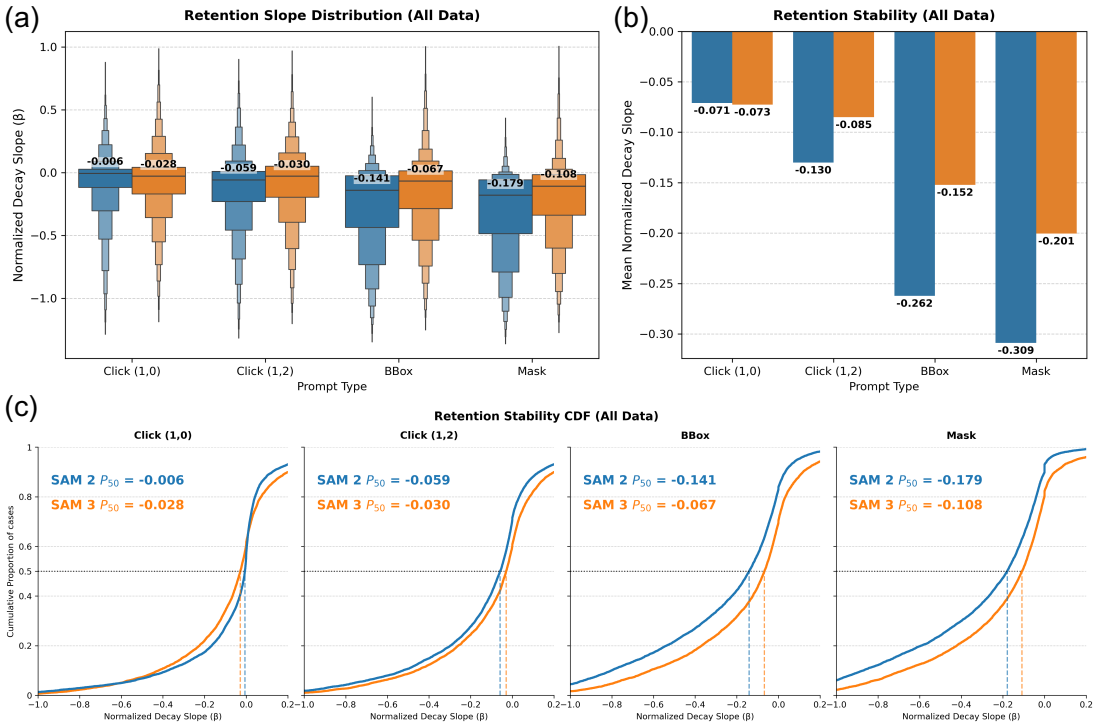

Figure 7: **Retention decay analysis on all cases.** Same layout as Figure 3, but computed over all volume–object pairs, including those with poor initialization ($DSC < 0.7$). The distributions broaden for both models, especially under click prompting, yet SAM 3 generally maintains less negative mean decay slopes for multi-click, bounding-box, and mask prompts.