UNCOVERING MODALITY DISCREPANCY: RETHINKING THE VALIDATION OF FOUNDATION MODELS FOR GENERAL-PURPOSE 3D MEDICAL SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation models have emerged as a transformative paradigm in 3D medical imaging, with the promise of unified quantitative analysis across diverse targets and imaging modalities. However, these models are predominantly developed and evaluated on datasets largely concentrated around a limited set of imaging modalities and anatomical regions. To provide an evaluation of out-of-the-box robustness of these foundation models, we curate a large-scale dataset comprising 490 whole-body PET/CT and 464 whole-body PET/MRI scans (approximately 675k 2D images and 12k 3D annotations) and conduct a thorough and comprehensive evaluation of representative 3D segmentation foundation models. Our analysis reveals a substantial gap between benchmark-reported performance and real-world generalization, with marked degradation on previously unseen data distributions and particularly severe failures on functional imaging modalities. These findings suggest that current foundation models remain far from achieving true universality. We suggest a reconsideration of how universality is defined and validated, extending evaluation beyond regional structural benchmarks toward whole-body structural and functional imaging. Bridging this gap will be essential for translating foundation models from controlled evaluation settings to real-world medical practice.

1 INTRODUCTION

Medical imaging is a cornerstone of modern healthcare, playing an indispensable role in disease diagnosis, treatment planning, and patient monitoring Shen et al. (2017). Among the diverse tasks in medical image analysis, volumetric segmentation of 3D medical imaging occupies a central position as it underpins quantitative measurements of organ volumes, tumor burden, and treatment response that are essential for clinical decision-making Ilesanmi et al. (2024). Manual segmentation has long been the gold standard. However, this procedure is highly time-consuming, labor-intensive, and requires significant domain expertise. Semi- or fully-automatic segmentation methods can significantly reduce the time and manual effort required, improve consistency across results, and enable the rapid analysis of large-scale datasets.

Deep learning models have shown strong potential in medical image segmentation by extracting intricate features from massive annotated datasets and achieve accurate segmentation across a diverse range of tasks Antonelli et al. (2022); Ma et al. (2022); Lalande et al. (2022); Gatidis et al. (2024). Despite these advances, a prominent bottleneck of existing models remains their strict task-specificity. These models are typically designed and trained for specific predefined tasks and often suffer from severe performance degradation when exposed to new tasks or out-of-distribution data. Consequently, their capacity to generalize to other clinical scenarios is highly constrained, necessitating the development of distinct models for each specific imaging modality or target organ, which poses a major obstacle to broad clinical translation given that routine clinical workflows demand the robust segmentation of highly heterogeneous structures across a wide array of imaging modalities Ma & Wang (2023).

Recent advancement of artificial intelligence has been defined by the emergence of foundation models (FMs), which are large-scale, pre-trained models capable of performing multi-task segmentation

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

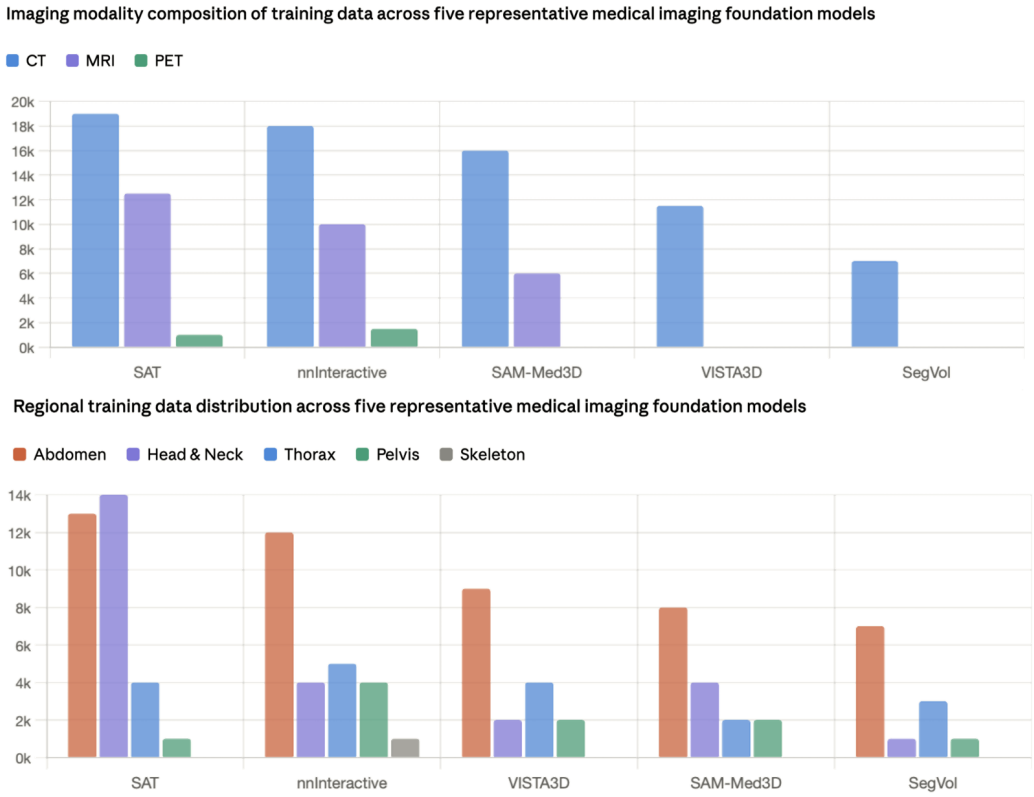

Figure 1: Modality and regional discrepancies in the training data of five representative medical imaging foundation models. (Top) The composition of imaging modalities reveals a structural bias. CT is the dominant modality across all models, supplemented by MRI, while functional imaging such as PET is severely underrepresented or entirely absent. (Bottom) The distribution of anatomical regions demonstrates a strong bias toward the abdomen (and the head and neck for SAT). Conversely, data representing the pelvis and skeleton are persistently scarce. Together, these distributions highlight systemic data imbalances that may limit the generalization of current models to comprehensive, whole-body, and multi-modal tasks.

within a unified framework Awais et al. (2025); Zhang & Metaxas (2024). The field of medical image analysis is undergoing a paradigm shift from task-specific models to general-purpose foundation models. Inspired by the success of the Segment Anything Model (SAM) in computer vision Kirillov et al. (2023), the community has witnessed a proliferation of efforts to adapt this paradigm to medical image segmentation for quantitative image analysis Mazurowski et al. (2023); Zhang et al. (2024). However, direct adaptation of natural-image-based models to the medical domain faces significant structural hurdles. A primary limitation is that many existing adaptations operate via 2D slice-wise segmentation Ma et al. (2024); Zhao et al. (2025a), when adapted to 3D medical images, the volumes are segmented slice-by-slice and reconstructed through post-processing. This strategy inherently overlooks the critical volumetric contextual information and inter-slice spatial correlations that are fundamental to radiological interpretation Zhang et al. (2022). To address the shortcomings of 2D models, a series of 3D general-purpose medical foundation models have emerged Du et al. (2024); He et al. (2025); Isensee et al. (2025); Wang et al. (2025); Zhao et al. (2025b) to leverage intrinsic volumetric features directly and achieve remarkable generalization capabilities across a wide array of 3D segmentation tasks.

While these 3D medical foundation models claim to provide a unified solution for diverse clinical tasks, their development and validation remain largely restricted. As shown in Figure 1, we quantify this severe imbalance by tallying the total 3D training volumes adopted during model construction across leading universal segmentation backbones. As visualized, CT and MRI scans constitute the

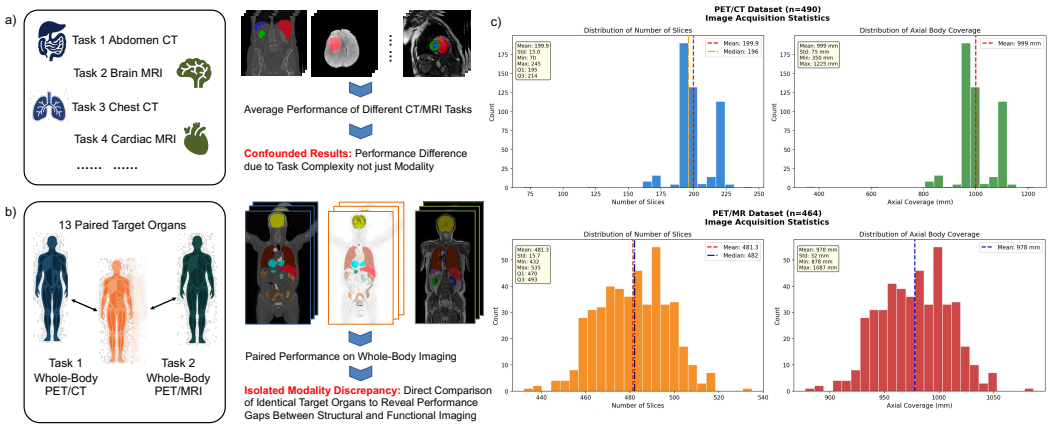

Figure 2: Overview and comparison of existing evaluation datasets and proposed UMD dataset. (a) Existing validation protocols typically assess models on heterogeneous datasets where modality is intrinsically entangled with specific anatomical tasks. This approach prevents an isolated measurement of modality-specific robustness, as performance variations are confounded by varying task complexities. (b) Our evaluation addresses this limitation via paired whole-body PET/CT and PET/MRI datasets with matched target organs. Direct segmentation of identical anatomical targets across co-registered structural-functional modality pairs isolates pure modality discrepancy, enabling rigorous, unconfounded quantification of cross-modality performance deficits. (c) Image acquisition statistics of PET/CT and PET/MRI datasets in proposed benchmark.

overwhelming majority of training corpora for every benchmarked model, while functional PET data occupies only a negligible fraction of all training samples. Beyond modality imbalance, the regional distribution of training data reveals an equally pronounced bias. This skewed data distribution motivates testing whether released models remain robust on underrepresented functional modalities and anatomical regions. Although such imbalance is a plausible contributor to cross-domain degradation, training-set composition alone does not establish the mechanism responsible for a particular model's failure.

Conventional cross-dataset evaluations can also obscure cross-modality robustness because different modalities are often paired with different cohorts, anatomical targets, and annotation protocols. Under this setup, a performance difference cannot be attributed to modality alone: task difficulty, organ morphology, acquisition, and imaging contrast vary simultaneously, as shown in Figure 2(a). This motivates a paired benchmark that controls patient anatomy and target definitions while preserving the model-specific inference process used in practice. Conventional cross-dataset evaluations can also obscure cross-modality robustness because different modalities are often paired with different cohorts, anatomical targets, and annotation protocols. Under this setup, a performance difference cannot be attributed to modality alone: task difficulty, organ morphology, acquisition, and imaging contrast vary simultaneously, as shown in Figure 2(a). This motivates a paired benchmark that controls patient anatomy and target definitions while preserving the model-specific inference process used in practice.

A significant limitation in current evaluations of general-purpose medical segmentation models is the reliance on heterogeneous datasets, where different modalities often correspond to disparate anatomical regions or clinical tasks. Here, we constructed a benchmark of 490 paired PET/CT and 464 paired PET/MRI for whole-body segmentation, enabling a direct comparison of model performance across co-registered modalities, as shown in Figure 2(b). This paired design, with multi-modal scans acquired simultaneously from the same patient cohort, ensures that segmentation results for identical organs can be used to objectively compare performance differences across modalities. By utilizing intra-subject controlled comparisons, we effectively isolate imaging modality as the primary independent variable, enabling a rigorous quantification of the inherent modality discrepancy. Furthermore, our benchmark serves as a pristine testbed to evaluate the true zero-shot generalization performance of these models on previously unseen data.

Our findings identify important robustness limitations in the evaluated systems and motivate broader, deployment-oriented validation. The promise of whole-body, modality-agnostic universality, we argue, remains substantially unaddressed by the evaluated releases under this protocol. The observed failures are consistent with limitations in current training and validation ecosystems, but our experiments do not isolate training-data bias from model- and implementation-specific factors. We conceptualize universality along two orthogonal axes. (1) **modality universality**: the capacity to process and reason across distinct 3D imaging modalities, spanning both structural and functional domains. (2) **regional universality**: the ability to perform competently across the full topographic extent of the human body, rather than within a restricted set of organ systems.

2 BENCHMARK DESIGN

To bridge the gap and uncover modality discrepancy, we present UMD, a segmentation benchmark dataset by retrospectively collecting 490 whole-body PET/CT and 464 whole-body PET/MRI scans, with approximately 675,000 2D slices. For each case, the structural (CT or MRI) and functional (PET) volumes were acquired from the same subject during a single diagnostic session and co-registered. This intra-subject paired nature guarantees that the underlying morphology, scale, and orientation of the target organs remain constant across modalities. We provide voxel-wise fine-grained annotations for 13 diverse organs, including the liver, left kidney, right kidney, brain, heart, spleen, aorta, lung, colon, urinary bladder, pancreas, esophagus, and stomach, which supports a comprehensive and systematic evaluation of general-purpose segmentation models across varied anatomical complexities and imaging modalities.

By providing paired, co-registered ground-truth labels for both structural and functional imaging volumes, our protocol enables a head-to-head comparison between different modalities, free from confounding factors such as inter-subject anatomical differences. Furthermore, as an originally collected dataset rather than a recompilation of existing public data sources, we eliminate the risk of data leakage, guaranteeing the validity and authenticity of the evaluation results. As a result, the evaluation performance can be reliably attributed to modality-specific generalization capabilities of these models.

To comprehensively evaluate the robustness of current 3D medical foundation models, we conducted a systematic comparison of five representative state-of-the-art general-purpose segmentation foundation models, including SAM-Med3D-turbo Wang et al. (2025), SegVol Du et al. (2024), SAT-Pro Zhao et al. (2025b), VISTA3D He et al. (2025), and nnInteractive Isensee et al. (2025). These models represent the cutting edge of general-purpose segmentation, utilizing diverse strategies such as large-scale pre-training on volumetric data, prompt-based interaction, and universal feature encoders. Notably, each of these frameworks has claimed to achieve superior zero-shot generalization performance across a wide array of 3D medical imaging tasks and anatomical structures.

To ensure a fair and rigorous evaluation, we utilized the official, latest, and largest-scale pre-trained weights, and the model-specific default inference workflow for all comparing foundation models. Following the standardized evaluation protocols associated with each model, we implemented two primary interaction strategies to facilitate zero-shot segmentation. For VISTA3D and SAT, we provided the specific textual category or class ID corresponding to each of the target organs. For other point-based models, we utilized simulated point prompts generated from the ground truth masks to mimic ideal user interaction. Our objective is a system-level, out-of-the-box evaluation of these foundation models as interactive segmentation tools in clinical applications.

3 RESULTS AND ANALYSIS

Table 1 summarizes the overall evaluation performance. Comparing literature-reported performance with performance on our benchmark reveals systematic overestimation across all models. Critically, even within CT alone, most models exhibit substantial performance gaps between their reported benchmark scores and their actual efficacy on previously unseen data from different acquisition protocols. SAM-Med3D-turbo, which reports a CT Dice score of 0.790, achieves only 0.404 on the benchmark CT data. SegVol drops from a reported 0.793 to 0.107, and VISTA3D collapses from 0.711 to 0.071, suggesting that current models may rely heavily on distribution-specific patterns rather than fully generalizable anatomical representations. The transition to functional PET imag-

Model	Setting	Reported Performance*			UMD Evaluation Performance			
		CT	MRI	PET	CT	PET	MRI	PET
SAM-Med3D-turbo	1 point prompt	0.790	0.754	/	0.404	0.410	0.433	0.397
SegVol	1 point prompt	0.793	/	/	0.107	0.075	0.134	0.148
nnInteractive	1 point prompt	∼0.55	∼0.55	/	0.481	0.342	0.511	0.217
VISTA3D	class id	0.711	/	/	0.071	0.005	0.003	0.006
SAT	text prompt	75.6	83.8	63.4	0.018	0.003	0.006	0.000
nnUNet	10 Training Samples	-	-	-	0.686	0.605	0.652	0.560

Table 1: Performance comparison of report average performance of different segmentation tasks from different modalities and paired evaluation performance on proposed benchmark. ∗ Literature values are reproduced only to describe the operating range reported by each source paper. They are not numerically comparable with UMD because target classes, cohorts, annotations, input sizes, prompts, and averaging procedures differ. / denotes that the test data do not contain corresponding modality. ∼ denotes that the specific performance for each individual modality is not reported, and only the average performance is presented.

ing further amplifies these failures. Semantic-guided models such as VISTA3D and SAT suffer near-complete failure on PET data, with Dice scores plummeting to near zero across nearly all organs, indicating that their learned representations are not grounded in anatomical understanding but rather in distribution-specific patterns. Point-based interactive models such as SAM-Med3D-turbo and nnInteractive maintain marginal continuity through spatial prompt localization, yet their performance remains far below clinical acceptability, particularly for organs with intricate geometries or low contrast, while even the best-performing model nnInteractive achieves only 0.342 on PET compared to 0.481 on CT.

Given the inherent differences in imaging characteristics across different modalities, we disentangle whether the aforementioned performance gap arises from the intrinsic segmentation difficulty of each modality itself, or from the distributional bias of these foundation models. We train nnU-Nets Isensee et al. (2021) of each modality as a reference of state-of-the-art task-specific models for comparison. As shown in Table 1, These results show that the structural-functional difficulty difference is smaller for this reference than for these foundation models. Most models exhibit a significant performance gap, indicating that current 3D foundation models are overfitted to their validation data distributions and lack the robustness required for diverse clinical scenarios. Specifically, semantic guided models like SAT and VISTA3D undergo a near total collapse upon evaluation on unseen data, revealing that their purported generalization is restricted to specific morphological patterns.This persistent performance gap across all frameworks confirms that structural proficiency does not inherently translate into functional understanding, highlighting a critical limitation in the current foundation model paradigm.

Comprehensive evaluation results separated by organ, model, and modality are provided in Table 2 and Table 3, which further confirms that the modality discrepancy is systemic rather than organ-specific. The few exceptions, where PET segmentation achieves non-trivial scores, are confined to organs with extreme radiotracer uptake such as the bladder, where the high signal intensity creates sufficient contrast for modality-agnostic spatial heuristics to succeed, with the signal sufficiently distinctive to be captured. These findings indicate that current foundation models remain strongly modality-dependent, having not yet developed the capacity to bridge the gap between anatomical density and metabolic activity. The benchmarks on which these models are evaluated and celebrated are themselves structurally biased, creating a self-reinforcing cycle where models optimized for a narrow data regime are validated on data drawn from that same regime.

Qualitative analysis further corroborates these quantitative findings and provides intuitive insight into the nature of model failure. Figure 3(a) illustrates the generalization illusion through SegVol. On its in-domain CT cases, the model achieves near-perfect segmentation of the liver (DSC: 0.97) and spleen (DSC: 0.73). However, when applied to previously unseen CT data acquired under different protocols, performance collapses catastrophically (liver DSC: 0.02, spleen DSC: 0.10). The segmentation masks are fragmented, spatially incoherent, and bear little resemblance to the target

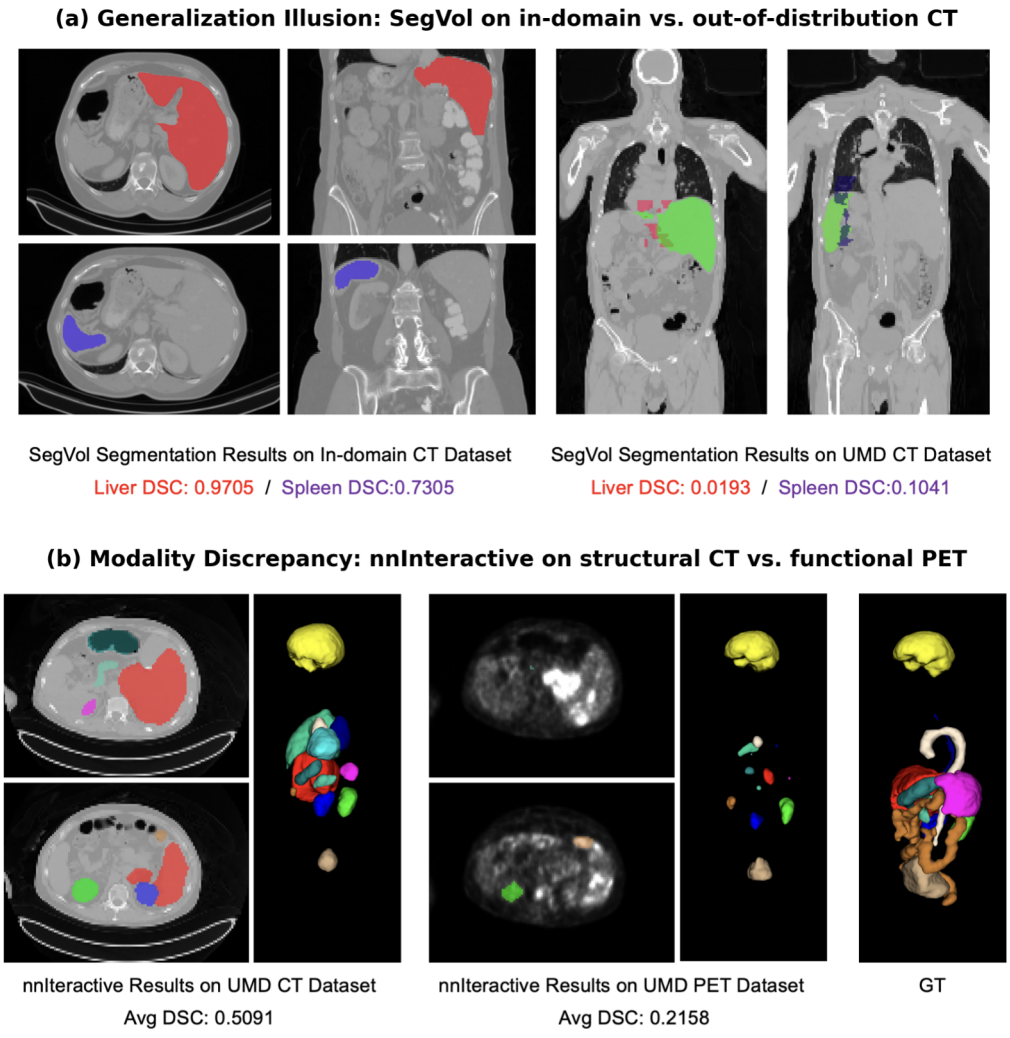

Figure 3: Qualitative visualization of failure modes in representative 3D segmentation foundation models. (a) Generalization illusion: SegVol achieves excellent performance on in-domain CT data (left) but undergoes complete segmentation collapse on out-of-distribution CT from the UMD benchmark (right), despite both being structural CT scans. (b) Modality discrepancy: nnInteractive produces reasonable multi-organ segmentation on UMD CT (left) but fails to maintain anatomical coherence on the co-registered PET scan from the same patient (middle), compared with ground truth (right).

anatomy, revealing that the model's in-domain competence derives from memorization of dataset-specific intensity patterns rather than genuine anatomical understanding. Figure 3(b) demonstrates the modality discrepancy through nnInteractive, the best-performing model in our evaluation. While achieving moderate multi-organ segmentation on structural CT (average DSC: 0.51), the same model produces severely degraded and anatomically implausible predictions on the co-registered PET scan of the same patient (average DSC: 0.22). This controlled comparison provides direct visual evidence that current foundation models are predominantly anchored to structural intensity profiles and face substantial difficulty transferring their learned representations to the metabolic signal distributions characteristic of functional imaging.

These empirical evidences underscore a critical methodological insight for the evaluation of foundation models: performance on aggregated, in-distribution benchmarks is a necessary but profoundly insufficient condition for claiming generality. Without controlled, out-of-distribution evaluation that

Model	SAM-Med3D-tb	SegVol	nnIteractive	VISTA3D	SAT
Liver	0.7039(0.1155)	0.0806(0.1982)	0.4993(0.3319)	0.9186(0.0378)	0.0057(0.0205)
	0.6835(0.1100)	0.1618(0.2539)	0.3904(0.3769)	0.0676(0.0475)	0.0020(0.0451)
Kidney-L	0.4701(0.1300)	0.1203(0.1967)	0.7930(0.1636)	0.0000(0.0000)	0.0001(0.0018)
	0.5678(0.1411)	0.0687(0.1329)	0.4464(0.3436)	0.0000(0.0000)	0.0041(0.0638)
Kidney-R	0.4387(0.1491)	0.1437(0.2279)	0.7780(0.1753)	0.0000(0.0001)	0.0061(0.0780)
	0.5596(0.1843)	0.0630(0.1178)	0.4187(0.3413)	0.0003(0.0020)	0.0082(0.0900)
Brain	0.6760(0.2378)	0.3024(0.3926)	0.7940(0.1432)	0.0000(0.0000)	0.0020(0.0451)
	0.6609(0.3030)	0.1399(0.2408)	0.8796(0.1250)	0.0000(0.0000)	0.0020(0.0451)
Heart	0.5073(0.0992)	0.0993(0.1863)	0.5607(0.2908)	0.0000(0.0000)	0.0020(0.0451)
	0.4581(0.1323)	0.1148(0.1994)	0.2214(0.2641)	0.0001(0.0013)	0.0041(0.0638)
Spleen	0.5110(0.1483)	0.1208(0.2065)	0.6064(0.2937)	0.0000(0.0000)	0.0061(0.0780)
	0.5827(0.1295)	0.1105(0.2000)	0.4378(0.3516)	0.0001(0.0014)	0.0061(0.0780)
Aorta	0.1343(0.0588)	0.0465(0.0900)	0.2356(0.1325)	0.0000(0.0000)	0.0024(0.0452)
	0.1635(0.0600)	0.0541(0.0894)	0.0662(0.0936)	0.0000(0.0000)	0.0020(0.0451)
Lung	0.6243(0.0767)	0.1426(0.2239)	0.4805(0.2361)	0.0002(0.0015)	0.0429(0.0151)
	0.5307(0.1522)	0.0236(0.0594)	0.3345(0.2679)	0.0002(0.0006)	0.0020(0.0451)
Colon	0.1316(0.0838)	0.0384(0.0769)	0.0701(0.0631)	0.0000(0.0000)	0.1465(0.0727)
	0.1298(0.0652)	0.0264(0.0519)	0.0685(0.0792)	0.0000(0.0000)	0.0000(0.0000)
Bladder	0.3960(0.1774)	0.0972(0.1704)	0.5234(0.2486)	0.0000(0.0000)	0.0021(0.0451)
	0.5900(0.2343)	0.1016(0.1950)	0.6676(0.2430)	0.0000(0.0000)	0.0020(0.0451)
Pancreas	0.1233(0.0813)	0.0743(0.1296)	0.4007(0.1399)	0.0000(0.0000)	0.0115(0.0252)
	0.1589(0.0767)	0.0299(0.0594)	0.0844(0.0894)	0.0002(0.0022)	0.0041(0.0638)
Esophagus	0.0904(0.0545)	0.0356(0.0690)	0.1721(0.0867)	0.0000(0.0000)	0.0057(0.0472)
	0.0620(0.0321)	0.0235(0.0416)	0.0749(0.0662)	0.0000(0.0000)	0.0020(0.0451)
Stomach	0.4459(0.1482)	0.0875(0.1680)	0.3536(0.2567)	0.0000(0.0000)	0.0022(0.0036)
	0.1945(0.1477)	0.0546(0.1127)	0.3563(0.3034)	0.0000(0.0000)	0.0041(0.0638)
Avg	0.4035(0.0417)	0.1071(0.1376)	0.4813(0.0737)	0.0706(0.0045)	0.0181(0.0207)
	0.4101(0.5016)	0.0747(0.0971)	0.3416(0.0999)	0.0053(0.0037)	0.0033(0.0440)

Table 2: Evaluation results (DSC) of representative general-purpose segmentation foundation models for zero-shot promptable PET and CT segmentation. Results in red and blue are the performance of PET and CT, respectively. Light gray and dark gray shaded cells indicate results that are significantly better with $p<0.01$ and $p<0.0001$.

systematically probes model behavior across the structural-functional divide, the field risks perpetuating what we term the generalization illusion: a state in which high reported metrics may create unwarranted confidence in model robustness while deployment on underrepresented modalities reveals significant and previously uncharacterized performance gaps. Moving beyond this illusion requires both the development of new evaluation paradigms and a fundamental reconceptualization of what constitutes a valid test of universality.

4 Universality Reconsidered

The performance differences revealed by our evaluation motivate a broader question about how universality is defined and validated in medical imaging foundation models. **Strong results across multiple datasets establish breadth only over the sampled evaluation conditions.** UMD adds two underrepresented dimensions, including functional modality and whole-body coverage, and shows that several current releases are not consistently robust in this setting.

Modality discrepancy: the under-representation of functional imaging. Current foundation models are overwhelmingly developed on structural imaging modalities, particularly CT and MRI, while functional imaging remains severely underrepresented. This imbalance reflects practical challenges in acquiring and annotating PET data, but it also creates a fundamental blind spot in model

Model	SAM-Med3D-tb	SegVol	nnInteractive	VISTA3D	SAT
Liver	0.7719(0.0755)	0.3986(0.2433)	0.8209(0.2412)	0.0384(0.0548)	0.0000(0.0005)
	0.6259(0.2050)	0.3731(0.2500)	0.1225(0.2252)	0.0749(0.0244)	0.0000(0.0000)
Kidney-L	0.4857(0.1422)	0.1090(0.1136)	0.6346(0.2242)	0.0000(0.0000)	0.0000(0.0000)
	0.5666(0.1142)	0.1047(0.0867)	0.1679(0.2398)	0.0000(0.0000)	0.0000(0.0000)
Kidney-R	0.5521(0.1488)	0.0917(0.1088)	0.7040(0.1546)	0.0000(0.0000)	0.0000(0.0000)
	0.5708(0.1257)	0.0884(0.0785)	0.2225(0.2494)	0.0057(0.0207)	0.0000(0.0000)
Brain	0.5797(0.1767)	0.2342(0.1919)	0.6405(0.3937)	0.0002(0.0010)	0.0378(0.0072)
	0.7707(0.0746)	0.2654(0.2306)	0.7860(0.3120)	0.0000(0.0000)	0.0000(0.0000)
Heart	0.6799(0.0893)	0.0987(0.1069)	0.5633(0.3423)	0.0000(0.0000)	0.0000(0.0000)
	0.6685(0.1012)	0.2382(0.2124)	0.1430(0.2047)	0.0000(0.0000)	0.0000(0.0000)
Spleen	0.5609(0.1130)	0.1564(0.1477)	0.6719(0.2582)	0.0000(0.0000)	0.0000(0.0000)
	0.4353(0.1606)	0.1655(0.1422)	0.1368(0.2115)	0.0000(0.0000)	0.0000(0.0000)
Aorta	0.0709(0.0458)	0.0866(0.0872)	0.2009(0.1384)	0.0000(0.0000)	0.0000(0.0000)
	0.0837(0.0437)	0.1186(0.0811)	0.0465(0.0662)	0.0000(0.0000)	0.0000(0.0000)
Lung	0.6046(0.0938)	0.2611(0.2298)	0.5700(0.1768)	0.0000(0.0000)	0.0005(0.0022)
	0.1892(0.1183)	0.0469(0.0718)	0.2685(0.2484)	0.0002(0.0007)	0.0000(0.0000)
Colon	0.1091(0.0586)	0.0478(0.0648)	0.1115(0.1110)	0.0000(0.0000)	0.0346(0.0556)
	0.1225(0.0661)	0.0416(0.0478)	0.0532(0.0632)	0.0000(0.0000)	0.0000(0.0000)
Bladder	0.6176(0.1912)	0.0452(0.0863)	0.5908(0.2813)	0.0000(0.0000)	0.0000(0.0000)
	0.6548(0.2074)	0.2613(0.2188)	0.5999(0.2067)	0.0000(0.0000)	0.0000(0.0000)
Pancreas	0.1642(0.0834)	0.0579(0.0625)	0.2961(0.1992)	0.0002(0.0025)	0.0043(0.0655)
	0.1702(0.0831)	0.0552(0.0559)	0.0554(0.0732)	0.0000(0.0000)	0.0043(0.0655)
Esophagus	0.0352(0.0209)	0.0837(0.0678)	0.1177(0.0936)	0.0000(0.0000)	0.0000(0.0000)
	0.0335(0.0169)	0.0435(0.0362)	0.0360(0.0448)	0.0000(0.0000)	0.0000(0.0000)
Stomach	0.4073(0.1306)	0.0701(0.0925)	0.7182(0.2040)	0.0000(0.0000)	0.0000(0.0000)
	0.2888(0.1211)	0.1204(0.0967)	0.1825(0.1667)	0.0000(0.0000)	0.0000(0.0000)
Avg	0.4334(0.0395)	0.1340(0.0411)	0.5107(0.0859)	0.0030(0.0042)	0.0059(0.0065)
	0.3974(0.0552)	0.1480(0.0427)	0.2170(0.0766)	0.0062(0.0027)	0.0003(0.0050)

Table 3: Evaluation results (DSC) of representative general-purpose segmentation foundation models for zero-shot promptable PET and MRI segmentation. Results in red and teal are the performance of PET and MRI, respectively. Light grey and dark gray shaded cells indicate results that are significantly better with $p<0.01$ and $p<0.0001$.

development. The discrepancy is not merely a matter of missing data. Structural and functional imaging encode fundamentally different representations of human biology. CT and MRI primarily capture anatomical morphology, whereas PET reflects metabolic and physiological activity. Consequently, models trained predominantly on structural signals learn priors centered on boundaries, textures, and geometric organization. These priors transfer poorly to functional imaging, where clinically meaningful information is often diffuse and heterogeneous. As the clinical significance of functional imaging is increasing towards more precise characterization of pathological processes, integrating metabolic and physiological information with structural anatomy is becoming indispensable Andreou et al. (2022); Rowe & Pomper (2022). Our evaluation demonstrates that this limitation persists even when segmentation targets are held constant. The substantial degradation observed across all evaluated models indicates that current foundation models have not yet achieved modality universality, despite strong performance on conventional structural-imaging benchmarks.

Regional discrepancy: the fragmentation of whole-body understanding. A related evaluation gap concerns anatomical coverage. The documented training datasets are concentrated in a limited number of body regions, notably the abdomen and brain. Although recent initiatives have expanded anatomical diversity, whole-body coverage remains uneven across the evaluated models. This fragmentation becomes particularly problematic in whole-body imaging applications. Many clinically important tasks, especially in oncology, require understanding relationships across multiple organs and organ systems simultaneously rather than interpreting isolated anatomical structures. Whole-

body PET imaging is a representative example, where disease burden, metastatic spread, and treatment response often manifest as distributed biological processes spanning the entire body Schwenck et al. (2023). In contrast, anatomical segmentation tasks are routinely framed as localized problems to segment a specific organ, evaluated in isolation from its systemic context. As a result, models trained primarily on localized segmentation tasks may perform adequately within familiar anatomical regions while struggling to generalize across the broader anatomical landscape encountered in routine clinical practice.

Modality and regional coverage are also not independently manipulated in our data. Whole-body PET jointly probes a modality that is sparse in documented training corpora and an imaging setting with broad anatomical coverage. We therefore treat these as complementary axes for future benchmark design. This observation motivates a broader perspective on universality. Rather than viewing generalization as a single property, we argue that universality should be considered along at least two orthogonal axes: 1) Modality universality: robustness across structural and functional imaging paradigms. 2) Anatomical universality: competence across the full spatial extent of the human body. Current foundation models have made substantial progress toward general-purpose segmentation within structurally biased training distributions. However, the evidence presented here suggests that genuine whole-body multimodal universality remains an open challenge. Recognizing this distinction is essential for designing future benchmarks and foundation models that more faithfully reflect the complexity of real-world clinical imaging.

5 Conclusion

Over the past several years, remarkable progress has been achieved in developing general-purpose segmentation foundation models. These systems have demonstrated impressive capabilities across diverse anatomical targets and have substantially advanced the field toward unified medical image analysis. However, our findings suggest that strong performance across existing benchmarks should not be conflated with genuine universality. Current evaluations predominantly assess models within a restricted subspace of medical imaging, largely centered on structural modalities and region-specific tasks. As a result, important dimensions of clinical reality remain insufficiently represented. More complete claims of general-purpose capability should therefore include robustness across structural and functional imaging and across broad anatomical coverage. These dimensions are increasingly relevant as clinical imaging moves toward integrated anatomical-metabolic assessment and system-level disease characterization.

Consequently, future progress should be driven not only by larger models or larger datasets, but also by a redefinition of evaluation itself. In the era of deep learning, evaluation benchmarks serve as the compass for the research community, shaping how progress is defined. When the benchmarks themselves reflect the modality and regional discrepancies described above, evaluation becomes a reflection of bias rather than a test of genuine capability. Benchmarks should evolve from heterogeneous collections of isolated tasks toward unified grounded testbeds that explicitly measure robustness across modalities and across the full spatial extent of the human body. Such evaluations are essential for distinguishing genuine generalization from benchmark-specific optimization. Recent efforts such as Touchstone Bassi et al. (2024) have begun to establish standardized evaluation protocols for abdominal CT segmentation, representing an important step toward more rigorous and reproducible evaluation. We view this as a promising beginning. An important next step may be to extend such benchmarking frameworks beyond regional structural imaging tasks toward whole-body multimodal evaluation settings that better reflect the diversity of real-world clinical imaging.

Ultimately, the goal is not merely to develop models capable of segmenting more organs, but to build foundation models that can interpret the human body as an integrated biological system. Achieving this vision will require closer convergence between structural and functional imaging, between local and global understanding, and between benchmark performance and real-world clinical utility. While current foundation models represent an important milestone on this trajectory, the evidence presented here suggests that the path toward true universality remains only partially explored. In this sense, universality should not be viewed as an established achievement, but as an evolving objective that must be continually re-examined as models expand to increasingly diverse imaging conditions and clinical scenarios. Such a perspective, we believe, is essential for truly reconsidering what universality means in medical imaging foundation models.

REFERENCES

Chrysafis Andreou, Ralph Weissleder, and Moritz F Kircher. Multiplexed imaging in oncology. *Nature Biomedical Engineering*, 6(5):527–540, 2022.

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.

Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2245–2264, 2025.

Pedro R Bassi, Wenxuan Li, Yucheng Tang, Fabian Isensee, Zifu Wang, Jieneng Chen, Yu-Cheng Chou, Saikat Roy, Yannick Kirchhoff, Maximilian Rokuss, et al. Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation? *Advances in Neural Information Processing Systems*, 37:15184–15201, 2024.

Tugba Akinci D'Antonoli, Lucas K. Berger, Ashraya K. Indrakanti, Nathan Vishwanathan, Jakob Weiß, Matthias Jung, Zeynep Berkarda, Alexander Rau, Marco Reisert, Thomas Küstner, Alexandra Walter, Elmar M. Merkle, Martin Segeroth, Joshy Cyriac, Shan Yang, and Jakob Wasserthal. Totalsegmentator mri: Sequence-independent segmentation of 59 anatomical structures in mr images, 2024.

Yuxin Du, Fan Bai, Tiejun Huang, and Bo Zhao. Segvol: Universal and interactive volumetric medical image segmentation. *Advances in neural Information Processing Systems*, 37:110746–110783., 2024.

Sergios Gatidis, Marcel Früh, Matthias P Fabritius, Sijing Gu, Konstantin Nikolaou, Christian La Fougère, Jin Ye, Junjun He, Yige Peng, Lei Bi, et al. Results from the autopet challenge on fully automated lesion segmentation in oncologic pet/ct imaging. *Nature Machine Intelligence*, 6(11): 1396–1405, 2024.

Yufan He, Pengfei Guo, Yucheng Tang, Andriy Myronenko, Vishwesh Nath, Ziyue Xu, Dong Yang, Can Zhao, Benjamin Simon, Mason Belue, et al. Vista3d: A unified segmentation foundation model for 3d medical imaging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20863–20873, 2025.

Ademola E Ilesanmi, Taiwo O Ilesanmi, and Babatunde O Ajayi. Reviewing 3d convolutional neural network approaches for medical image segmentation. *Heliyon*, 10(6), 2024.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Fabian Isensee, Maximilian Rokuss, Lars Krämer, Stefan Dinkelacker, Ashis Ravindran, Florian Stritzke, Benjamin Hamm, Tassilo Wald, Moritz Langenberg, Constantin Ulrich, et al. nninteractive: Redefining 3d promptable segmentation. *arXiv preprint arXiv:2503.08373*, 2025.

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*, pp. 4015–4026, 2023.

Alain Lalande, Zhihao Chen, Thibaut Pommier, Thomas Decourselle, Abdul Qayyum, Michel Salomon, Dominique Ginhac, Youssef Skandarani, Arnaud Boucher, Khawla Brahim, et al. Deep learning methods for automatic evaluation of delayed enhancement-mri. the results of the emidec challenge. *Medical Image Analysis*, 79:102428, 2022.

Jun Ma and Bo Wang. Towards foundation models of biological image segmentation. *Nature Methods*, 20(7):953–955, 2023.

Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, Shucheng Cao, Qi Zhang, Shangqing Liu, Yunpeng Wang, Yuhui Li, Jian He, and Xiaoping Yang. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6695–6714, 2022.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:1–9, 2024.

Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023.

Christophe Nioche, Fanny Orlhac, Sarah Boughdad, Sylvain Reuzé, Jessica Goya-Outi, Charlotte Robert, Claire Pellot-Barakat, Michael Soussan, Frédérique Frouin, and Irène Buvat. Lifex: a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. *Cancer research*, 78(16):4786–4789, 2018.

Steven P Rowe and Martin G Pomper. Molecular imaging in oncology: current impact and future directions. *CA: a cancer journal for clinicians*, 72(4):333–352, 2022.

Johannes Schwenck, Dominik Sonanini, Jonathan M Cotton, Hans-Georg Rammensee, Christian la Fougère, Lars Zender, and Bernd J Pichler. Advances in pet imaging of cancer. *Nature Reviews Cancer*, 23(7):474–490, 2023.

Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.

Haoyu Wang, Sizheng Guo, Jin Ye, Zhongying Deng, Junlong Cheng, Tianbin Li, Jianpin Chen, Yanzhou Su, Ziyan Huang, Yiqing Shen, et al. Sam-med3d: a vision foundation model for general-purpose segmentation on volumetric medical images. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel T. Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

Shaoting Zhang and Dimitris Metaxas. On the challenges and perspectives of foundation models for medical image analysis. *Medical image analysis*, 91:102996, 2024.

Yichi Zhang, Qingcheng Liao, Le Ding, and Jicong Zhang. Bridging 2d and 3d segmentation networks for computation-efficient volumetric medical image segmentation: An empirical study of 2.5 d solutions. *Computerized Medical Imaging and Graphics*, pp. 102088, 2022.

Yichi Zhang, Zhenrong Shen, and Rushi Jiao. Segment anything model for medical image segmentation: Current applications and future directions. *Computers in Biology and Medicine*, pp. 108238, 2024.

Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Sid Kiblawi, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature methods*, 22(1): 166–176, 2025a.

Ziheng Zhao, Yao Zhang, Chaoyi Wu, Xiaoman Zhang, Xiao Zhou, Ya Zhang, Yanfeng Wang, and Weidi Xie. Large-vocabulary segmentation for medical images with text prompts. *NPJ Digital Medicine*, 8(1):566, 2025b.

A IMAGING BACKGROUND.

Computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET) provide complementary but physically different representations of the body. CT measures X-ray attenuation and expresses reconstructed values on the approximately standardized Hounsfield-unit scale. It generally provides high-resolution structural information, with strong contrast at interfaces such as air–lung, soft tissue–fat, and soft tissue–bone. MRI derives image contrast from nuclear-magnetic-resonance signals whose appearance depends on the pulse sequence, scanner, and acquisition parameters. MRI therefore provides rich soft-tissue contrast, but unlike CT, its intensity values do not have a universal quantitative scale across sequences and sites. PET is a functional imaging technique: after administration of a radiotracer, it estimates the spatial distribution of tracer uptake and thereby visualizes metabolic or molecular processes. PET commonly has lower spatial resolution and less sharply defined organ boundaries than CT or MRI, and normal physiological uptake can be intense in structures such as the brain, heart, and urinary bladder. These physical and practical differences also affect the availability of large-scale datasets. CT and MRI are widely used in clinical practice and have generated large volumes of imaging data. In contrast, PET examinations generally involve radiotracer administration, specialized imaging infrastructure, and higher per-examination costs, which limit their routine use and the scale at which data can be collected and publicly released. Consequently, publicly available PET datasets remain substantially smaller than those available for CT and MRI. This relative scarcity of large-scale PET data presents an important challenge for developing and evaluating data-intensive medical AI models.

B DATA ACQUISITION AND ANNOTATION

Patients fasted for at least 6 hours with blood glucose levels confirmed at < 11.1 mmol/L prior to receiving an intravenous injection of ^{18}F-FDG. The scanning field was from the mid-thigh level to the top of the head with the subject in the supine position. All paired multimodal scans cover consistent whole-body anatomical ranges, enabling direct intra-subject comparison between structural anatomical contrasts and metabolic signals.

To construct reliable segmentation ground truth, we adopted a two-stage annotation pipeline. Initial coarse segmentation masks were automatically generated via pre-trained segmentation tools Wasserthal et al. (2023); D'Antonoli et al. (2024), and all preliminary annotations were subsequently reviewed, corrected, and refined by radiologists using the LIFEx Nioche et al. (2018) platform.

C IMPLEMENTATION DETAILS

All experiments were conducted on NVIDIA A100 GPUs with 80GB of memory. We evaluate the zero-shot performance of these models for promptable segmentation without any fine-tuning, thereby assessing their out-of-the-box segmentation capability.

SAM-Med3D Wang et al. (2025) is a promptable segmentation foundation model designed specifically for volumetric medical images. In contrast to applying a two-dimensional model independently to successive slices, it uses a fully learnable 3D image encoder, prompt encoder, and mask decoder to process volumetric context directly. The model accepts spatial prompts, including foreground and background points, and predicts a 3D binary mask for the prompted target. SAM-Med3D-turbo is a later, further-trained checkpoint released by the authors to improve the coverage and robustness of the original model. We evaluated this released turbo checkpoint following its official pipeline at `https://github.com/uni-medical/SAM-Med3D`.

SegVol Du et al. (2024) is a universal and interactive volumetric segmentation model developed primarily from large-scale CT data. Its training pipeline combines self-supervised pre-training on approximately 90,000 unlabeled CT volumes with supervised training on approximately 6,000 labeled CT volumes, covering more than 200 anatomical categories. SegVol supports semantic prompts as well as spatial point and box prompts, and introduces a zoom-out zoom-in inference strategy that first localizes the target at a coarse scale and then refines the mask at a higher resolution. We evaluated SegVol following its official pipeline at `https://github.com/BAAI-DCAI/SegVol`.

nnInteractive Isensee et al. (2025) is an open-set 3D interactive segmentation framework derived from the design principles of nnU-Net. It supports several forms of user guidance including foreground and background points, scribbles, bounding boxes, and lasso prompts provided through intuitive 2D interactions while producing a full volumetric segmentation. The model was trained on more than 120 heterogeneous 3D datasets spanning CT, MRI, PET, and microscopy, and is intended to segment previously unseen structures without a predefined class vocabulary. Its inference framework combines an initial prediction with interaction-driven refinement and automatically focuses computation on the relevant image region. We evaluated nnInteractive following its official pipeline at `https://github.com/MIC-DKFZ/nnInteractive`.

VISTA3D He et al. (2025) is a unified 3D segmentation and annotation model that combines automatic class-conditioned segmentation with interactive correction. A shared 3D image encoder feeds an automatic branch activated by a class-ID prompt and an interactive branch activated by point prompts. The released model provides automatic predictions for 127 supported anatomical and lesion classes and uses 3D supervoxels to transfer information from pre-trained 2D representations. Because the UMD target organs belong to the model's supported semantic vocabulary, we evaluated VISTA3D through its automatic class-ID interface following its official pipeline at `https://github.com/Project-MONAI/VISTA/tree/main/vista3d`.

SAT Zhao et al. (2025b) is a knowledge-enhanced, text-prompted foundation model for large-vocabulary 3D medical image segmentation. It organizes anatomical terminology through a medical knowledge tree and aligns a text encoder with volumetric image features, allowing a user to specify the desired structure by name rather than by spatial interaction. The source study assembled more than 22,000 scans from 72 datasets, encompassing 497 segmentation categories, three imaging modalities, and eight body regions. SAT-Pro is the larger reported variant, with 447 million parameters, and is intended to transfer semantic knowledge across datasets and label conventions. We evaluated SAT-Pro using the organ-name text prompt expected by its official pipeline at `https://github.com/zhaoziheng/SAT`.